International Journal of Interactive Multimedia and Artificial Intelligence

September 2021, Vol. VI, Number 7
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

This picture titled "Uncertainty" is authored by José Mª Chow Casas, winner of the call "2021: the year of..." launched by the Virtual Museum of the Engineering and Technology School of UNIR.
2021

Special Issue on Current Trends in Intelligent Multimedia Processing Systems

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE

ISSN: 1989-1660 –VOL. 6, NUMBER 7

EDITORIAL TEAM

Editor's Note

In today's world, we have witnessed an onset of multimedia content being uploaded/downloaded and shared through a multitude of platforms both online and offline. In support of this trend, multimedia processing and analyzing has become very popular in all kinds of information extraction and attracts research interest from both academia and industry. This is to be expected as the multimedia digital world is worth trillions of dollars worldwide. However, multimedia information is hard to encode, interpret and recognize because it is combined with many complex components. Recently, there are many research areas related to the overall notion of intelligent multimedia processing. Therefore, the collected papers in this special issue provide a systematic overview and state-of-the-art research in the field of intelligent multimedia processing and analyzing system and outline new developments in fundamental, theorems, approaches, methodologies, software systems, recommendations, and real-world applications in this area. The collected research works in this special issue are described as follows.

In the first paper, titled "A Novel Fog Computing Approach for Minimization of Latency in Healthcare using Machine Learning", the authors presented a novel Intelligent Multimedia Data Segregation (IMDS) scheme using Machine learning (k-fold random forest) in the fog computing environment. The designed model segregates the multimedia data and calculates total latency in terms of transmission, computation, and network. With the simulated results, the developed model achieved 92% classification accuracy, and an approximately 95% reduction in latency compared with the pre-existing model and improved the quality of services in e-healthcare.

In the second paper, titled "An enhanced texture-based feature extraction approach for classification of biomedical images of CT-Scan of Lungs", the authors considered a feature vector by concatenation of features extracted from the local mesh peak valley edge pattern (LMePVEP) technique. A dynamic threshold-based local mesh ternary pattern technique and texture of the image in five different directions are then considered in the developed model. The concatenated feature vector is then used to classify images of two datasets. The proposed framework has improved the accuracy by 12.56%, 9.71% and 7.01% on average for dataset 1 and 9.37%, 8.99% and 7.63% on average for dataset 2 over three popular algorithms.

In the next paper "Alzheimer Disease Detection Techniques and Methods: A Review", the authors presented a systematic review of Alzheimer's disease based on Neuroimaging and cognitive impairment classification, which is mainly focused on computer-aided diagnosis. This study revealed that the classification criterion based on the features shows promising results to diagnose the disease and helps in clinical progression. The most widely used machine learning classifiers for AD diagnosis including Support Vector Machine, Bayesian Classifiers, Linear Discriminant Analysis, and K-Nearest Neighbor along with Deep learning are then studied and investigated. The possible challenges along with future directions are also discussed in the paper.

The work "Imputation of Rainfall Data Using the Sine Cosine Function Fitting Neural Network" proposes a novel pre-processing mechanism for non-precipitation data by using principal component analysis (PCA). The PCA in the developed model is used to extract the most relevant features from the meteorological data. The output of the PCA is combined with the rainfall data from the nearest neighbour gauging stations and then used as the input to the neural network for missing data imputation. In addition, a sine cosine algorithm is presented to optimize the neural network for infilling the missing rainfall data. The proposed SC-FITNET model outperformed LSTM, SC-FFNN and FFNN imputation in terms of mean absolute error (MAE), root mean square error (RMSE) and correlation coefficient (R), with an average accuracy of 90.9%.

In the paper titled "Design and Development of an Energy Efficient Multimedia Cloud Data Center with Minimal SLA violation", the authors highlight a novel virtual machine (VM) selection policy based on identifying the Maximum value among the differences of the Sum of Squares Utilization Rate (MdSSUR) parameter to reduce the energy consumption (EC) of multimedia cloud data centers with minimal service level agreement violation (SLAV). The proposed MdSSURVM selection policy has been evaluated using real workload traces in CloudSim. The simulation results demonstrate that the designed MdSSUR VM selection policy achieves the rate of improvements of the EC, the number of VM migrations, and the SLAV by 28.37%, 89.47%, and 79.14%, respectively.

In the paper titled "A Generalized Wine Quality Prediction Framework by Evolutionary Algorithms", the authors propose the generalized wine quality prediction framework to provide a mechanism for finding a useful hybrid model for wine quality prediction. Based on the developed framework, the generalized wine quality prediction algorithm using the genetic algorithms is proposed. It first encodes the classifiers as well as their hyperparameters into a chromosome. The fitness of a chromosome is then evaluated by the average accuracy of the employed classifiers. The genetic operations are performed to generate new offspring. The evolution process is continuing until reaching the stop criteria. As a result, experiments on the wine datasets were made to show the merits and effectiveness of the proposed approach.

The authors of the paper titled "Optimal QoE Scheduling in MPEG-DASH Video Streaming" designed a series of click density experiments to verify whether different resolutions have different quality of experience (QoE) effects in different video scenes. To evaluate true user's experience, the authors convert viewing QoE into a satisfaction quality score, called Q-score, for different resolutions of each video segment. Additionally, the authors developed an optimal segment assignment (OSA) algorithm for the Q-score optimization in a constraint network bandwidth. The experimental results showed that the playback schedule by applying the OSA algorithm significantly improved users' viewing satisfaction.

In the next paper, "Pulmonary nodule classification in lung cancer from 3D thoracic CT scans using fastai and MONAI", the authors construct a convolutional neural network to classify pulmonary nodules as malignant or benign in the context of lung cancer. To construct and train the model, the fastai deep learning framework is extended and investigated to 3D medical imaging tasks, combined with the MONAI deep learning library. The authors train and evaluate the model using a large, openly available data set of annotated thoracic CT scans. The designed model then achieves a nodule classification accuracy of 92.4% and a ROC AUC of 97% when compared to a "ground truth" based on multiple human raters' subjective assessment of malignancy. Also, the developed model achieves a test set accuracy of 75% for predicting patient-level diagnoses of cancer.

In the paper titled "Modeling of Performance Creative Evaluation Driven by Multimodal Affective Data", the authors proposed a Performance Creative-Multimodal Affective (PC-MulAff) model based on the multimodal effective features for performance creative

DOI: 10.9781/ijimai.2021.08.014

evaluation. The multimedia data acquisition equipment is used to collect the physiological data of the audience, including the multimodal affective data such as the facial expression, heart rate and eye movement. The designed model thus calculates effective features of multimodal data combined with director annotation and defines Performance Creative-Affective Acceptance (PC-Acc) based on multimodal affective features to evaluate the quality of performance creative. Results showed that the accuracy of the mPC-MulAff model is 7.44% and 13.95% higher than that of the single textual and single video evaluation.

In the next work, "Modified YOLOv4-DenseNet Algorithm for Detection of Ventricular Septal Defects in Ultrasound Images", the authors first solve the object detection problem of the ventricular septal defect (VSD) by using a modified YOLOv4–DenseNet framework regarding the echocardiographic images that are used for diagnosing congenital heart diseases (CHDs). The results revealed that the YOLOv4–DenseNet outperformed YOLOv4, YOLOv3, YOLOv3–SPP, and YOLOv3–DenseNet in terms of metric mAP-50. The F1-score of YOLOv4-DenseNet and YOLOv3-DenseNet were better than those of others. Thus, the modified model establishes the feasibility of using deep learning for echocardiographic image detection of VSD investigation.

In the last paper, titled "Integration of Genetic Programming and TABU Search Mechanism for Automatic Detection of Magnetic Resonance Imaging in Cervical Spondylosis", the authors adopted a heuristic programming, genetic programming (GP), to build the core of the refereeing engine by combining the TABU search (TS) with the evolutionary GP. To validate the accuracy of the proposed model, the authors implemented experiments and compared the prediction results with the radiologist's diagnosis to the same magnetic resonance image (MRI). The experiment found that using clinical indicators to optimize the TABU list in GP+TABU got better fitness than the other two methods and the accuracy rate of the proposed model can achieve 88% on average. Thus, the proposed model can help radiologists reduce the interpretation effort and improve the relationship between doctors and patients.

Jerry Chun-Wei Lin[1], Gautam Srivastava[2], and Vincent S. Tseng[3]

[1] Western Norway University of Applied Sciences, Bergen (Norway)

[2] Brandon University, Brandon (Canada)

[3] National Yang Ming Chiao Tung University, Hsinchu (Taiwan)

TABLE OF CONTENTS

EDITOR'S NOTE 4

A NOVEL FOG COMPUTING APPROACH FOR MINIMIZATION OF LATENCY IN HEALTHCARE USING MACHINE LEARNING 7

AN ENHANCED TEXTURE-BASED FEATURE EXTRACTION APPROACH FOR CLASSIFICATION OF BIOMEDICAL IMAGES OF CT-SCAN OF LUNGS 18

ALZHEIMER DISEASE DETECTION TECHNIQUES AND METHODS: A REVIEW 26

IMPUTATION OF RAINFALL DATA USING THE SINE COSINE FUNCTION FITTING NEURAL NETWORK 39

DESIGN AND DEVELOPMENT OF AN ENERGY EFFICIENT MULTIMEDIA CLOUD DATA CENTER WITH MINIMAL SLA VIOLATION 49

A GENERALIZED WINE QUALITY PREDICTION FRAMEWORK BY EVOLUTIONARY ALGORITHMS 60

OPTIMAL QOE SCHEDULING IN MPEG-DASH VIDEO STREAMING 71

PULMONARY NODULE CLASSIFICATION IN LUNG CANCER FROM 3D THORACIC CT SCANS USING FASTAI AND MONAI 83

MODELING OF PERFORMANCE CREATIVE EVALUATION DRIVEN BY MULTIMODAL AFFECTIVE DATA 90

MODIFIED YOLOV4-DENSENET ALGORITHM FOR DETECTION OF VENTRICULAR SEPTAL DEFECTS IN ULTRASOUND IMAGES 101

INTEGRATION OF GENETIC PROGRAMMING AND TABU SEARCH MECHANISM FOR AUTOMATIC DETECTION OF MAGNETIC RESONANCE IMAGING IN CERVICAL SPONDYLOSIS 109

OPEN ACCESS JOURNAL

ISSN: 1989-1660

The International Journal of Interactive Multimedia and Artificial Intelligence is covered in Clarivate Analytics services and products. Specifically, this publication is indexed and abstracted in: *Science Citation Index Expanded, Journal Citation Reports/ Science Edition, Current Contents®/Engineering Computing and Technology.*

A Novel Fog Computing Approach for Minimization of Latency in Healthcare using Machine Learning

Amit Kishor[1]*, Chinmay Chakraborty[2], Wilson Jeberson[1]

[1] Department of Computer Science & Information Technology, Sam Higginbottom University of Agriculture, Technology and Sciences (SHUATS), Allahabad, U.P. (India)
[2] Assistant Professor, Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Jharkhand (India)

Received 28 October 2020 | Accepted 23 December 2020 | Published 21 December 2020

unir
LA UNIVERSIDAD EN INTERNET

Abstract

In the recent scenario, the most challenging requirements are to handle the massive generation of multimedia data from the Internet of Things (IoT) devices which becomes very difficult to handle only through the cloud. Fog computing technology emerges as an intelligent solution and uses a distributed environment to operate. The objective of the paper is latency minimization in e-healthcare through fog computing. Therefore, In IoT multimedia data transmission, the parameters such as transmission delay, network delay, and computation delay must be reduced as there is a high demand for healthcare multimedia analytics. Fog computing provides processing, storage, and analyze the data nearer to IoT and end-users to overcome the latency. In this paper, the novel Intelligent Multimedia Data Segregation (IMDS) scheme using Machine learning (k-fold random forest) is proposed in the fog computing environment that segregates the multimedia data and the model used to calculate total latency (transmission, computation, and network). With the simulated results, we achieved 92% as the classification accuracy of the model, an approximately 95% reduction in latency as compared with the pre-existing model, and improved the quality of services in e-healthcare.

Keywords

Cloud Computing, Data Segregation Scheme, Fog Computing, Latency, Machine Learning, Multimedia Healthcare Data Analytics, Multimedia Transmission, Quality Of Service.

DOI: 10.9781/ijimai.2020.12.004

I. Introduction

As per the International Data Corporation report, there will be 41.6 billion to 1 trillion IoT devices and that will generate a huge amount of data in zettabytes by 2025. There is a big demand of wireless communication due to many reasons such as the tremendous increase in the popularity of IoT devices, extensive use of social media, the dissemination of different mobile application, the population growth of the world, and the present lifestyle that is highly dependent on the latest technology in every aspect. A huge number of multimedia data is generated by IoT devices used in healthcare it is very important to process multimedia data in the healthcare sector, Cloud servers are mostly used world-wide to handle the immense data generated by these IoT devices. The extraction of information's about patient health from supplied analyzed multimedia data is plays a very important and crucial role. Analysis of data, storage of data, pre-processing of data is done by cloud servers. Mainly the cloud computing is the probably viable solution for establishing communication between IoT and healthcare [1]. The healthcare data generated by IoT devices is analyzed, filtered, pre-processed and aggregated only on the cloud. Cloud computing has its limitations. As the data transmission rate increases, due to the receiving of these excessive volumes of data, the response time is increasing in the cloud environment. A higher service delay has occurs to end-users. A high volume of data transmission over the network increases the probability of occurrence of an error and the delay. The loss of data packets and transmission latency is directly related to the quantity of data transmitted through IoT devices to the cloud. Due to this reason, it causes a low quality of service (QoS) produced to the end-user. Cloud computation and data storage are generally not desired in most of the time-sensitive applications of the IoT. Extreme time-bounded problems must be completed nearer to the IoT devices. As the healthcare infrastructures' main requirements are minimization in latency and reduction in network bandwidth, for this it requires data in real-time for a time-critical scenario [2]. The connection is established between end devices and the cloud through routers and gateways. Thus, a huge wide variety of routers are positioned among the cloud and the healthcare IoT's. Due to routers, computational delay increases. As the distance is larger, the large numbers of routers are connected between cloud and IoT devices, because of that a long route is travelled by data from source to destination and it consumes high bandwidth.

For the utilization of the complete advantages of the IoT with fog, it is essential to make available enough networking and infrastructure to produce minimum latency and rapid responding time for IoT applications. Fog computing is introduced as a prime catalyser for the execution and processing of the data generated by IoT devices. It is more effective to shift the applications, execution, and processing capabilities nearer to IoT devices that generated the data. The fog computing concept is properly well suited to resolve these issues.

Fog computing is an emerging concept that uses the processing and execution capabilities closer to the end-user to achieve an improved quality of service as previously used in the cloud paradigm [3]-

* Corresponding author.
E-mail address: amit_kishor@rediffmail.com

[4]. The fog computing layer is placed in between IoT and cloud; it brings low latency and low network usage. Fog computing provides storage, pre-processing, execution, networking, and computational services to their end-users at the edge of the network and closer to end-users. Despite the number of advantages of fog computing, the available research-work is still immature in this domain, and numbers of researchers are still working on the challenges of fog-computing and basic architecture [3]-[6]. The main challenges of fog devices are privacy, security, and consumption of energy. Fog computing is used to overcome the limitations of cloud computing. Fog Computing is used for such applications that require minimum latency and it works on geo-distribution, which is fast and transferable, and has a broad level distribution control system. It enables distributed and computation with low latency at the edge of the network to provides support to IoT applications. Sufficient amounts of available data can be stored, computed, and processed over the fog-networks and that can be controlled by end-users [7]. An open research aim is to improve the quality of services of fog computing by introducing a fog layer between cloud and IoT devices.

The motivation came from a study about how to generate minimum response time with a better quality of service for time-sensitive healthcare IoT based applications. The cloud alone is not able to satisfy the aforementioned requirements due to their limitations. The patient's physical status varies with time and needs rapid response as an action to monitor remote patients. This is possible when there is a very good network available. Otherwise, it will take more time to respond. In fact, due to unpredictable networks, there is high latency, the health data of patients are not considered as real-time data. This shows that the data become unreliable, worthless, and insufficient. The delay may increase for these IoT time-sensitive data from milli-seconds (ms) to seconds and then reaches to minutes [8]-[9], When the size of health data increases, therefore the situation become worsening in handling real-time operation [10]-[11]. The QoS requirements for medical health data [12]-[13] are shown in Table I and the QoS service requirements are shown in Table II for e-healthcare services.

TABLE I. QoS Requirement Parameter for Healthcare and Medical Data Transmission

Services for Healthcare	Data Rate (kbps)	Maximum Delay	Loss of Packet (%)
Audio	4.0–25.0	150.0–400.0 ms	3.0
Video	32.0–384.0	150.0–400.0 ms	1.0
Electro-cardio-gram (ECG)	1.0–20.0	Approx. 1 second	0.0
File -Transfer (FTP)	-	-	0.0

The main contributions of the research work-study are as follows:

1. An analytical model based on fog computing is proposed to transfer healthcare sensor data to end-users in real-time.
2. A random forest algorithm is implemented which reduces and avoids the "over-fitting" issues.
3. The proposed research scheme minimizes the total latency between healthcare sensors and cloud servers. A performance comparison is conducted for the proposed analytical model with existing models on different parameters.
4. To improve the quality of service for e-healthcare.

The remaining part of the paper is organized as follows; Section II describes related work. Section III, introduces a proposed system model for IoT-Fog-Cloud applications. Section IV of the paper is about the work done for the proposed system model. The analysis based on simulation results are provided in Section V. Section VI, comprises the conclusion and provides the future scope of the paper.

TABLE II. QoS Requirement Parameter Services for E-Healthcare

Types of e-healthcare services	e-healthcare system examples	Media type used	Maximum delay
Audio-based communication in real-time	Conferencing (audio) among patients/end-users or end-user/doctors	Audio	< 150.0 ms one way end to end
Video-based communication in real-time	Conferencing (video) among patients/end-users or end-user/doctors	Video	< 250.0 ms one way end to end
Robotic services in real-time	Remote based tele-surgery	Control of data, audio, video by robotics	< 300.0 milliseconds round trip time
Monitoring in real-time	Patient's essential sign transmission and video steaming in an urgent scenario	Sensors (to collect biomedical data)	< 300.0 milliseconds for real-time ECG
Real-time diagnosis	Transfer the medical images to remote areas in an urgent scenario	Images, text, data	-
Real-time messaging	Alarm based indication for urgency	Text, data, small images	No

II. Related Work

Silva et al. [14] used fog computing technology to manage patient records. Fog computing is used to overcome the problem of cloud computing for data management with challenges such as availability, performance, and secrecy. Alarm et al. [15] proposed a method to store the health data on the cloud. The data is generated by IoT-devices. They presented a system for data management in the cloud, based on the management of IoT. The collection of data is done in real-time and an alert system is there with a prior defined rule for notification. Nishtala et al. [16] used a combination of heuristic and reinforcement learning technology called Hipster used to control the latency-critical workloads in the cloud. Hipster aims to improve the efficiency of used resources concerning the quality of service. Latency for large computations is not discussed by the author for the cloud environment. Gia et al. [17] proposed research for continuous monitoring of time-sensitive health patient's through fog computing and concern cost is low. They provided automatic notification and analysis. The sensor-node (energy efficient) system is developed with a layer of fog. Medical practitioners access the data collected through sensors. Naas et al. [18] raised the major problem for IoT applications in time-sensitive cases which is resolved by the author with a technique proposed named iFogStor in fog environment. The author proposed a schema called GAP (Generalized Assignment Problem) for the placement of the data in fog computing. For the solution of GAP two methods are used, first is an accurate solution and the other one is the heuristics method. Rahmani et al. [19] discussed the different services such as real-time processing of local data, data-mining (embedded), and some higher-level services. They presented a prototype called UT-GATE for smart e-health gateway and through which they discussed the features. They have shown the enhancement in performance of overall systems. Wu et al. [20] proposed a schema as security services in fog computing in information-concentric social networks. The main contribution is the introduction of fog computing concepts with required parameters end-to-end communication, low latency, and computing resources at the network edge and also improving the security services by content-aware and matching. Although the network delays, as well as computation delay, are not being discussed by the author. Brogi et al. [21] presented a model for the deployment of QoS-aware in IoT used applications by the use of fog computing technology. The model

is used to produce the latency and bandwidth of accessible resources but the author missed discussing network and computation latency.

Shahzad et al. [22] proposed a method to monitor the medical condition in real-time compare to the private cloud. A system is designed and known as BTS (bounded telemonitoring system) for monitoring of patients in real-time. The information for patients is captured in the boundary of the private cloud. They try to provide medical data of patient's with security. Kao et al. [23] introduced the time-critical data analysis in mobile computing for latency minimization the author presented a novel technique with the name of Hermes. The optimization technique based on NP-hard is used for the task data. Li et al. [24] introduced the SPSRP's (service popularity-based smart-resources partitioning) architecture for implementation in IoT and fog computing and also created a mathematical model for the popularity of service and cost of computation on Fog Nodes (FNs). The authors reduced fault tolerance, response-time, and delay time. The calculation for the cost of computing on FNs at the arrival of services from IoT by applying Zipf's law is provided.

Dinh et al. [25] used a service-oriented schema related to cost-effectiveness for providing the service of the IoT-Fog-Cloud network. The authors also used to measure VNF (Virtual Network Function) with development in the capabilities to enhance the availability of SFC (service function chaining) with the proposed metric. Mahmud et al. [26] discussed the problem that occurred in the use of healthcare due to the large volume of transmission of data and high latency. As a solution to these issues, the author presented an IoT-healthcare structure based on fog and explored the cloud-fog service over the traditional cloud. An improvement result is shown for network-traffic, power usage, and the cost. Ahsan et al. [27] highlighted the security, protection, and integrity of the data is a major concern in cloud computing. The author proposed a fog-centric scheme for the storage of data in the cloud. Data security issues had been discussed. Xor-combination is implemented to provide the protection and security of data in the cloud. The proposed method is used to prevent the attack of unauthorized access and malicious users. Hash technique was used in a new form to detect the data alteration with a high occurrence of probability. They also prevented a cyber attack. Rafique et al. [28] used a technique with modification and combination of the PSO (Particle Swarm Optimization) and CSO (Cat Swarm Optimization) to reduce the response time. With the combination of the above two algorithms, they produce NBIHA (Novel bio-inspired hybrid-algorithm) used to overcome the response-time in IoT-Fog-Cloud applications. Li et al. [29] introduced the factors of network delays and designed a framework based on IoT-Fog for estimation of latency. They can predict the delay occurred in the cloud-fog inter-node and proposed a GNP (Global Networking Position) a landmark-based algorithm to predict the latency with good accuracy. Thota et al. [30] presented sensor architecture by using a fog computing platform. Sensors were used to collect patient data and after that sensor send data to edge devices with security. They provided authentication and security of medical data, and unauthorized access was prevented by using asynchronous communication.

Tahani [31], used the scheduling algorithm MAX-MIN on medical data, and then the author used a new method for distribution of task to reduce the waiting time in queue, called TCVC (Task Classification and Virtual Machine Categorization). Raafat et al. [32], presented a model for resource allocation in fog and cloud environments when the data is generated by edge devices. They calculated the overall latency of the model in a fog environment using a genetic algorithm. Pan et al. [33], presented and discussed the current technologies summary report and the compatibility among the cloudlet, home cloud, nebula, fog computing, MEC (mobile-edge-computing). They discussed the different issues related to the aforementioned technologies. But no practical issue is discussed. Chakraborty et al. [36]-[37] measure QoS over heterogeneous networks. Nilashi et al.[38] presented a heart disease prediction model by using fuzzy-SVM. They improve accuracy and computation time. Tarik et al. [39] presented a model for diabetic patients. They analyzed the fasting blood sugar as attributes

TABLE III. Existing Literature Survey

Authors	Proposed Techniques	Advantages	Limitations
Alam et al. [15]	A real-time data collection in fog computing	Data collected in real-time. Transmission delay and computation delay is calculated	There is no calculation for network delay
Nishtala et al. [16]	Hipster : to control time-sensitive issues	Improved efficiency by using network and computation delay	Latency for large data is not discussed and also no method is designed for transmission delay
Naas et al. [18]	iFogStor : GAP for fog computing and heuristic approach	System efficiency is improved. Also they resolve the issue occurred with time-sensitive data.	The transmission delay is not calculated
Wu et al. [20]	Security services as well as content-aware filtering based on fog computing on the edge network	Shifts the task to edge end device from remote locations	There is no discussion about network and computation delay
Brogi et al. [21]	QoS-aware model deployment in IoT by fog computing	Deployed a QoS-aware model in IoT	No explanation for network and computation delay
Shahzad et al. [23]	Hermes: NP-hard technique	Task data is optimized by using NP-hard	There is no explanation for network and computation delay
Li et al. [24]	SPSRP for fog nodes (FNs) and IoT-device	Minimizes the response and delay time in the fog environment	Computation and network delay is not discussed
Dinh et al. [25]	Deployment of cost-effective schema through fog computing for IoT-application	Discusses the issues that occurred due to failure of software and hardware	
Mahmud et al. [26]	IoT-healthcare structure for cloud-fog	Discusses the issue of high volume data. Improved the performance of network traffic and power	Network delay is not being discussed
Ahsan et al. [27]	A fog-centric schema for data storage	Discusses the storage and security of data in the cloud	No discussion about transmission and network delay
Rafique et al. [28]	PSO and CSO techniques	Reduces the response time in the IoT-Fog-Cloud environment	There is no discussion about network and computation delay
Proposed scheme	**IMDS**	**Reduce the overall latency by using transmission, network, and computation delay**	-

for predicting of the diabetics. Mahmud et al. [40] highlighted the recent techniques to capture the different types of patient data in the research. They also captured the video data of the patients. Tarik et al. [41] presented a method for healthcare analysis of patients through meta-heuristic algorithms. This method is very useful for doctors and patients when the patients are suffering from different diseases. Jerry et al., [42] presented a model named BILU-NEMH for extraction and classification of data. They used hypergraph and deep learning concepts to enhance the performance of the designed model. Jerry et al. [43], highlighted the problem faced by sequence labeling and they proposed a model for enhancing the sequence labeling with latent variable conditional random fields. This model is very useful in the stage of the pre-processing of data. Machine understanding becomes strengthens through this model. Ahmed et al. [44], presented a machine learning-based classifier, and the method is mapped with OpenCL. The classification can be accelerated by the use of the proposed method for heterogeneous networks. They also highlighted the solution method for the data imbalance problem. Table III shows the survey on existing literature and Table IV shows the comparative analysis.

TABLE IV. The Comparative Analysis

Authors/Year	Transmission Delay (T_D)	Network Delay(N_D)	Computation Delay(C_D)
Alam et al. [15], 2016	Yes	No	Yes
Nishtala et al. [16], 2017	No	Yes	Yes
Naas et al. [18], 2017	No	Yes	Yes
Wu et al. [20], 2017	Yes	No	No
Brogi et al. [21], 2017	Yes	No	No
Shahzad et al. [23], 2017	Yes	No	No
Li et al. [24], 2018	Yes	No	No
Dinh et al. [25], 2018	Yes	No	No
Mahmud et al. [26], 2018	Yes	Yes	No
Ahsan et al. [27], 2019	No	No	Yes
Rafique et al. [28], 2019	Yes	No	No
Proposed IMDS Algorithm	**Yes**	**Yes**	**Yes**

After analyzing the available research and studying the comparative comprehensive analysis of the reduction in total latency (transmission delay, network delay, and computation delay) among IoT-Fog-Cloud networks, we found that there is a research gap and the available techniques for reducing the latency in healthcare used by the researcher are incomplete. Hence, a novel technique must be required to fill this research gap.

To achieve the imperative execution, the issue of minimization of latency in healthcare cloud and IoT was developed and for the aforementioned aim the system model is presented, the main aim is for the formation of the fog network, to effectively allocate, and distribute the task data. To create the network of fog and unload its task data, a fog node (FN) should search neighboring or adjacent FNs. The neighboring FNs in the system will dynamically appear and disappear. It is well known that, In the healthcare system for monitoring high-risk patients, regular monitoring of patients is required. To maintain the regular monitoring system by the human being is very difficult, tedious and it seems to be an unpractical approach. As a result, carelessness towards the high-risk patient occurs. To avoid such situations, the aim is to evaluate the patient health data to track the probability of any high risk, the system required an analysis of a large volume of healthcare data set and attributes. Random forest is applied for the detection, segregation, and analysis of data. Random forest is selected to avoid the over-fitting problem. The predicted data is sent to the end-user within minimum time. Here, to find the availability of adjacent FN for computation is very difficult. In addition to it, the total numbers of adjacent or neighboring FNs with their locations are unidentified and extremely unpredictable, so it is very challenging to manage the fog network creation and task data distribution process. So under such an unpredictable condition, also considering neighboring FNs is accountable for the appearance of new FNs, which also produces a higher data rate and increasing computational capabilities.

TABLE V. Description of Data Used in the Proposed Model

Sr. No.	Variable Name	Description
1	Age	Patient's age (in years)
2	Sex	Male/Female as 1/0
3	CP	Chest Pain type (result 1: Angina, result 2: A typical way of Angina, result 3: Not-angina, result 4: Angina symptom nil)
4	Trestbps	Blood Pressure values in resting in mm Hg
5	Chol	Cholesterol results in mg/dl
6	FBS	Blood Sugar results in fasting>120 mg/dl (1 as true; 0 as false)
7	Restecg	ECG resting results (result 0 for normal; result 1 and 2 for abnormal)
8	Thalach	Heart Rate (maximum) as recorded
9	Exang	Prompted Angina exercise(1 as yes; 0 as no)
10	Oldpeak	ST Depression prompted by Exercise as compare to rest
11	Slope	The slope respect to peak of exercise (result 1,2, and 3 for up sloping, flat, and downsloping)
12	CA	Major vessels number (total)
13	Thalrest	Values (at rest) of heart rate
14	NUM	Status of heart disease (result 0 = no heart disease; result greater than 0 = heart disease)

III. Proposed Model

The Fog computing environment based IoT- healthcare system model shown in Fig. 1. The proposed model collects the patient data as per table V. The data is transmitted from IoT or sensor devices and then data is classified into three categories such as low sensitive risk, normal, and high sensitive risk by applying random forest machine learning algorithm. Healthcare sensor data offload their task data to fog servers. After processing healthcare data the time-sensitive data are sent to the end-user in minimum time. FNs are used to distribute and allocate the task data packets in different available nodes and end-user. A principal FN manager is used and that maintains the topological details of task data packet distribution and allocation. Network topology is used to connect the nodes and every FNs are then linked with the principal FN. Here the study shows a continual

task data packet allocation system using fog computing environment in machine learning. FNs transfer the task data packet to other FNs in the network to minimize the latency and reduce the network traffic. Here, the processing unit comprises the task data packets in a transmission queue, and then the task data is sent for computation in the computation queue, and these impacts the response time. The entire FN collects details, composes a decision, provides task data information and maintains the queue position. A network table is created by the principal FN and it considers all information that was distributed by the other nodes. The principal FN sends a request to other FNs to determine their availability to process the required data. After getting the availability, the task data will be sent to the nearest FN, where the allotment is performed based on requisite data and time. The work aims to reduce or minimize latency and traffic of network with the selection of time-sensitive data. The process of interacting is that IoT sends data to FN and received data is served by the same FN if data is small otherwise FN serves it partially and sends the remaining data to neighboring FNs to serve. The neighboring FNs compute the data if they are not currently processing any data otherwise data have to wait in the waiting queue for processing. After processing of data from neighboring nodes, these send back the data to the original FN (which transfers the task data first) then the FN sends it to the end-user or cloud. Therefore the selection of the best neighboring nodes is very important for task processing otherwise it increases the waiting time in the computation queue.

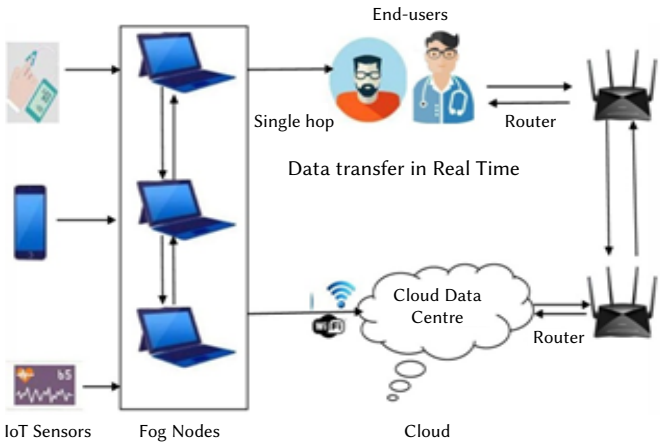


Fig. 1. IoT- HealthCare System Model.

The principal node will maintain an availability table in which all the detail of FNs will be available and maintain all the current updates of the status of FNs. The availability of nodes is shown in table VI, containing the approximate waiting time and the processing speeds of FNs. If any FN became free early then it updates the approximately waiting time in the availability nodes table. The principal node will regularly update the approximately waiting time. The selection of best neighboring FN can be performed by selecting the highest processing speed with a minimum of approximately waiting time.

TABLE VI. Availability of Nodes

Node Description	Approx Weighting Time (microseconds)	The processing speed of FNs (in G.Hz)
132.115.76.84	245	2.201
132.115.76.86	237	2.202
……	…..	….
……	…..	…..

The random forest machine learning algorithm [34]-[35] was used to achieve the aim of the minimization of latency. This aim can be achieved by reducing the used delay as transmission, network, and computation. The proposed model used IoT-Fog-Cloud application's dynamic behavior. A decision-oriented process has been performed using a random forest machine-learning algorithm to overcome the issue of task data demand at a distinct time interval among distinct users and the processing capabilities of FNs. In real-time, the random forest algorithms are used to monitor and care the health data. The main purpose is to minimize the delay that occurs in health monitoring. The FNs identify and select best neighboring FN for computation and process. The quality of service is also a major concern of the entire system. The FNs were used to select the task data communicated by the IoT application in the proposed research. Thereafter, it starts the processing of health task data, and the remaining part of data is transferred to the best neighboring FN and then these processed data is sent to either end-user or cloud in real-time. All the executions are required to be processed in minimum time.

IV. Materials and Methods

Healthcare heart disease data are taken from UC Irvine's machine learning repository [46]. In the simulation heart disease data set encompasses 303 instances and 14 attributes. Although, the UCI repository encompasses 76 attributes in the actual heart disease dataset. In total 14 attributes have been taken for simulation of the proposed algorithm. The testing of the algorithm is performed on these attributes. The attributes are categorized into qualitative and quantitative attributes as shown in the Table V, which shows the data description used in the proposed model. The selection of high-risk data is based on qualitative attributes.

To achieve the objective of the research we applied a k-fold random forest machine learning algorithm. The reason to apply random forest is having better contributions among other classifiers such as SVM (support vector machine), BN (Bayes Network), MP (Multilayer perception), etc. [33]. Feature selection becomes easier in a random forest based algorithm. Estimation of missing values is completed effectively. It avoids over-fitting problems despite that it is a collection of decision trees. Many of research work said that random forest has a quality for prediction of accuracy is excellent for both normal and abnormal data. In a random forest method, the optimization of features is governed by bootstrapped data and this can be performed by k-fold cross-validation (k=10). To avoid overfitting the other scheme such as early stopping and ensembling can also be used. Fig. 2 represents the flow chart of the intelligent multimedia data segregation (IMDS) scheme. The distance travel and the number of hops covered from the sensor to the cloud server is minimum for the high-risk data because it is processed near to sensor devices known as fog computing. By the use of this process, there is a reduction in transmission time due to the total latency time reduces.

In the proposed IMDS algorithm based on k-fold random forest machine learning techniques, the model collects the data at the initial level. Data is pre-processed after collection. Then data is divided into k-fold. Herein k-fold cross-validation is applied. The cross-validation process is evaluating the model by dividing the original sample into small k-chunks. The partition process of the original data in k-chunks used a random approach but the size is always equal. In k-fold, k-1 chunks are used for training the model, and the remaining single chunk is used to test the model. The Gini index is calculated for accurate measurement. Training and testing of data are completed with a ratio of 70 and 30. We can also train our proposed model by using meta-heuristic optimization techniques [45].

Proposed IMDS Algorithm:

```
from random import randrange
from csv import reader
from math import sqrt
def load_csv(filename):
def str_column_to_float(dataset, column):
def str_column_to_int(dataset, column):
def cross_validation_split(dataset, n_folds):
dataset_split = list()
dataset_copy = list(dataset)
fold_size = int(len(dataset) / n_folds)
def test_split(index, value, dataset):
    binwidth = int((max(df["survival_score"])-
    min(df["survival_score"]))/3)bins=range(min(df["survival_
    score"]),max(df["survival_score"],binwidth)
    group_name= ['normal','low_risk','high_risk']
def gini_index(groups, classes):
gini += (1.0 - score) * (size / n_instances)
return gini
def build_tree(train, max_depth, min_size, n_features):
root = get_split(train, n_features)
split(root, max_depth, min_size, n_features, 1)
return root
def predict(node, row):
if row[node['index']] < node['value']:
                    if isinstance(node['left'], dict):
                                return predict(node['left'], row)
                    else:       return node['left']
else:  if is instance(node['right'], dict):
                                return     predict(node['right'],
row)
                    else:  return node['right']
def random_forest(train, test, max_depth, min_size, sample_size,
n_trees, n_features):
trees = list()
for i in range(n_trees):
                    sample = subsample(train, sample_size)
                    tree = build_tree(sample, max_depth,
min_size, n_features)
                    trees.append(tree)
predictions = [bagging_predict(trees, row) for row in test]
return(predictions)
filename = 'sonar.all-data.csv'
dataset = load_csv(filename)
for i in range(0, len(dataset[0])-1):
                    str_column_to_float(dataset, i)
    str_column_to_int(dataset, len(dataset[0])-1)
    n_folds = 10, max_depth = 10, min_size = 1, sample_size = 1.0
    n_features = int(sqrt(len(dataset[0])-1))
for n_trees in [1, 10, 10]:
                    scores = evaluate_
algorithm(dataset, random_forest, n_folds,
    max_depth, min_size, sample_size, n_trees, n_features)
print('Mean Accuracy: %.3f%%' % (sum(scores)/float(len(scores))))
```

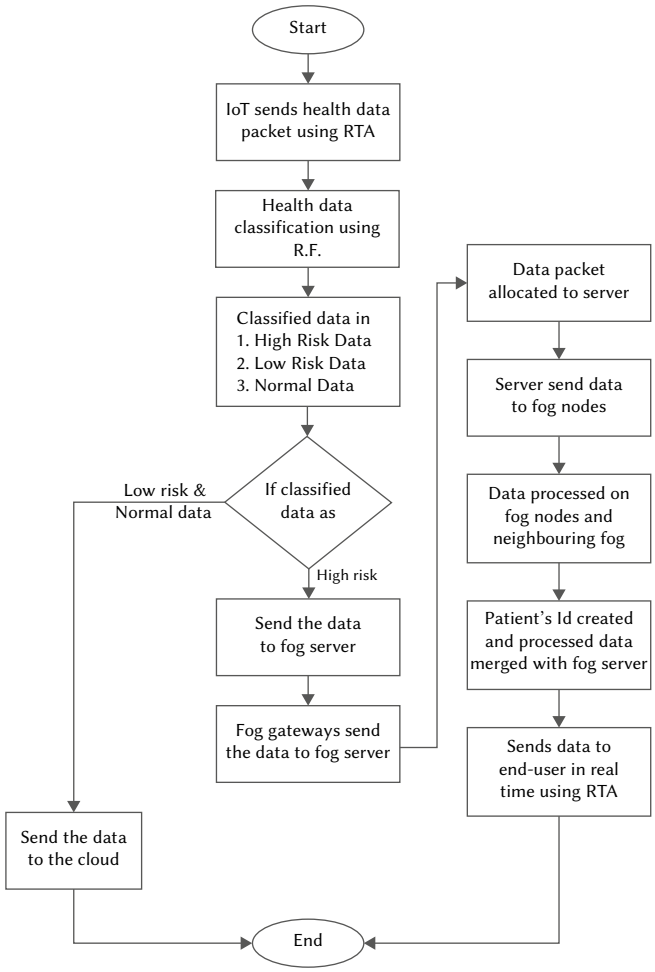


Fig. 2. Flow chart for IMDS scheme.

A. A mathematical Framework For Latency Calculation

Here we assume that there are different kinds of sensors used to forwarding their health data to an FN. In this research, for implementation of fog computing, we are concern with finding the low latency in health informatics. Considering a fog network, containing a cloud layer, a fog layer, and a sensor layer, here the data will be transferred regularly between all the tiers. The sensor layer is containing smart in nature and small in size IoT devices and they don't have enough capability of computations. The fog networks are placed closer to IoT devices to process the patient's data. It is considered that different kinds of sensor devices send their data to a FN (i) and the data size will be XP packets/second. FN (i) performs the task of controlling, storing, analyzing, and processing the health data received from sensors. Here the FNs (i) cooperate with other adjacent or neighboring nodes. After receiving XP packets of the task at FN (i) from the end-user node (e) then the FN distributes the task to adjacent or neighboring FNs (j) for computation. After computation the task is returned to the main FN. Here, the transmission delay for the request of FN and response time from neighboring FN is calculated as the transmission delay.

Transmission delay (T_D): Transmission delay (T_D) is the round trip time (RTT) in relaying of data fragment between end-users nodes (wearable sensors) to FNs can be calculated by M/D/1 system as follows

$$T_D = FN_{RT} + FN_{RPT} \quad (1)$$

Transmission delay between end-user node (e) to FN (i) is as follows

FN request (FN_{RT}) from e to i is,

$$T_{ei} = \frac{\lambda_{ei}}{2\mu_{ei}(\mu_{ei}-\lambda_{ei})} + \frac{1}{\mu_{ei}} \quad (2)$$

FN response time (FN_{RPT}) from i to e is,

$$T_{ie} = \frac{\lambda_{ie}}{2\mu_{ie}(\mu_{ie}-\lambda_{ie})} + \frac{1}{\mu_{ie}} \quad (3)$$

Where

$$\mu_{ei} = B_e \log_2\left(1 + \frac{g_{ei}P_{tx,e}}{B_e N_o^e}\right) \quad (4)$$

and $g_{ei} = \gamma_1 d_{ei}^{-\gamma_2}$

$$\mu_{ie} = B_i \log_2\left(1 + \frac{g_{ie}P_{tx,i}}{B_i N_o^i}\right) \quad (5)$$

and $g_{ie} = \gamma_3 d_{ie}^{-\gamma_4}$

Hence the transmission delay is as

$$T_{D1} = T_{ei} + T_{ie}$$
$$= \frac{\lambda_{ei}}{2\mu_{ei}(\mu_{ei}-\lambda_{ei})} + \frac{1}{\mu_{ei}} + \frac{\lambda_{ie}}{2\mu_{ie}(\mu_{ie}-\lambda_{ie})} + \frac{1}{\mu_{ie}}$$
$$= \frac{\lambda_{ei}}{2\mu_{ei}(\mu_{ei}-\lambda_{ei})} + \frac{\lambda_{ie}}{2\mu_{ie}(\mu_{ie}-\lambda_{ie})} + \frac{1}{\mu_{ei}} + \frac{1}{\mu_{ie}} \quad (6)$$

T_D between FN (i) and neighboring node (j) can be expressed as follows

FN request (FN_{RT}) from i to j is,

$$T_{ij} = \frac{\lambda_{ij}}{2\mu_{ij}(\mu_{ij}-\lambda_{ij})} + \frac{1}{\mu_{ij}} \quad (7)$$

FN response time (FN_{RPT}) from j to i is,

$$T_{ji} = \frac{\lambda_{ji}}{2\mu_{ji}(\mu_{ji}-\lambda_{ji})} + \frac{1}{\mu_{ji}} \quad (8)$$

$$\text{Where } \mu_{ij} = B_j \log_2\left(1 + \frac{g_{ji}P_{tx,j}}{B_j N_o^j}\right) \quad (9)$$

And $g_{ij} = \gamma_5 d_{ij}^{-\gamma_6}$

$$\mu_{ji} = B_i \log_2\left(1 + \frac{g_{ij}P_{tx,i}}{B_i N_o^i}\right) \quad (10)$$

And $g_{ji} = \gamma_7 d_{ji}^{-\gamma_8}$

Hence the transmission delay is as

$$T_{D2} = \frac{\lambda_{ij}}{2\mu_{ij}(\mu_{ij}-\lambda_{ij})} + \frac{\lambda_{ji}}{2\mu_{ji}(\mu_{ji}-\lambda_{ji})} + \frac{1}{\mu_{ji}} + \frac{1}{\mu_{ij}} \quad (11)$$

Transmission delay between end user node (e) and neighboring FN (j) is

FN request (FN_{RT}) from i to j is

$$T_{ej} = \frac{\lambda_{ej}}{2\mu_{ej}(\mu_{ej}-\lambda_{ej})} + \frac{1}{\mu_{ej}} \quad (12)$$

FN response time (FN_{RPT}) from i to e is,

$$T_{je} = \frac{\lambda_{je}}{2\mu_{je}(\mu_{je}-\lambda_{je})} + \frac{1}{\mu_{je}} \quad (13)$$

$$\text{Where, } \mu_{je} = B_j \log_2\left(1 + \frac{g_{je}P_{tx,j}}{B_j N_o^j}\right) \quad (14)$$

And $g_{ej} = \gamma_9 d_{ej}^{-\beta\gamma_{10}}$

$$\mu_{ej} = B_e \log_2\left(1 + \frac{g_{ej}P_{tx,e}}{B_e N_o^e}\right) \quad (15)$$

$$g_{je} = \gamma_{11} d_{je}^{-\gamma_{12}}$$

The transmission delay between the end node (e) and the neighboring FN (j) is

$$T_{D3} = \frac{\lambda_{ej}}{2\mu_{ej}(\mu_{ej}-\lambda_{ej})} + \frac{\lambda_{je}}{2\mu_{je}(\mu_{je}-\lambda_{je})} + \frac{1}{\mu_{ej}} + \frac{1}{\mu_{je}} \quad (16)$$

Hence the total transmission delay will be

$$T_D = T_{D1} + T_{D2} + T_{D3}$$
$$= \frac{\lambda_{ei}}{2\mu_{ei}(\mu_{ei}-\lambda_{ei})} + \frac{\lambda_{ie}}{2\mu_{ie}(\mu_{ie}-\lambda_{ie})} + \frac{1}{\mu_{ei}} + \frac{1}{\mu_{ie}} + \frac{\lambda_{ij}}{2\mu_{ij}(\mu_{ij}-\lambda_{ij})} + \frac{\lambda_{ji}}{2\mu_{ji}(\mu_{ji}-\lambda_{ji})} + \frac{1}{\mu_{ji}} + \frac{1}{\mu_{ij}} + \frac{\lambda_{ej}}{2\mu_{ej}(\mu_{ej}-\lambda_{ej})} + \frac{\lambda_{je}}{2\mu_{je}(\mu_{je}-\lambda_{je})} + \frac{1}{\mu_{ej}} + \frac{1}{\mu_{je}}$$
$$= \frac{\lambda_{ei}}{2\mu_{ei}(\mu_{ei}-\lambda_{ei})} + \frac{\lambda_{ie}}{2\mu_{ie}(\mu_{ie}-\lambda_{ie})} + \frac{\lambda_{ij}}{2\mu_{ij}(\mu_{ij}-\lambda_{ij})} + \frac{\lambda_{ji}}{2\mu_{ji}(\mu_{ji}-\lambda_{ji})} + \frac{\lambda_{ej}}{2\mu_{ej}(\mu_{ej}-\lambda_{ej})} + \frac{\lambda_{je}}{2\mu_{je}(\mu_{je}-\lambda_{je})} + \frac{1}{\mu_{ei}} + \frac{1}{\mu_{ie}} + \frac{1}{\mu_{ji}} + \frac{1}{\mu_{ij}} + \frac{1}{\mu_{ej}} + \frac{1}{\mu_{je}} \quad (17)$$

Here, μ_{ei} is the transmission service rate from the IoT device e to FN i, μ_{ie} is the transmission service rate from the FN i to the IoT device e, μ_{je} is the transmission service rate from neighbouring FN j to e, μ_{ej} is the transmission service rate from the IoT device e to node j, μ_{ij} is the transmission service rate from the FN i to the node j, T_{ei} is the T_D from IoT device e to the FN i, T_{ie} is the T_D from the FN i to IoT device e, T_{ej} is the T_D from the IoT device e to the neighbouring FN j, T_{je} is the T_D from the neighbouring FN j to the IoT device e, T_{ji} is the T_D from the neighbouring FN j to the FN i, T_{ij} is the T_D from the FN i to neighbouring FN j, g_{ei}, g_{ie}, g_{je}, g_{ej}, g_{ij} and g_{ji} are the channel gains for μ_{ei}, μ_{ie}, μ_{je}, μ_{ej}, μ_{ij}, and μ_{ji}, B_e, B_i, and B_j are the bandwidth for the IoT device e , for node i, and for node j, $\gamma_1, \gamma_3, \gamma_5, \gamma_7, \gamma_9$ and γ_{11} are the path loss exponent, $\gamma_2, \gamma_4, \gamma_6, \gamma_8, \gamma_{10}$ and γ_{12} are the Path loss constant, $P_{tx,e}$, $P_{tx,i}$ and $P_{tx,j}$ are the transmission power for node e , node i, and node j, d_{ei}, d_{ie}, d_{ej}, d_{je}, d_{ij}, and d_{ji} are the distance between e and i, i and e, e and j, j and e, i and j, and j and i, N_o^e , N_o^i , and N_o^j are the noise densities from nodes e to i and j, i to j and e, and j to i and e, λ_{ei} and λ_{ie} are arrival rates of task data from node e to node i, and from node i to node e, λ_{ej} and λ_{je} are the arrival rate of task data from node e to node j, and from node j to node e, λ_{ij} and λ_{ji} are the arrival rate of task data from node i to node j, and from node j to node i

Network Delay (N_D): Networks delay (N_D) incurred the delay which depends upon the total number of packets from the sensor network to fog network and fog network to sensor network. Network delay depends upon every hop delay as well as total packet sent from the end-user node e to FN i, FN i to neighboring node j, and from FN j to end-user node e and also assuming that there is equal latency for each hop delay. The network delay is calculated as:

N_D = N_D from node e to i + N_D from node i to j + N_D from node j to e

$$= \frac{d_\propto h_c\, e}{X_P} + \frac{d_\propto h_c\, i}{X_P} + \frac{d_\propto h_c\, j}{X_P} = \frac{d_\propto h_c}{X_P}(e + i + j) \quad (18)$$

Where h_c and $d_\propto$ are the hop count and hop delay.

Computation Delay (C_D): When task computation is done by FN, there exists a waiting queue in the task computation queue due to the prior task available in the queue for processing. The neighbouring FNs are not just receiving the task from a single source node they receive it from multiple nodes and also from end-users. Hence, the computation queue can be computed as an M/D/1 system, neglecting the loss of packets, with the task data arrival rate and the computation latency of

FNs that can be expressed as

$$C_{ei} = \frac{\lambda_{ei}}{2\mu_i(\mu_i-\lambda_{ei})} + \frac{1}{\mu_i} + \frac{\lambda_{ei}}{C_s} \quad (19)$$

$$C_{ij} = \frac{\lambda_{ij}}{2\mu_j(\mu_j-\lambda_{ij})} + \frac{1}{\mu_j} + \frac{\lambda_{ij}}{C_s'} \quad (20)$$

Total computation delay can be calculated as

$$C_D = C_{ei} + C_{ij} = \frac{\lambda_{ei}}{2\mu_i(\mu_i-\lambda_{ei})} + \frac{1}{\mu_i} + \frac{\lambda_{ei}}{C_s} + \frac{\lambda_{ij}}{2\mu_j(\mu_j-\lambda_{ij})} + \frac{1}{\mu_j} + \frac{\lambda_{ij}}{C_s'}$$

$$= \frac{\lambda_{ei}}{2\mu_i(\mu_i-\lambda_{ei})} + \frac{\lambda_{ij}}{2\mu_j(\mu_j-\lambda_{ij})} + \frac{1}{\mu_i} + \frac{1}{\mu_j} + \frac{\lambda_{ei}}{C_s} + \frac{\lambda_{ij}}{C_s'} \quad (21)$$

Where μ_i and μ_j are the hardware parameter at node i and node j, C_s and C_s' are the speeds of CPU at node i and node j.

Here, the first term used as a waiting time in the computation queue, and the second term is used as delay occurred for tracking the proper application used for task computation. The tracking delay depends upon the quality of the hardware used.

The total latency (T_L) or total delay time can be calculated as the sum of transmission delay, network delay, and computation delay

$$T_L = T_D + N_D + C_D$$

$$= \frac{\lambda_{ei}}{2\mu_{ei}(\mu_{ei}-\lambda_{ei})} + \frac{\lambda_{ie}}{2\mu_{ie}(\mu_{ie}-\lambda_{ie})} + \frac{\lambda_{ij}}{2\mu_{ij}(\mu_{ij}-\lambda_{ij})} + \frac{\lambda_{ji}}{2\mu_{ji}(\mu_{ji}-\lambda_{ji})} + \frac{\lambda_{ej}}{2\mu_{ej}(\mu_{ej}-\lambda_{ej})} + \frac{\lambda_{je}}{2\mu_{je}(\mu_{je}-\lambda_{je})} + \frac{1}{\mu_{ei}} + \frac{1}{\mu_{ie}} + \frac{1}{\mu_{ji}} + \frac{1}{\mu_{ij}} + \frac{1}{\mu_{ej}} + \frac{1}{\mu_{je}} + \frac{h_c}{X_P}(d_e\, e + d_i\, i + d_j\, j) + \frac{\lambda_{ei}}{2\mu_i(\mu_i-\lambda_{ei})} + \frac{\lambda_{ij}}{2\mu_j(\mu_j-\lambda_{ij})} + \frac{1}{\mu_i} + \frac{1}{\mu_j} + \frac{\lambda_{ei}}{C_s} + \frac{\lambda_{ij}}{C_s'} \quad (22)$$

V. Results and Discussion

In this section, we discussed the performance of the model. In this model, data is transferred from one layer to another layer started from IoT devices and reaches to cloud through a fog environment. The time consumed by data in travelling is calculated. As data is classified, it is processed and as per requirement, the data is sent to the end-user or cloud. To complete the research task, we use the tool of python editor. The result will be visualized after the completion of the simulation process. Here, the data set is divided into tenfold as we applied a k-fold random forest learning algorithm. 70% of the data set will use for training purposes whereas 30% of data used for testing purposes. Python 3.7 is used as a platform for implementing this work. Random forest algorithm classifies the data in high risk, low risk, and normal with the accuracy of 92% in the proposed work. It has taken 14 seconds as computation time.

For the simulation, we performed several tests for monitoring devices with five different configurations. The evaluation of latency, usage of the network, and consumption of RAM were performed by the simulations. The ifogsim [11] simulator is used to simulate the fog network and nodes. The evaluation of the transmission delay, computation delay, and network delay is simulated through the ifogsim [11] simulator. This simulator creates the physical topologies and they are programmed with Java API. JSON file format is used to store the updated and modified topologies. By varying the size of topology, the performance of simulation is evaluated.

The FNs are swapping the data packets among the system entities during the simulation. The wi-fi connection is established between Fog and IoT devices. In the process of testing the performance, different topologies parameters are used concerning the different number of fog and IoT devices. IoT-sensor, FNs, and cloud data centers as servers are used as physical topology parameters in the simulating tool. By varying the size of topology, the performance of simulation was evaluated.

All configurations (number five) such as configure.1, configure.2, configure.3, configure.4, and configure.5 are simulated with physical topologies on the simulated tool. This system is generated for the performance analysis of proposed work in the fog computing environment. The IoT_sensor device has 1200 million instructions as a CPU length, a network-length of 21000 bytes, and inter-arrival time (average) at data packet arrival of 20 ms.

The details of the used fog device, IoT-sensor, and link of the network are shown in Table VII and Table VIII.

TABLE VII. Details of Fog Device Parameters

Type of device	Processing speed (G.Hz)	Ram (MegaBytes)
Fog-device	2.60	2.0
Cloud-server	4.0	4.0

TABLE VIII. Details of Network Link Parameters

Source node	Destination node	Latency (ms)
IoT_sensor1	Fog-device	45.0
IoT_sensor2	Fog-device	45.0
IoT_sensor3	Fog-device	50.0
Fog-device	Cloud-server	75.0

A comparison in transmission delay between fog and cloud environment is shown in Fig. 3. First of all, a link (tuple) is generated by IoT-sensors, and the connection is established with available routers, gateways, and connected FNs. After getting the data packet on fog servers, processing and distributing to other neighbouring FNs, then data packets are received by the end-user.

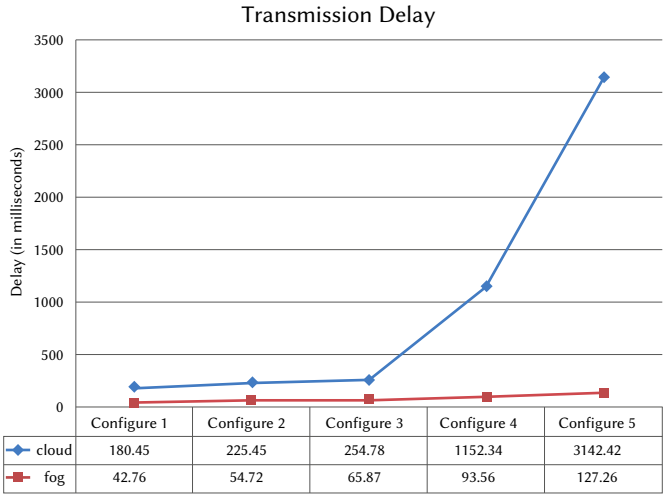
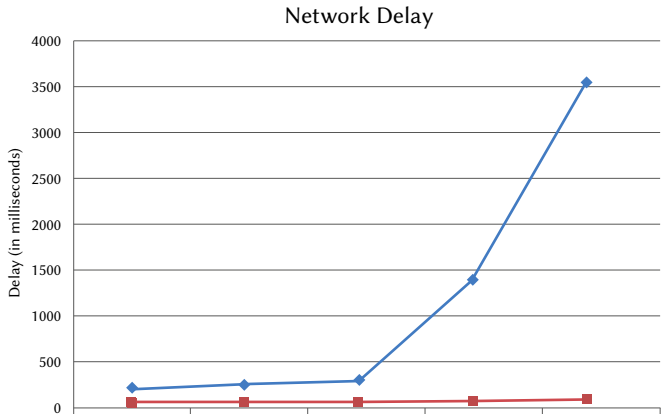


Fig. 3. Fog computing and cloud computing comparison for T_D.

The comparison of network delay between fog and cloud computing is shown in Fig. 4. When the transmission of data occurs between

IoT-sensors and fog servers, the hop counts decreases. Fig. 4. shows the reduction in network latency. When there is a large volume of data transmitted between IoT-sensors and cloud servers, there is an important increase in network latency for the cloud network while this is kept low for the fog network.

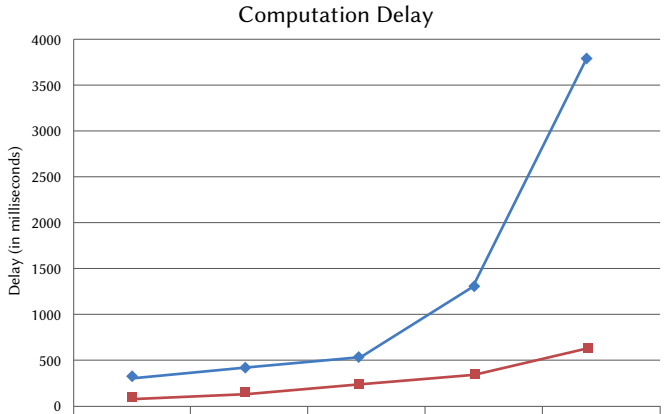
The comparison between fog and cloud computing computation delay is shown in Fig. 5. When task data reaches to FNs, it starts computing the data, and the computation depends upon the parameters such as the speed of the processing unit, hardware performance, and size of the data packet.



	Configure 1	Configure 2	Configure 3	Configure 4	Configure 5
cloud	225.18	247.77	292.46	1390.25	3554.95
fog	58.23	63.46	67.45	70.28	94.25

Fig. 4. Fog computing and cloud computing comparison for N_D.

Fig. 6 shows the consumption of usage of networks in fog and cloud computing environments. FNs are deployed over certain regions to overcome network congestion.



	Configure 1	Configure 2	Configure 3	Configure 4	Configure 5
cloud	295.28	415.56	520.89	1295.76	3805.15
fog	68.12	115.35	225.25	335.37	625.34

Fig. 5. Fog computing and cloud computing comparison for C_D.

In these simulation results, the various physical topology configurations are used in the fog computing environment. As a result, the average result of transmission delay is 76.834 ms, the average result of network delay is 70.734 ms, and 273.886 ms for average computation delay. The usage of the network is also minimized with the average result is kilo bytes. The existing state-of-art is compared with the proposed algorithm that minimized latency by 94-95%. We compared the proposed model by Hermes [23], iFogStor [18], and Hipster [16], where an improvement in the minimization of latency is by 16% with the model presented by Hermes, an 86% reduction in latency is demonstrated as a comparison to cloud computing by iFogStor, and Hipster improves the latency for web-searching by 80-90%. Raafat [32] shows the reduction in overall service latency by 21.9% to 46.6% in the fog environment.

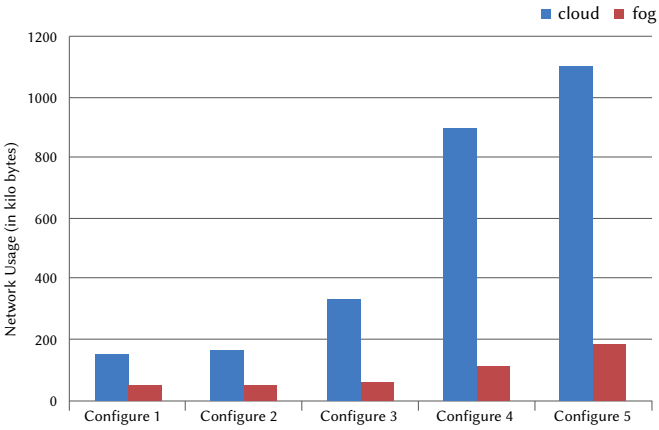


Fig. 6. Fog computing and cloud computing comparison for Network Usage.

VI. Conclusion

Classification of health data and minimization of latency is the most challenging task in e-healthcare, where the fog server is receiving a high volume of task data. Due to the complicated nature of data, fog computing technology becomes essential and important to minimize latency in e-healthcare. In this paper, we presented a novel intelligent multimedia data segregation (IMDS) scheme using machine learning (k-fold random forest) in the fog computing environment. The latency parameters such as transmission delay, network delay, and computation delay are evaluated and it shows the reduction in the high latency. The proposed model is improving the quality of service in e-healthcare and suitable for heterogeneous networks. The latency and usage of the network is a part of QoS. Hence, minimizing the latency and usage of network improves the QoS. In the future, the quality of services in e-healthcare and latency for high-risk data can be improved by using 5G as higher internet connectivity. A smart healthcare system can be implemented in a different hospital through the fog model.

Acknowledgment

The authors would like to thanks to Department of Computer Science & Information Technology, Sam Higginbottom University of Agriculture, Technology and Sciences (SHUATS), Allahabad, U.P., India to give this platform to work. Thanks to all my seniors in the department.

Conflicts of interest/Competing interests – There are no conflicts of interests.

References

[1] C. S. Nandyala and H. K. Kim, "From cloud to fog and IoT-based real-time U-healthcare monitoring for smart homes and hospitals", *International Journal of Smart Home*, vol. 10, no. 2, 2016, pp. 187-196.

[2] S. Cirani, G. Ferrari, N. Iotti, M. Picone M, "The IoT hub: A fog node for seamless management of heterogeneous connected smart objects", 12th Annual IEEE International Conference on Sensing, Communication, and Networking-Workshops (SECON Workshops), 2015.

[3] L.M. Vaquero, L. Rodero-Merino, "Finding your way in the fog: Towards a comprehensive deffinition of fog computing", ACM SIGCOMM

Computer Commun. Review, vol. 44, no. 5, 2014, pp. 27–32.

[4] F. Bonomi, R. Milito, J. Zhu, S. Addepalli S, "Fog computing and its role in the internet of things", in Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, 2012, pp. 13–16.

[5] S. Yi, C. Li C, Q. Li Q, "A survey of fog computing: Concepts, applications and issues", in *Proceedings of the 2 Workshop on Mobile Big Data, ACM,* 2015, pp. 37–42.

[6] P. G. Lopez, A. Montresor, D. Epema, A. Datta, T. Higashino, A. Iamnitchi, M. Barcellos, P. Felber, E. Riviere, "Edge-centric computing: Vision and challenges", SIGCOMM Computer Communication Review, vol. 45, no. 5, 2015, pp. 37–42.

[7] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities", *IEEE Internet of Things Journal*, vol. *3, no.* 6, 2016, pp. 854-864.

[8] T. N. Gia, M. Jiang, A. M. Rahmani, T. Westerlund, P. Liljeberg, H. Tenhunen, "Fog computing in healthcare internet of things: A case study on ecg feature extraction", in *2015 IEEE international conference on computer and information technology; ubiquitous computing and communications; dependable, autonomic and secure computing; pervasive intelligence and computing, IEEE,,* 2015, pp. 356-363.

[9] S.C. Hung, D. Liau, S.Y. Lien, K.C. Chen, "Low latency communication for Internet of Things", in *IEEE/CIC International Conference on Communications in China (ICCC), IEEE, 2015,* pp. 1-6.

[10] G. Lee, W. Saad, M. Bennis, "An online optimization framework for distributed fog network formation with minimal latency", *IEEE Transactions on Wireless Communications*, vol. *18, no.* 4, 2019, pp. 2244-2258.

[11] H. Gupta, A. V. Dastjerdi, S. K. Ghosh, R. Buyya, "iFogSim: A toolkit for modeling and simulation of resource management techniques in the Internet of Things, Edge and Fog computing environments", *Practice and Experience*, vol. *47, no.* 9, 2017, pp.1275-1296.

[12] L. Skorin-Kapov and M. Matijasevic, "Analysis of QoS requirements for e-health services and mapping to evolved packet system QoS classes", *International journal of telemedicine and applications.* 2010.

[13] J. R. Gallego, A. Hernandez-Solana, M. Canales, J. Lafuente, A. Valdovinos, J. Fernandez-Navajas, "Performance analysis of multiplexed medical data transmission for mobile emergency care over the UMTS channel", *IEEE transactions on information technology in biomedicine. Vol. 9, no. 1,* 2005, pp.13–22.

[14] C. A. Silva, G. S. Aquino, S.R.M. Melo, D. J. B. Egidio, "A Fog Computing-Based Architecture for Medical Records Management", *Wireless Communications and Mobile Computing,* 2019.

[15] M. G. R. Alam, Y. K. Tun, C. S. Hong, "Multi-agent and reinforcement learning based code offloading in mobile fog", in *International Conference on Information Networking (ICOIN) , IEEE,* 2016, pp. 285-290.

[16] R. Nishtala, P. Carpenter, V. Petrucci, X. Martorell, "Hipster: Hybrid task manager for latency-critical cloud workloads", in *IEEE International Symposium on High Performance Computer Architecture (HPCA), IEEE* 2017, pp. 409-420.

[17] T. N. Gia, M. Jiang, V. K. Sarkar, A. M. Rahmani, T. Westerlund, P. Liljeberg, H. Tenhunen H, "Low-cost fog-assisted healthcare IoT system with energy-efficient sensor nodes" in Proceedings of *13th international wireless communications and mobile computing conference (IWCMC) , IEEE,* 2017, pp. 1765-1770.

[18] M. I. Naas, P. R. Parvedy, J. Boukhobza, L. Lemarchand, "iFogStor: an IoT data placement strategy for fog infrastructure", in *IEEE 1st International Conference on Fog and Edge Computing (ICFEC),* 2017, pp. 97-104.

[19] A. M. Rahmani, T. N. Gia, B. Negash, A. A. I. Azimi, M. Jiang, P. Liljeberg, " Exploiting smart e-Health gateways at the edge of healthcare Internet-of-Things: A fog computing approach", *Future Generation Computer Systems*, vol. *78*, 2018, pp. 641-658.

[20] J. Wu, M. Dong, K. Ota, J. Li, Z. Guan, "FCSS: Fog computing based content-aware filtering for security services in information centric social networks", *IEEE Transactions on Emerging Topics in computing, 2017.*

[21] A. Brogi and S. Forti, "QoS-aware deployment of IoT applications through the fog", IEEE Internet of Things Journal, vol. 4, no. 5, 2017, pp.1185–1192.

[22] A. Shahzad, Y. S. Lee, M. Lee, Y. G. Kim, N. Xiong, "Real-Time Cloud-Based Health Tracking and Monitoring System in Designed Boundary for Cardiology Patients", *Journal of Sensors*, 2018.

[23] Y. H. Kao, B. Krishnamachari, M. R. Ra, F. Bai, "Hermes: Latency optimal task assignment for resource constrained mobile computing". *IEEE Transactions on Mobile Computing*, vol. *16, no.* 11, 2017, pp. 3056-3069.

[24] G. Li, J. Wu, J. Li, K. Wang, T. Ye, "Service popularity-based smart resources partitioning for fog computing-enabled industrial Internet of Things", *IEEE Transactions on Industrial Informatics, vol. 14, no. 10,* 2018, pp.:4702–4711.

[25] R. Mahmud, F. L. Koch, R. Buyya, "Cloud-fog interoperability in IoT-enabled healthcare solutions", in *Proceedings of the 19th international conference on distributed computing and networking,* 2018, pp. 1-10.

[26] N. T. Dinh and Y. Kim, "An Efficient Availability Guaranteed Deployment Scheme for IoT Service Chains over Fog-Core Cloud Networks", *Sensors. 18(11):3970,* 2018.

[27] M. M. Ahsan, I. Ali, M. Imran, M. Y. I. Idris, S. Khan, A. Khan, "A Fog-centric Secure Cloud Storage Scheme", *IEEE Transactions on Sustainable Computing,* 2019.

[28] H. Rafique, M. A. Shah, S. U. Islam, T. Maqsood, S. Khan, C. Maple, "A Novel Bio-Inspired Hybrid Algorithm (NBIHA) for Efficient Resource Management in Fog Computing", *IEEE Access, vol. 7,* 2019, pp. 115760–115773.

[29] J. Li, T. Zhang, J. Jin, Y. Yang, D. Yuan, L. Gao, "Latency Estimation for Fog-based Internet of Things", in *27th International Telecommunication Networks and Applications Conference (ITNAC), IEEE,* 2017, pp. 1-6. 2017.

[30] C. Thota, R. Sundarasekar, G. Manogaran, R. Varatharajan, M. K. Priyan, " Centralized fog computing security platform for IoT and cloud in healthcare system", in *Fog computing: Breakthroughs in research and practice, IGI global,* 2018, pp. 365-378.

[31] T. Aladwani, "Scheduling IoT Healthcare Tasks in Fog Computing Based on their Importance", in *Proc of 16th International Learning & Technology Conference, Procedia Computer Science*, 2019, pp. 560-569.

[32] O. A. Raafat, A. K. Mazin, L. Taha, E. F. Khaled, "Scheduling Internet of Things Requests to Minimize Latency in Hybrid Fog-Cloud Computing", *Future Generation Computer Systems*, *111*, 2020, pp. 539-551.

[33] J. Pan J and J. McElhannon, "Future edge cloud and edge computing for internet of things applications", IEEE *Internet of Things Journal, vol. 5, no. 1,* 2017. pp. 439–49.

[34] Z. Alam, M. S. Rahman, M. S. Rahman, "A Random Forest based predictor for medical data classification using feature ranking", *Informatics in Medicine Unlocked*, *15*, 100180, 2019.

[35] C. Chakraborty, "Computational approach for chronic wound tissue characterization", *Informatics in Medicine Unlocked*, *17*, 100162, 2019.

[36] C. Chakraborty C and R. Roy, "Markov Decision Process based Optimal Gateway Selection Algorithm" International Journal of Systems, Algorithms & Applications (IJSAA), 2012, pp. 48-52.

[37] C. Chakraborty, R. Roy, S. Pathak, S. Chakrabarti, "An Optimal Probabilistic Traffic Engineering Scheme for Heterogeneous Networks", *Int. Journal of Fuzzy Systems, vol. 3, no. 2,* 2011, pp. 35-39.

[38] M. Nilashi, H. Ahmadi, A. A. Manaf, T. A. Rashid, S. Samad, L. Shahmoradi, A. Nahla, A. Elanz, "Coronary Heart Disease Diagnosis Through Self-Organizing Map and Fuzzy Support Vector Machine with Incremental Updates", *International Journal of Fuzzy Systems*, 2020, pp.1-13.

[39] T.A. Rashid, S. M. Abdullah, R. M. Abdullah, "An intelligent approach for diabetes classification, prediction and description", *Innovations in Bio-Inspired Computing and Applications,* Springer, Cham, 2016, pp. 323-335.

[40] H. Mahmud, M. Mohammadi,, N. Ali, T. A. R. Khan, K. A. S. Nawzad, R. M. D. Omer, L. Joan, " Technologies in medical information processing", *Advances in Telemedicine for Health Monitoring, Technologies, Design and Applications*, 31, 2020.

[41] T. A. Rashid, M. K. Hassan, M. Mohammadi, K. Fraser, "Improvement of variant adaptable LSTM trained with metaheuristic algorithms for healthcare analysis", *Advanced Classification Techniques for Healthcare Analysis*, IGI Global, 2019, pp. 111-131.

[42] J. C. W. Lin, Y. Shao, P. Fournier-Viger, F. Hamido, "BILU-NEMH: A BILU neural-encoded mention hypergraph for mention extraction", *Information Sciences*, *496*, 2020, pp. 53-64.

[43] J. C. W. Lin, Y. Shao, J. Zhang, U. Yun, "Enhanced sequence labeling based on latent variable conditional random fields", *Neurocomputing*, *403*, 2020, pp. 431-440.

[44] U. Ahmed, J. C. W. Lin, G. Srivastava, M. Aleem, "A load balance multi-scheduling model for OpenCL kernel tasks in an integrated cluster", *Soft Computing*, 2020, pp. 1-14.

[45] T. A. Rashid, P. Fattah, D. K. Awla, "Using accuracy measure for improving

the training of LSTM with metaheuristic algorithms", *Procedia Computer Science, 140*, 2018, pp. 324-333.

[46] A. Janosi, M. Pfisterer, W. Steinbrunn, R. Detrano, J. Schmid, S. Sandhu, K. Gupta, S. Lee , V. Froelicher, UCI Machine Learning Repository 2019. https://archive.ics.uci.edu/ml/datasets/heart+disease

Amit Kishor

Er. Amit Kishor pursuing a Ph.D. in Computer Engineering from the Department of Computer Science and Information Technology, in the Faculty of Engineering and Technology, Sam Higginbottom University of Agriculture, Technology and Sciences, Naini, Allahabad, India. He is also working as an Assistant Professor in the Department of Computer Science and Engineering & I.T., Subharti Institute of Engineering and Technology, Swami Vivekanand Subharti University, Meerut. India His area of interest is cloud computing, Algorithms, Artificial Intelligence, and Data Structures. He has published more than 20 papers in reputed international journals. He is also a member of an International body of International Association of Engineers (IEANG).

Chinmay Chakraborty

Dr. Chinmay Chakraborty is working as an Assistant Professor (Sr.) in the Dept. of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, India. He worked at the Faculty of Science and Technology, ICFAI University, Agartala, Tripura, India as a Sr. lecturer. He worked as a Research Consultant in the Coal India project at Industrial Engineering & Management, IIT Kharagpur. He worked as a project coordinator of the Telecommunication Convergence Switch project under the Indo-US joint initiative. He also worked as a Network Engineer in System Administration at MISPL, India. His main research interests include the Internet of Medical Things, Wireless Body Area Network, Wireless Networks, Telemedicine, m-Health/e-health, and Medical Imaging. Dr. Chakraborty has published sixty papers at reputed international journals, conferences, book chapters, and books. He is an Editorial Board Member in the different Journals and Conferences. He is serving as a Guest Editor of MDPI-Future Internet Journal, Internet Technology Letters, Wiley, and has conducted a session of SoCTA-19, ICICC – 2019, and also a reviewer for international journals including IEEE Access, Elsevier, Springer, Taylor & Francis, IGI, IET, TELKOMNIKA Telecommunication Computing Electronics and Control, and Wiley. Dr. Chakraborty is co-editing Eight books on Smart IoMT, Healthcare Technology, and Sensor Data Analytics with CRC Press, IET, Pan Stanford, and Springer. He has served as a Publicity Chair member at renowned international conferences including IEEE Healthcom, IEEE SP-DLT. Dr. Chakraborty is a member of Internet Society, Machine Intelligence Research Labs, and Institute for Engineering Research and Publication. He received a Young Research Excellence Award, Global Peer Review Award, Young Faculty Award, and Outstanding Researcher Award.

Wilson Jeberson

Prof.(Dr.) Wilson Jeberson was awarded Ph.D. degree in Computer Science and Communication, from Sam Higginbottom Institute of Agriculture, Technology & Sciences, University, Allahabad, India. He has received the MCA in computer Application and MBA in Management from Madurai Kamaraj University Tamilnadu, India. He had worked as Programmer at National Informatics Centre (NIC), Govt. of India, from 1999 to 2000. He also worked as Senior Software Engineer cum DBA at Quintessence Technologies Limited - Trivandrum, Kerala, India, from 2000 to 2002 and as Senior System Analyst at Netcare Technologies-Trivandrum, Kerala, India from 2002 to 2003.Currently he is working as Professor & Head, Department of Computer Science & Information Technology in Sam Higginbottom University of Agriculture, Technology & Sciences. Prayagraj, India from 2003. He has published more than 70 papers in reputed International journals and more than 15 Papers in National & International Proceedings. Some of the best papers are published in Open Computer Science (Formerly: Central European Journal of Computer Science) Published by De Gruyter Open (Formerly: Springer Verlag) titled "Survey of Context Information Fusion for Ubiquitous Internet-of- Things (IoT)", International Journal Elsevier, Science Direct titled "Covert Communication Using Arithmetic Division Operation", :International Journal Elsevier , Science Direct Titled "Fortune at the bottom of the Classifier Pyramid: A novel approach to Human Activity Recognition". A research paper was selected as best one among the top 20 papers presented in the Second GMSARN International Conference at Pattaya, Thailand and was published in International Journal of AIT , Thailand titled "An approach to ICT enabled solution architecture for critical Social Security issues and challenges for e-Governance". He has co-authored two books which are published in Narosa Publications and another in Springer International. He has presented research papers in international conferences in Singapore and Thailand and also attended a training program in Galilee International institute of Management, Israel. He is also member of many national and international bodies including Computer Society of India (CSI) and International Association of Engineers (IAENG), Hong Hong. He was the Principal investigator for two funded projects one funded by Computer Society of India Education Directorate, Chennai-India and another by SHIATS. He is one of the editorial board members of various journals including Scientific & Academic Publishing Co., Journal Name: Software Engineering, USA.

An Enhanced Texture-Based Feature Extraction Approach for Classification of Biomedical Images of CT-Scan of Lungs

Varun Srivastava[1], Shilpa Gupta[1], Gopal Chaudhary[1], Arun Balodi[2], Manju Khari[3], Vicente García-Díaz[4]*

[1] Bharati Vidyapeeth's College of Engineering, Paschim Vihar, New Delhi (India)
[2] Atria Institute of Technology, Bengaluru, Karnataka (India)
[3] Netaji Subhas University of Technology, East Campus, Delhi (India)
[4] Department of Computer Science, University of Oviedo, Oviedo (Spain)

Received 14 August 2020 | Accepted 9 October 2020 | Published 2 November 2020

Abstract

Content Based Image Retrieval (CBIR) techniques based on texture have gained a lot of popularity in recent times. In the proposed work, a feature vector is obtained by concatenation of features extracted from local mesh peak valley edge pattern (LMePVEP) technique; a dynamic threshold based local mesh ternary pattern technique and texture of the image in five different directions. The concatenated feature vector is then used to classify images of two datasets viz. Emphysema dataset and Early Lung Cancer Action Program (ELCAP) lung database. The proposed framework has improved the accuracy by 12.56%, 9.71% and 7.01% in average for data set 1 and 9.37%, 8.99% and 7.63% in average for dataset 2 over three popular algorithms used for image retrieval.

Keywords

Image Retrieval, Local Mesh Peak Valley Edge Patterns, Local Patterns, Bio-medical Image Classification, Texture-based Retrieval.

DOI: 10.9781/ijimai.2020.11.003

I. Introduction

Content based image retrieval (CBIR) relies upon the extraction of features present in an image to fetch related images from a database. Out of the multiple techniques of CBIR based feature extraction, texture-based features are one of the most popular and earliest [1]. Various algorithms like local binary pattern (LBP), local ternary pattern (LTP), local mesh patterns (LMeP), etc. have been proposed to extract texture from a given image for CBIR.

Also, multiple image retrieval techniques based on texture information have been proposed in the literature [2]-[11]. A comparison of state-of-the-art biomedical image retrieval techniques has been given in [5]. The texture-based feature vector has been introduced by the concept of local binary patterns (LBPs). LBP compares the intensity of central pixel to its P neighbours at radius R [12]. If the intensity of neighbour is greater than central pixel then feature vector is assigned a value 1 or otherwise 0. This constitutes a feature vector of size P which is then multiplied by a weight vector and finally summed up to form the intensity value at that pixel. Similarly, the technique of local ternary pattern is introduced where instead of one threshold (central pixel intensity); two thresholds are used [13]. These techniques are modified by various researchers to propose other texture-based approaches. If the central pixel intensity is compared to neighbours in different directions, we retrieve local mesh ternary patterns [2] or if the edge information is incorporated into the texture information, then local maximum edge binary patterns are formed [11].

Recently, researchers have tried to merge texture-based features with other features like frequency-based features, tags, shape based features etc. as well. In this way average retrieval rate (ARR) and average retrieval precision (ARP) are improved to a great deal for medical image retrieval. Authors in [14] presented a retrieval approach in which colour, texture and edge information are used as features. Further Principal Component Analysis (PCA) is applied to reduce these features which are then fed into a support vector machine (SVM) and a fuzzy classifier model (FCM) for statistical similarity measurement. Image retrieval techniques are also very popular for biomedical datasets. Some of them are compared in [15]. Diabetic retinopathy is performed in [16] by calculating LBP for different planes. Edges of an object in the image are detected in red, green and blue planes and the corresponding LBP relationships are extracted for each plane. These LBP feature vectors are then combined to generate the final feature set. In [17], textual and shape-based features are extracted and combined to form the final feature matrix. Authors in [18] generated a visual model of an input biomedical image. A SVM model is generated for classification of that visual model. The visual and semantic similarities are then used to form the feature vector. A combination of text and image texture to form the feature vector for image retrieval has also been proposed in the literature. The associated tags (text) with the images are used along with the texture to extract images in [19], [20] and [21]. Texture-based algorithms have also been extended in many ways. Authors in [22] used RGB components along with luminescence and chrominance components whose texture pattern has been computed.

* Corresponding author.
E-mail address: garciavicente@uniovi.es

A histogram of these texture patterns is then used as a feature vector. In [23], the centre symmetric local binary patterns are calculated and are then combined with co-occurrence matrix calculated in 0, 45, 90 and 135 degrees. The feature vector obtained thereby is used for image retrieval. A robust approach is proposed for biomedical image retrieval in [24] using Zernike moments. Spatial and wavelet-based features are combined in [25] to form the feature vector for extraction of similar mammograms. Texture in 0, 45, 90, 135 and 180 degrees are extracted from an image in [26]. Further the technique of local ternary patterns (LTP) is applied in these directions to form the feature vector. A second-order image moments in the local neighbourhood is computed in [27]. A covariance matrix is then computed followed by Eigen value decomposition. The principal Eigen value is then normalized using gray value of the center pixel. A refined histogram is then obtained for image retrieval. A texture block coding-based tree data structure is proposed in [28] for effective image retrieval. The precision, recall and *F* score were compared and were better as compared to the similar algorithms. Instead of a 2D 3*3 block, a 3*3*3 3D block is used in [29] to extract 3D ternary patterns. These 3D ternary patterns are then used for image retrieval. In [34], authors extracted feature vector by first decomposing an original image using wavelet decomposition for five consecutive times and then used probabilistic principal component analysis to reduce its features. The reduced set of features is classified using feed forward neural network or k nearest neighbour algorithm. A brief overview of various types of cancers and a summary of recent work to identify the extent of possible damage in the corresponding tissue has been discussed in [35].

Recent techniques involve the use of deep convolutional networks for extraction of similar images from biomedical dataset. Few such algorithms are given in [30] and [31] where deep convolutional features are extracted to retrieve similar images. Authors in [32] used wavelet decomposition along with principal component analysis to classify various MR images of brain. In [33], authors proposed a local mesh peak valley edge pattern feature vector to extract similar images from a given image dataset. The algorithm computed j^{th} order derivatives in forward and backward directions to compute peak and valley texture from an image.

The proposed algorithm is a hybrid of two very popular image retrieval approaches viz. local mesh peak valley edge pattern (LMePVEP) and local mesh ternary pattern (LMeTerP) to retrieve similar images. A new feature set obtained by computing texture in different directions is introduced in this paper and combined with the aforementioned two feature sets. The concatenation of feature set has yielded a very comprehensive feature vector with much more accuracy as compared to LMePVEP and LMeTerP. Thereby the proposed algorithm is found better than two recent biomedical image retrieval algorithms for classification of diseased and healthy patient's images. The paper has been organized in the following manner. Section II discusses the various steps of the proposed methodology. Section III summarizes the results and compares the given algorithm with three other algorithms used for image retrieval. Section IV presents the conclusion for the given work.

II. Materials and Methods

A. Local Mesh Peak Valley Edge Patterns

Local mesh peak valley edge pattern (LMePVEP) as given in [33] computes the texture details from a given image in terms of peak and valleys. It compares the central pixel intensity with *P* neighbors at radius *R* and calculates the first order gradient in both forward and backward direction. If $I^f_{P,R}$ represents a first order forward derivative vector, then for central pixel with intensity $I(g_c)$, the j^{th} order gradient in forward direction can be computed using eq. (1).

$$I^f_{P,R}(g_c, g_i) = I(g(\alpha)) - I(g(i)) \quad i = 1,2,\ldots.P \tag{1}$$

$$\text{where } \alpha = 1 + \text{mod}((i + P + j - 1), P) \tag{2}$$

and j= 1, 2 …. P/2, which represents the distance for first order derivation

Similarly, j^{th} order backward derivative $I^b_{P,R}$, can be defined as

$$I^b_{P,R}(g_c,g_i) = \begin{cases} I(g(P + i - j)) - Ig(i) & if\ j \geq i \\ I(g(i - j)) - Ig(i) & otherwise \end{cases} \tag{3}$$

where values of i and j vary as in eq. (1) and eq. (2) respectively.

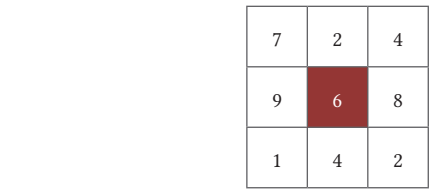
Local mesh peak valley edge pattern LMePVEP (P, R) is then calculated using eq. (4) in the form of a column vector of size *1*P.*

$$\text{LMePVEP}_{(P,R)} = \begin{vmatrix} f_2(I^f_{P,R}(g_c,g_1), & I^b_{P,R}(g_c,g_1)), \\ f_2(I^f_{P,R}(g_c,g_2), & I^b_{P,R}(g_c,g_2)), \\ \ldots\ldots\ldots\ldots & \ldots\ldots\ldots\ldots \\ f_2((I^f_{P,R}(g_c,g_P)), & I^b_{P,R}(g_c,g_P)) \end{vmatrix} \tag{4}$$

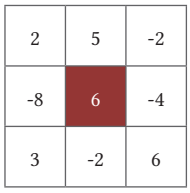
where the function $f_2(x, y)$ used in eq. (4) is defined as:

$$f_2(x, y) = \begin{cases} 1 & \text{if } x > 0 \text{ and } y > 0 \\ 2 & \text{if } x < 1 \text{ and } y < 1 \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

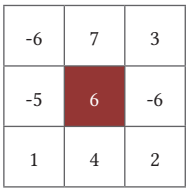
Therefore, the possible values of all components of LMePVEP vector can be 0, 1 or 2. Fig. 1 represents an illustration for computation of Combined LMePVEP values. The illustration considers a 3x3 block of original image and the LMePVEP values for 8 neighbors at radius R as 1 are computed thereby.



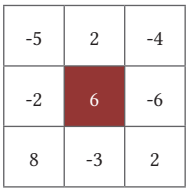
(a)



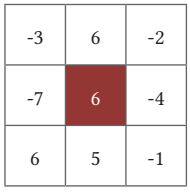
(b)



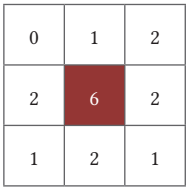
(c)



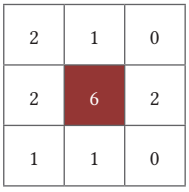
(d)



(e)



(f)



(g)

Fig. 1. Calculation of LMePVEP patterns (a) Original 3x3 image block. (b) Forward first order derivative of the original image with j=1 (c) Forward first order derivative of the original image with j=2 (d) Backward first order derivative of the original image with j=1 (e) Backward first order derivative of the original image with j=2 (f) Combined LMePVEP values for j=1 (g) Combined LMePVEP values for j=2.

A 3*3 image block is considered for the extraction of LMePVEP values as shown in Fig. 1(a). Forward first order derivative values

for j=1 and j=2 are computed for the central pixel using eq. (1) and eq. (2) and these derivative values are shown in Fig. 1(b) and Fig. 1(c) respectively. Similarly, the backward first order derivative using for j=1 and j=2 are computed using eq. (3) as shown in Fig. 1(d) andFig. 1(e) respectively. Then Combined LMePVEP coefficients are computed using eq. (5) which are shown in Fig. 1(f) and Fig. 1(g) for j=1 and j=2 respectively. These Combined LMePVEP values are then separated to obtain peak and valley patterns. Here pixels that contain value 2 represent peak and the pixels having values as 1 represent valley. Fig. 2(a) represents a matrix having valley information from Fig. 1(f). Such pixels having valley information (i.e. value 1) are assigned a value 1 and rest all are made zero. Similarly,Fig. 2(c) represents a matrix having peak information only. Matrix in Fig. 2(c) is obtained by considering pixels having value 2 in Fig. 1(f) and making the intensity of the rest of pixelsequal to 0. In a similar manner we obtain the coded information of valley and peak in matrices of Fig. 2(b) and Fig. 2(d) from Fig. 1(g) for j=1 and j=2. A weight matrix W consisting of P values (in this case 8) is generated as shown in Fig. 2(e) using eq. (6) and an inner dot product of weight matrix and peak and valley values obtained in Fig. 2(a) to Fig. 2(d) is calculated. The sum of these individual results is used to replace the central shaded pixel. These values are final weighted LMePVEP values that replace the central pixel to obtain the final LMePVEP texture. This process is repeated for the entire image and all the pixels and finally four images based on weighted LMePVEP values are obtained.

$$W(i)=1.2^{i-1} \quad \text{where } i=1\ldots P \tag{6}$$

Here for each image, we kept the value of j as 1 and 2 to obtain four corresponding images having LMePVEP texture information.

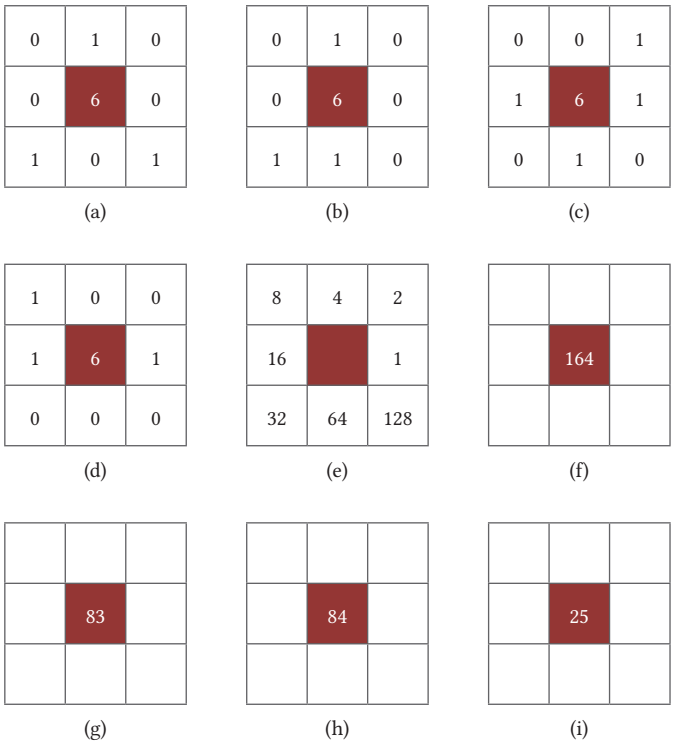
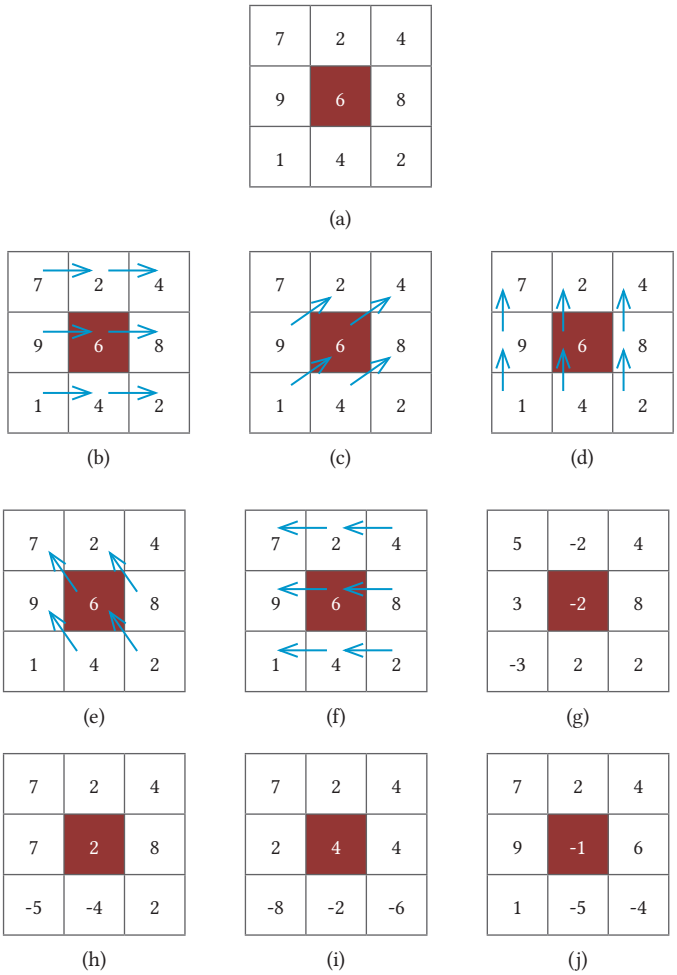


Fig. 2. (a) Local mesh valley pattern extracted from original image for j=1 (b) Local mesh valley pattern for j=2 (c) Local mesh peak pattern for j=1 (d) Local mesh peak pattern for j=2 (e) Weight Matrix (f) Final Weighted LMePVEP value of central pixel for j=1 in forward direction (g)Weighted LMePVEP value of central pixel for j=1 in backward direction (h)Weighted LMePVEP value of central pixel for j=2 in forward direction (i)Weighted LMePVEP value of central pixel for j=2 in backward direction.

B. Texture in Various Directions

Now we compute five different images based on the texture in five different directions for an image block given in Fig. 3.3(a). In [40], the texture in different directions is computed in the images obtained in section II. However, in this paper, we compute texture in the original grayscale image. In the resultant images, one image will contain texture in one direction. To achieve that, for each case, the pixel intensity lying on the tail of the directional arrow is replaced by the difference of pixel intensity lying on the tail and on the head of the directional arrow, and these difference blocks are shown in Fig. 3(g) to Fig. 3(k) for 0, 45, 90, 135 and 180 degrees, respectively. Thereafter, we sum the values in each matrix of Fig. 3(g) to Fig. 3(k) to obtain five matrices as shown in Fig. 3(l). These five central pixel values represent different directional coefficients. The directional coefficients are then binarized by comparing them with a threshold T which is the median value of values present in matrix of Fig. 3(a). For the matrix in Fig. 3(a) the median value is 4 as shown in Fig. 4. This median is taken to remove any noise value if present in the original image. Thus, after comparison if the value is more than 4, then the central pixel will have value as 1 and otherwise it will have a value 0. Thus, the binarized directional coefficients are obtained as shown in Fig. 3(m).

This process is continued for all overlapping 3x3 image blocks which are one pixel distant from each other in horizontal and vertical directions. Ultimately five binary images are obtained.



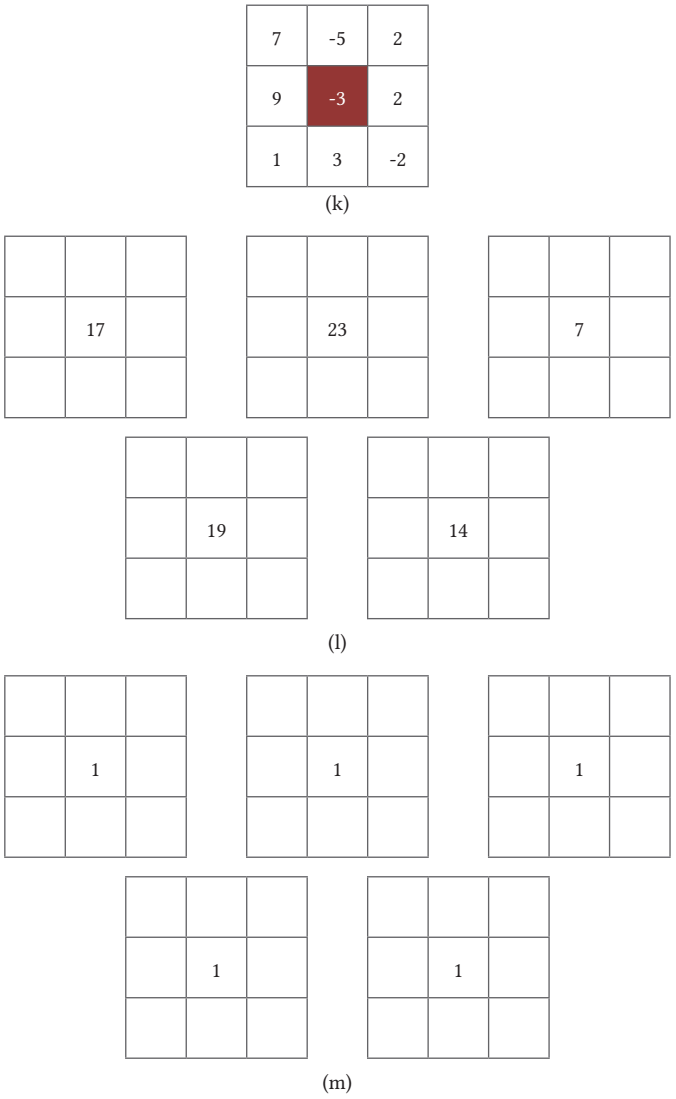


Fig. 3. (a) 3x3 input image block (b) pixel locations used to compute texture gradient in 0 degree (c) pixel locations used to compute texture gradient in 45 degree (d) pixel locations used to compute texture gradient in 90 degree (e) pixel locations used to compute texture gradient in 135 degree (f) pixel locations used to compute texture gradient in 180 degree (g) difference block for 0 degree (h) difference block for 45 degree (i) difference block for 90 degree (j) difference block for 135 degree (k) difference block for 180 degree (l) Summed up values for each matrix of Fig. 3(g) to Fig. 3(k) (m) Central pixel intensities for each direction obtained after comparing with a threshold value of 4 (median) to replace the central pixel intensity of 3x3 matrix in Fig. 3(a) and accordingly obtain five images.

C. Dynamic Threshold Based Local Mesh Ternary Pattern

This technique is derived from [39]. Here we compute the median of nine values in the 3x3 matrix extracted from an image. The median is taken as the threshold value for further computation.

For the computation of this feature vector, firstly Local Mesh Patterns (LMePs) are computed using eq. (7). This calculation is done up to P/2 values of j.

$$LMeP^{j}_{(P,R)} = \sum_{i=1}^{P} f_1(g_{\alpha}|_R - g_i|_R) \quad (7)$$

Here $\alpha = 1 + mod((i+P+j-1),P)$ *for* $j = 1,2, \ldots$ *up to* $P/2$.

Also g represents the intensity value of a pixel, thus $g_{\alpha}|_R$ = intensity value of α^{th} neighbor pixel (out of P neighbors) at radius R and $g_i|_R$= intensity value of the central pixel. Further function f_1 computes the difference between intensity value of $g_{\alpha}|_R$ and $g_i|_R$ pixel.

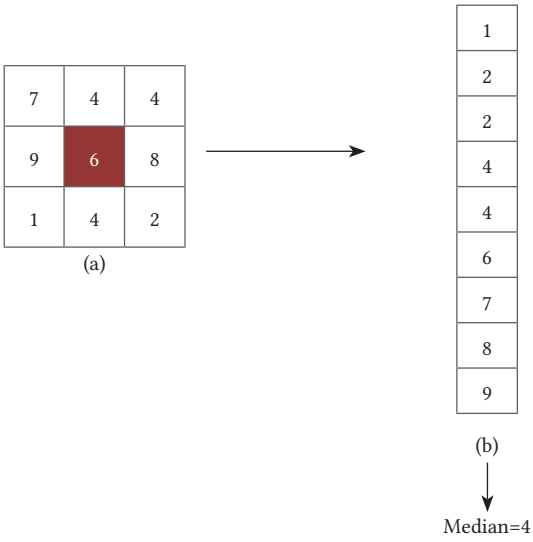


Fig. 4. Threshold computation as per [39] for a 3*3 matrix from an image.

Then, if a matrix sub-component as shown in Fig. 5(a) is considered from an image, then after applying eq. (7), matrix components as given in Fig. 5(b) to Fig. 5(d) are obtained. Further, to compute the upper and lower ternary values, two thresholds are used. In this case, since the median value is 4, the two thresholds as given in [39] are two units greater and lesser than 4. Thus, the two thresholds are taken as 6 and 2.

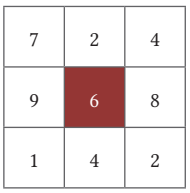
The values obtained in Fig. 5(b), Fig. 5(c) and Fig. 5(d) are then contrasted with thresholds 6 and 2 using eq. (8). If the values in these matrices aregreater than 6, then +1 is the output, if the valuesareless than 2, then -1 is output, and otherwise 0 is the output as given in matrices of Fig. 5(e), Fig. 5(f) and Fig. 5(g).

$$f2(x, med., t) = \begin{cases} +1, & \text{for } x \geq med. \\ 0, & \text{for } |x\text{-}med.| < t \\ -1, & \text{for } x \leq med. - t \end{cases} \quad (8)$$

Now the upper and lower LTPs are separated by separating the positive and negative values in matrices obtained in Fig. 5(e), Fig. 5(f) and Fig. 5(g) to obtain the six matrices of Fig. 5(h). For e.g. the matrix in fig. 5(e) is used to constitute two matrices, one matrix in which only positive values are considered and rest all are kept 0, whereas the other matrix is constituted by keeping the negative values and rest all are made 0. In this way, we obtain the six matrices of Fig. 5(h) from the three matrices of Fig. 5(e), Fig. 5(f) and Fig. 5(g).

Here t =2 and med. (median value) is 4.

These are then multiplied with weight matrix in Fig. 5(i). The multiplied values are then summed up to replace the central pixel intensity of six matrices as shown in Fig. 5(j). This process is repeated for all pixels that are one pixel distant and accordingly we obtain six different images for a single image.



2	5	-2
-8		-4
3	-2	6

(b)

-6	-7	3
-5		-6
1	4	2

(c)

-3	-1	5
-7		-1
7	0	0

(d)

-1	0	-1
-1		-1
0	-1	1

(e)

-1	-1	0
-1		-1
-1	1	-1

(f)

-1	-1	0
-1		-1
1	-1	-1

(g)

0	0	0
0		0
0	0	1

1	0	1
1		1
0	1	0

0	0	0
0		0
0	1	0

1	1	0
1		1
1	0	1

0	0	0
0		0
1	0	0

1	1	0
1		1
0	1	1

(h)

8	4	2
16		1
32	64	128

(i)

	128	

	91	

	64	

	189	

	32	

	221	

(j)

Fig. 5: Calculation of dynamic LMeTerP. (a) Intensity values in a 3*3 sub-part of an image. LMePs for (b) j=1 (c) j=2 (d) j=3. LMeTerP values using function f_2 of eq.(8) for (e) j=1 (f) j=2 (g) j=3 (h) Extraction of Upper and Lower ternary values for Fig. 5(d—f) (i) A standard weight matrix (j) Final values of dynamic LMeTerP after dot product multiplication of values in Fig. 5(h) with values in Fig. 5(i) and then summation.

D. Computation of Final Feature Vector

In subsection A, B and C of section II we obtain four, five and six images respectively for a single image. Thus, a total of fifteen images with texture details are obtained for a single image as given in Fig. 9. Histograms of these fifteen images are thereby computed and combined to form the feature vector.

The proposed feature extraction approach is unveiled in Fig. 6. The feature vector length comparison for different state of the art texture-based algorithms is summarized in Table I.

TABLE I. Comparison of Feature Vector Lengths of Various Texture-basedCBIR Mechanisms

Method	Feature vector length
LBP	256
LTP	2 x 256
LMePVEP	4 x 256
LMeTerP	3 x 2 x 256
Proposed algorithm	15 x 256

The extracted set of features is then fed into a classifier to classify between various classes or severity levels of a disease. Two classifiers are used in the proposed methodology viz. back propagation neural networks (BNN) and k nearest neighbors (KNN). In back propagation neural network, the number of neurons in hidden layer has been kept to 10.

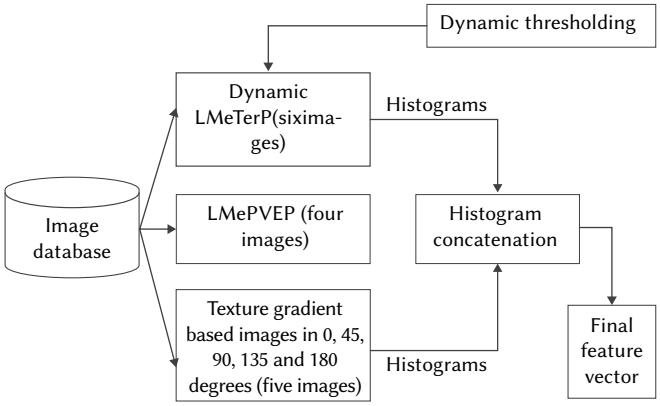


Fig. 6. Schematic representation of feature extraction.

III. Results and Discussion

The proposed framework has been implemented using MATLAB version 2018a platform on a computational device with Windows environment. The computational device has a RAM of 8 GB and 2.0 GHz octa-core processor.

The results are evaluated using two databases. The first database is of Lung CT images for 39 subjects divided into two groups. One group is of no or minimal emphysema and another of mild, average, and high emphysema [36]. A total of 124 images were there in .tiff extension. Fig. 7 shows two CT-Scan images, one from each group.

The second dataset is also of CT images of lungs obtained from ELCAP repository [37]. The dataset had images with nodules present in different location of lungs. A total of 10 different groups are there as per different location of nodules. Fig. 8 shows one image from each group of ELCAP database.

(a) (b)

Fig. 7. CT images of patients with emphysema condition as - (a) minimal or no emphysema (b) High emphysema level.

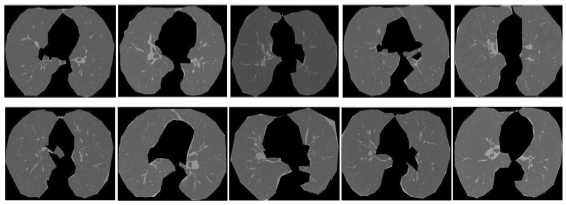
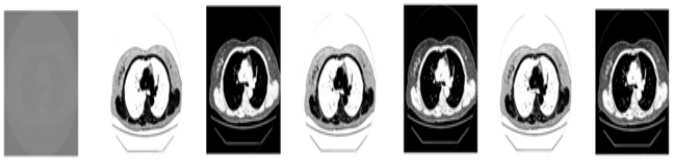
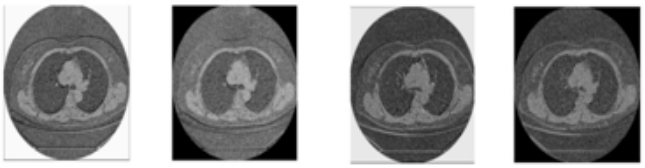


Fig. 8. Images obtained from ELCAP lung database.

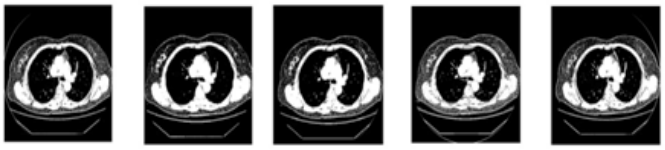
Fifteen images are obtained when we apply the given techniques on a single image. We then generate the histograms of these images and concatenate them to form the final feature vector. All these fifteen images are given in Fig. 9 for a given original image from dataset 1.



(a)



(b)



(c)

Fig. 9. (a) Image taken from emphysema database followed by six images obtained from dynamic threshold based LMeTerP. These six images are for j= 1 for upper LTP, j= 1 for lower LTP, j= 2 for upper LTP, j= 2 for lower LTP, j= 3 for upper LTP and j= 3 by for lower LTP respectively. (b) Four LMePVEP images for j=1 in forward direction, j in backward direction, j=2 in forward direction and j=2 in backward direction, respectively. (c) Five images obtained in 0, 45, 90, 135 and 180 degree direction respectively.

The feature vector is then fed into KNN or BNN for further classification as either an image for diseased individual or as an image for healthy individual. The feature vector of the proposed algorithm has been compared with feature vectors of Hybrid Intelligent Technique (HIT) technique for brain MR image classification [32], LMePVEP technique in [33], Dynamic LMeTerP technique of [39] and Shearlet Based Texture and Radial Basis Transform (SBT-RBF) [42]. The performance is analyzed using Accuracy metric as given in eq. (9) [32].

$$\text{Accuracy} = \frac{No.of\ True\ Positive\ samples}{Total\ number\ of\ samples} \quad (9)$$

Here the true positive samples are the ones that are correctly classified into their respective class. The feature vectors of four algorithms are analyzed on two classifiers viz. K-Nearest Neighbor (KNN) and Back-propagation Neural Network (BNN). The comparison is done by varying percentage of images in testing and training dataset as given in Table II, Table III and table IV respectively.

TABLE II. Accuracy (in %) When the Dataset Is Split for 75 and 25 Percent Training and Testing Data

Classifier	HIT for MR Image classification	LMePVEP	Dynamic LMeTerP	SBT-RBF	Prop. Algo.
	Dataset 1				
KNN	75.1	81.2	85.2	91.3	92.5
BNN	83.4	84.2	85.9	92.8	93.4
	Dataset 2				
KNN	55.5	54.4	55.5	76.5	79.3
BNN	65.6	67.3	67.8	70.4	74.2

TABLE III. Accuracy (in %) When the Dataset Is Split for 80 and 20 Percent Training and Testing Data

Classifier	HIT for MR Image classification	LMePVEP	Dynamic LMeTerP	SBT-RBF	Prop. Algo.
	Dataset 1				
KNN	81.1	87.2	91.3	93.4	95.3
BNN	79.3	82.5	86.2	93.5	93.4
	Dataset 2				
KNN	69.4	68.2	72.2	70.2	73.4
BNN	64.4	66.9	68.3	71.4	74.2

TABLE IV. Accuracy (in %) When the Dataset Is Split for 85 and 15 Percent Training and Testing Data

Classifier	HIT for MR Image classification	LMePVEP	Dynamic LMeTerP	SBT-RBF	Prop. Algo.
	Dataset 1				
KNN	84.5	83.1	84.3	90.5	91.2
BNN	81.1	83.4	85.4	92.3	92.9
	Dataset 2				
KNN	74.07	73.1	74.07	76.2	77.2
BNN	65.0	66.3	66.7	72.8	73.5

As we can see from Table II, table III and table IV, the feature vector of the proposed algorithm outperforms all the other algorithms which are used in the formation of its feature vector for all the cases. In all ratios of testing and training dataset and for both the classifiers the proposed feature vector is better in terms of accuracy.

IV. Conclusion

The paper combines three different approaches of bio-medical image retrieval to propose a feature vector which is robust and enhances the accuracy of image retrieval. From section III, we can calculate the average accuracy for the proposed algorithm in the case of dataset 1 as 93.31% and that of algorithms HIT, LMePVEP, Dynamic LMeTerP

and SBT-RBF to be 80.75%, 83.6%, 86.3% and 92.3% respectively. For dataset 2 these values are 65.66%, 66.03%, 67.4%, 72.91, and 75.03% for algorithmsHIT, LMePVEP, Dynamic LMeTerP, SBT-RBF and the proposed algorithm, respectively.

Thereby the proposed framework has improved the accuracy by 12.56%, 9.71%, 7.01% and 1.01% in average for data set 1 and 9.37%, 8.99%, 7.63% and 2.11% in average for dataset 2 over HIT, LMePVEP, Dynamic LMeTerP and SBT-RBFalgorithms respectively.

References

[1] Eakins, John, and Margaret Graham. "Content-based image retrieval.", JISC technology education programme,1999.

[2] Deep, G., L. Kaur, and S. Gupta. "Local mesh ternary patterns: a new descriptor for MRI and CT biomedical image indexing and retrieval." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, Vol. 6, no. 2, 2018, pp. 155-169.

[3] Haralick, Robert M., Karthikeyan Shanmugam, and Its' HakDinstein. "Textural features for image classification." IEEE Transactions on systems, man, and cybernetics, Vol. 6, 1973, pp. 610-621.

[4] Choi, Hyeokho, and Richard G. Baraniuk. "Multiscale image segmentation using wavelet-domain hidden Markov models." IEEE Transactions on Image Processing, Vol. 10, no. 9, 2001, pp. 1309-1321.

[5] Ghosh, Payel, Sameer Antani, L. Rodney Long, and George R. Thoma. "Review of medical image retrieval systems and future directions." In 24th International Symposium on Computer-Based Medical Systems (CBMS), pp. 1-6, 2011, IEEE.

[6] Zhang, L., Zhou, Z., & Li, H., "Binary Gabor pattern: An efficient and robust descriptor for texture classification". 19th IEEE International Conference on Image Processing (ICIP),2012,pp. 81-84.

[7] Nanni, L., Lumini, A., &Brahnam, S., "Local binary patterns variants as texture descriptors for medical image analysis". Artificial Intelligence in Medicine,Vol. 49(2), 2010, pp.117-125.

[8] Do, M., N., Vetterli, M., "Wavelet-based texture retrieval using generalized Gaussian density and Kullback-leibler distance". IEEE Trans. on Image Processing, Vol.11(2), 2002, pp. 146-158.

[9] Murala, Subrahmanyam, R. P. Maheshwari, and R. Balasubramanian. "Directional binary wavelet patterns for biomedical image indexing and retrieval." Journal of Medical Systems, Vol. 36, no. 5,2012, pp. 2865-2879.

[10] Murala, Subrahmanyam, R. P. Maheshwari, and R. Balasubramanian. "Directional local extrema patterns: a new descriptor for content based image retrieval." International journal of multimedia information retrieval, Vol. 1, no. 3, 2012, pp. 191-203.

[11] Murala, S., Maheshwari, R., P., and Balasubramanian, R., "Local maximum edge binary patterns: a new descriptor for image retrieval and object tracking". Signal Processing, Vol. 92(6), 2012, pp.1467-1479.

[12] Ojala, Timo, Matti Pietikainen, and Topi Maenpaa. "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns."IEEE Transactions on pattern analysis and machine intelligence, Vol.24.7, 2002, pp. 971-987.

[13] Tan, Xiaoyang, and Bill Triggs. "Enhanced local texture feature sets for face recognition under difficult lighting conditions." IEEE transactions on image processing, Vol. 19.6,2010, pp. 1635-1650.

[14] Rahman, Md Mahmudur, Prabir Bhattacharya, and Bipin C. Desai. "A framework for medical image retrieval using machine learning and statistical similarity matching techniques with relevance feedback." IEEE transactions on Information Technology in Biomedicine, Vol. 11.1, 2007, pp. 58-69.

[15] Rajeswari, J., and M. Jagannath. "Advances in biomedical signal and image processing–A systematic review." Informatics in Medicine Unlocked, Vol. 8, 2017, pp. 13-19.

[16] Galshetwar, Gajanan M., et al. "Edgy salient local binary patterns in inter-plane relationship for image retrieval in Diabetic Retinopathy." Procedia Computer Science, Vol. 115, 2017, pp. 440-447.

[17] Jyothi, B., MadhaveeLatha, Y., Mohan, P. K., & Reddy, V. S. K.,"Integrated multiple features for tumor image retrieval using classifier and feedback methods",Procedia Computer Science, Vol. 85, 2016, pp.141-148.

[18] Ma, Ling, et al. "A new method of content based medical image retrieval and its applications to CT imaging sign retrieval." Journal of biomedical informatics, Vol. 66, 2017, pp. 148-158.

[19] Xu, Songhua, and Michael Krauthammer. "A new pivoting and iterative text detection algorithm for biomedical images." Journal of biomedical informatics, Vol. 43(6),2010, pp. 924-931.

[20] Farruggia, Alfonso, Rosario Magro, and Salvatore Vitabile. "A text based indexing system for mammographic image retrieval and classification." Future Generation Computer Systems, Vol. 37,2014, pp. 243-251.

[21] Simpson, Matthew S., Daekeun You, Md Mahmudur Rahman, ZhiyunXue, Dina Demner-Fushman, Sameer Antani, and George Thoma. "Literature-based biomedical image classification and retrieval." Computerized Medical Imaging and Graphics, Vol. 39, 2015, pp. 3-13.

[22] Charles, YesubaiRubavathi, and Ravi Ramraj. "A novel local mesh color texture pattern for image retrieval system." AEU-International Journal of Electronics and Communications, Vol. 70(3),2016, pp. 225-233.

[23] Verma, Manisha, and Balasubramanian Raman. "Center symmetric local binary co-occurrence pattern for texture, face and bio-medical image retrieval." Journal of Visual Communication and Image Representation, Vol.32,2015, pp.224-236.

[24] Kumar, Yogesh, et al. "An efficient and robust approach for biomedical image retrieval using Zernike moments." Biomedical Signal Processing and Control, Vol. 39, 2018, pp.459-473.

[25] Singh, Vibhav Prakash, and Rajeev Srivastava. "Automated and effective content-based mammogram retrieval using wavelet based CS-LBP feature and self-organizing map." Biocybernetics and Biomedical Engineering, Vol. 38(1),2018, pp.90-105.

[26] Deep, G., L. Kaur, and S. Gupta. "Directional local ternary quantized extrema pattern: A new descriptor for biomedical image indexing and retrieval." Engineering Science and Technology, an International Journal, Vol. 19(4), 2016, pp. 1895-1909.

[27] Tiwari, Ashwani Kumar, Vivek Kanhangad, and Ram Bilas Pachori. "Histogram refinement for texture descriptor based image retrieval." Signal Processing: Image Communication, Vol 53,2017, pp. 73-85.

[28] Li, Wenbo, Haiwei Pan, Pengyuan Li, XiaoqinXie, and Zhiqiang Zhang. "A medical image retrieval method based on texture block coding tree." Signal Processing: Image Communication, Vol. 59, 2017, pp. 131-139.

[29] Murala, Subrahmanyam, and QM Jonathan Wu. "Spherical symmetric 3D local ternary patterns for natural, texture and biomedical image indexing and retrieval." Neurocomputing, Vol. 49 (1), 2015, pp. 1502-1514.

[30] Qayyum, Adnan, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. "Medical image retrieval using deep convolutional neural network." Neurocomputing, Vol. 266, 2017, pp.8-20.

[31] Alzu'bi, Ahmad, Abbes Amira, and Naeem Ramzan. "Content-based image retrieval with compact deep convolutional features." Neurocomputing, Vol. 249, 2017, pp.95-105.

[32] El-Dahshan, El-Sayed Ahmed, Hosny, Tamer and Salem Abdel-Badeeh M, Hybrid intelligent techniques for MRI brain images classification, Digital Signal Processing, Vol. 20(2), 2010, pp.433-441, Elsevier.

[33] Murala, Subrahmanyam, and QM Jonathan Wu. "MRI and CT image indexing and retrieval using local mesh peak valley edge patterns." Signal Processing: Image Communication, Vol. 29(3), 2014, pp.400-409.

[34] S. Varun, and R. K.Purwar, "A Five-Level Wavelet Decomposition and Dimensional Reduction Approach for Feature Extraction and Classification of MR and CT Scan Images," Applied Computational Intelligence and Soft Computing, Vol. 2017, Article ID 9571262,2017,.doi: https://doi.org/10.1155/2017/9571262

[35] Purwar, Ravindra Kr, and Varun Srivastava. "Recent Advancements in Detection of Cancer Using Various Soft Computing Techniques for MR Images." Progress in Advanced Computing and Intelligent Engineering. Springer, Singapore,pp.99-108, 2018.

[36] Emphysema dataset, L. Sørensen, S. B. Shaker, and M. de Bruijne, "Quantitative Analysis of Pulmonary Emphysema using Local Binary Patterns", IEEE Transactions on Medical Imaging, Vol. 29(2), 2010, pp. 559-569.

[37] ELCAP dataset, A. P. Reeves, A. M. Biancardi, D. Yankelevitz, S. Fotin, B. M. Keller, A. Jirapatnakul, J. Lee. "A Public Image Database to Support Research in Computer Aided Diagnosis," In 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3715-3718, Sept. 2009.

[38] Sverzellati,Nicola,DavidA.Lynch,MassimoPistolesi,Hans-UlrichKauczor, Phillippe A. Grenier, Carla Wilson, and James D. Crapo. "Physiologic and quantitativecomputedtomographydifferencesbetweencentrilobular and panlobularemphysemainCOPD."ChronicObstructivePulmonaryDiseases: Journal of the COPDFoundation, Vol. 1(1), 2014, p. 125.

[39] Srivastava Varun, Ravindra K. Purwar, and Anchal Jain. "A dynamic threshold-based local mesh ternary pattern technique for biomedical image retrieval." International Journal of Imaging Systems and Technology, Vol. 29(2),2019, pp. 168-179.

[40] Srivastava Varun and Ravindra Purwar. "An extension of local mesh peak valley edge based feature descriptor for image retrieval in bio-medical images." ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, Vol. 7(1),2018, pp.77-89.

[41] Srivastava, Varun, and Ravindra Kr Purwar. "Classification of CT scan images of lungs using deep convolutional neural network with external shape-based features." Journal of digital imaging, Vol. 33(1), 2020, pp.252-261.

[42] Semunigus, Wogderes. "Pulmonary Emphysema Analysis using Shearlet based textures and radial basis function network" International Journal of Advances in Signal and Image Sciences, Vol. 6(1), 2020, pp: 1-11.

Varun Srivastava

Mr. Varun Srivastava is currently associated as an Assistant Professor with Computer Science Department of Bharati Vidyapeeth's College of Engineering. His main research areas are Biomedical Image Processing and Pattern Recognition. He has many research articles to his credit in these domains.

Shilpa Gupta

Ms. Shilpa Gupta is currently associated with Computer Science Department of Bharati Vidyapeeth'College of Engineering as an Assistant Professor. Her main areas of research are Machine learning and Deep networks. She has contributed many renowned publications in these domains.

Gopal Chaudhary

Dr. Gopal Chaudhary is currently working as an assistant professor in Bharati Vidyapeeth's College of Engineering, Guru Gobind Singh Indraprastha University, Delhi, India. He holds a Ph.D. in Biometrics at the division of Instrumentation and Control engineering, Netaji Subhas Institute of Technology, University of Delhi, India. He received the B.E. degree in electronics and communication engineering in 2009 and the M.Tech. degree in Microwave and optical communication from Delhi Technological University (formerly known as Delhi College of Engineering), New Delhi, India, in 2012. He has 30 publications in refereed National/International Journals & Conferences (Elsevier, Springer, Inderscience) in the area of Biometrics and its applications. His current research interests include soft computing, intelligent systems, information fusion and pattern recognition. He has organized many conferences and special issues.

Arun Balodi

Dr. Arun Balodi has done his Ph. D. from Indian Institute of Technology Roorkee and is currently heading the Electronics and Communication Engineering Department of Atria Institute of Technology, Bengaluru. His main areas of research are Image processing and Machine learning.

Manju Khari

Dr. Manju Khari is an Assistant Professor in Netaji Subhas University of Technology, East Campus, Delhi, India. She is also the Professor- In-charge of the IT Services of the Institute and has experience of more than twelve years in Network Planning & Management. She holds a Ph.D. in Computer Science & Engineering from National Institute Of Technology Patna and She received her master's degree in Information Security from Ambedkar Institute of Advanced Communication Technology and Research, formally this institute is known as Ambedkar Institute Of Technology affiliated with Guru Gobind Singh Indraprastha University, Delhi, India. Her research interests are software testing, information security, optimization, Image processing and machine learning. She has 70 published papers in refereed National/International Journals & Conferences (viz. IEEE, ACM, Springer, Inderscience, and Elsevier) and 10+ edited books from reputed publishers. She is also co-author of two books published by NCERT of Secondary and senior Secondary School.

Vicente García-Díaz

Dr. Vicente García-Díaz is a Software Engineer, PhD in Computer Science. He has a Master in Occupational Risk Prevention. He has the qualification of University Expert in Blockchain Application Development. He is an Associate Professor in the Department of Computer Science at the University of Oviedo. He is also part of the editorial and advisory board of several journals and has been editor of several special issues in books and journals. He has supervised 100+ academic projects and published 100+ research papers in journals, conferences, and books. His teaching interests are primarily in the design and analysis of algorithms and the design of domain-specific languages. His current research interests include machine learning, decision support systems, eHealth and eLearning.

Alzheimer Disease Detection Techniques and Methods: A Review

Sitara Afzal[1], Muazzam Maqsood[1]*, Umair Khan[1], Irfan Mehmood[2], Hina Nawaz[1], Farhan Aadil[1], Oh-Young Song[3]*, Yunyoung Nam[4]*

[1] Department of Computer Science, COMSATS University Islamabad, Attock Campus (Pakistan)
[2] Department of Media Design and Technology, University of Bradford, Bradford (UK)
[3] Department of Software, Sejong University, Seoul 05006 (Korea)
[4] Department of Computer Science and Engineering, Soonchunhyang University, Asan 31538 (Korea)

Received 25 August 2020 | Accepted 17 October 2020 | Published 21 April 2021

Abstract

Brain pathological changes linked with Alzheimer's disease (AD) can be measured with Neuroimaging. In the past few years, these measures are rapidly integrated into the signatures of Alzheimer disease (AD) with the help of classification frameworks which are offering tools for diagnosis and prognosis. Here is the review study of Alzheimer's disease based on Neuroimaging and cognitive impairment classification. This work is a systematic review for the published work in the field of AD especially the computer-aided diagnosis. The imaging modalities include 1) Magnetic resonance imaging (MRI) 2) Functional MRI (fMRI) 3) Diffusion tensor imaging 4) Positron emission tomography (PET) and 5) amyloid-PET. The study revealed that the classification criterion based on the features shows promising results to diagnose the disease and helps in clinical progression. The most widely used machine learning classifiers for AD diagnosis include Support Vector Machine, Bayesian Classifiers, Linear Discriminant Analysis, and K-Nearest Neighbor along with Deep learning. The study revealed that the deep learning techniques and support vector machine give higher accuracies in the identification of Alzheimer's disease. The possible challenges along with future directions are also discussed in the paper.

Keywords

Alzheimer's Disease Review, Mild Cognitive Impairment, Neuroimaging, Machine Learning Classifiers, Deep Learning.

DOI: 10.9781/ijimai.2021.04.005

I. Introduction

THE most frequent form of dementia is Alzheimer's disease. About 1 out of 85 people in the world are suspected to have Alzheimer's Disease by 2050 [1]. In this disease, neurons are lost because of the accumulation of abnormal proteins in the form of plaques tau tangles of neuro-fibrillary in the brain of a person [1]. AD occurs in the temporal lobe of the brain and hippocampus [1] thus changes the brain even before the symptoms of dementia occur [2]. It has been proposed that this inescapable decay can be a profitable marker of neurodegeneration, as estimated with sMRI (Structural Magnetic Resonance Imaging). Moreover, functional MRI (fMRI) and fluorodeoxyglucose positron-emission tomography (FDG-PET) [3], [4] detect the alterations in function, metabolism, and connectivity. In the starting stages of AD, it becomes hard to distinguish between the patterns with quantitative analysis and even with radiological readings, this is because of the nuance in the pattern. So, it is challenging to diagnose and monitor disease at its early stage.

Individuals in the underlying phase of Alzheimer's Disease are considered to have Mild-Cognitive-Impairment (MCI) [5], even though not every individual grows Alzheimer's Disease having MCI. Mild-Cognitive-Impairment is a provisional phase from typical to Alzheimer's Disease, where an individual has gentle changes in the psychological capacity, which is evident to the individual and family members, yet the individual can perform daily routine activities.

Around 15 to almost 20% of individuals, matured to 60+ or more have Mild-Cognitive-Impairment, and 30 to almost 35% of people with MCI grow to AD within 4 years [6]. This transition takes time between six to three years, yet ordinarily, it takes a year and a half. MCI patients would then be able to be classified as Mild-Cognitive-Impairment converters or non-converters; means that the individual may or may not have changed to AD within a year and a half. There are additionally different subtypes of MCI that are occasionally referenced in the writing, for example, first MCI. The utmost critical hazard influences for Alzheimer's Disease are family ancestries as well as the nearness of linked qualities in an individual's genes. An Alzheimer's Disease evaluation depends on a medical assessment, just as an exhaustive meeting of the patient and their family members [7]. The evaluation of Alzheimer's Disease must be completed using dissection, which is not clinically useful [8]. Without this ground-truth information, individuals require additional models to diagnose Alzheimer's Disease. These measures can enhance the comprehension of AD, and make analysis workable for existing individuals.

* Corresponding author.
E-mail addresses: muazzam.maqsood@cuiatk.edu.pk (M. Maqsood), oysong@sejong.edu (O. Y. Song), ynam@sch.ac.kr (Y. Nam).

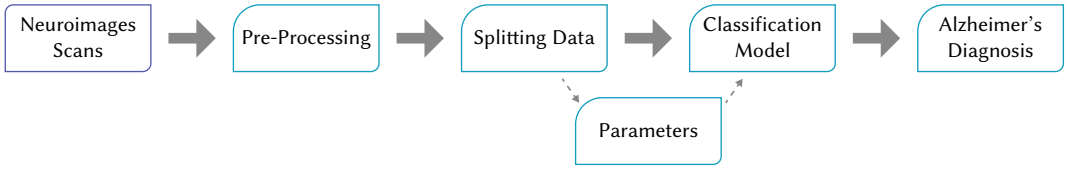


Fig. 1. The general flow of Computer-Aided Diagnosis (CAD) system in the diagnosis of Alzheimer's.

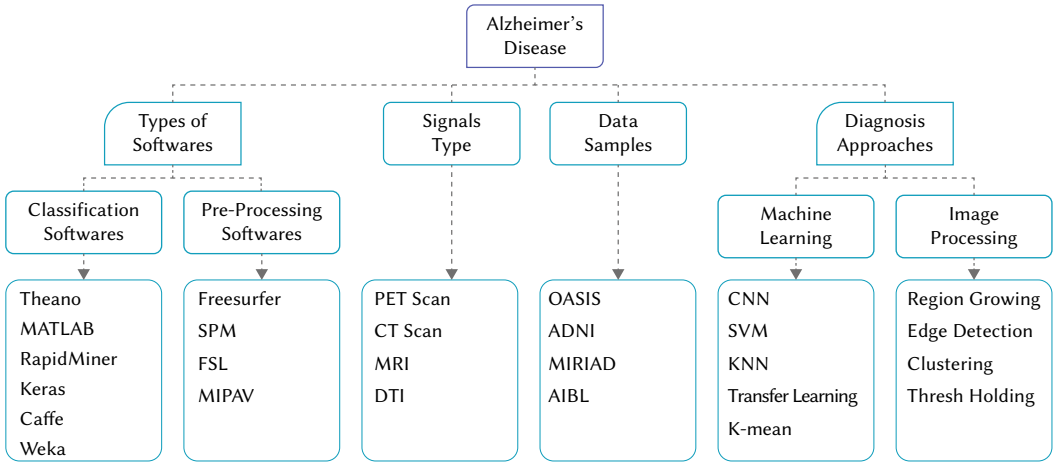


Fig. 2. The overall flow diagram of the survey.

Various neuroimaging studies, which have used region of interest (RoI) to find out the subtle changes related to AD, were solely dependent on previous knowledge to suggest the selection of features, ignoring the changes in the brain besides the region studied. Fig 1 demonstrates the general flow of the CAD system in the diagnosis of Alzheimer's. In contrast, machine learning is a systematic approach that develops automatic as well as objective classification and analyzes huge amounts of data either complex or simple and efficiently distinguishes between the subtle changes occurring in the brain images.

Mostly, feature extraction and classification methods, i.e.; classification frameworks make predictive models [9] to help in decision making and facilitating automation of medical decisions. Moreover, imaging markers or records can be created by using classification frameworks with improved affectability and an individual's specificity. This makes a more individualized and tolerant customized approach, which is basic in the present period of customized medication. It permits to encourage the thought of hereditary or way of life dangers, by using progressed computational control. In the past decades, a lot of work has been done on the neuroimaging-based classification of AD, to make it computerized to diagnose at its early stages. Its rapid study has motivated us to summarize different AD-related work from feature extraction using different neuroimaging data to different classification methods. Moreover, we discuss problems related to the limited sample size and data setting variability by analyzing different studies. The most significant and primary test in AD appraisal is to decide if somebody has Mild-Cognitive-Impairment and to anticipate if a Mild-Cognitive-Impairment individual will build up into neurodegenerative sickness. Different phases of the disease, for example, early/late MCI are equally important to diagnose. Identifying AD utilizing Deep learning is generally a test for analysts due to:

- Poor quality images during the pre-processing phase.
- Unavailability of publicly available large data samples for research.
- Limited labeled dataset for AD.
- Deficiency of essential data points, specifically for the identification of ROI in the cerebrum.

The general flow of the CAD system in the diagnosis of Alzheimer's is demonstrated in Fig. 1. We will generalize the existing AD classification studies. Furthermore, critical parts, which were previously not explored in AD will also be discussed. The past review papers on AD [4], [10], limited AD classification to MRI but the pathological changes in the brain-related to AD can be diagnosed by several different modalities of imaging which include FDG-PET and amyloid-PET. This is the reason that a comprehensive review of AD classification is needed.

This paper also discusses cross-validation strategies i.e. independent training and testing of classification algorithms. It uses two strategies, split-n-train and k-fold cross-validation [11], [12] for unbiased results. We further discuss the different imaging modalities used for AD identification along with machine learning techniques and algorithms used in this domain. We conclude our review by highlighting the limitations and research challenges along with possible future research directions for researchers.

II. Scope of this Review

Computer-Aided-Diagnosis (CAD) of Alzheimer's has opened an important area for the early detection of Alzheimer's Disease. In this survey, the papers are reviewed from repositories of IEEE, ACM libraries, Science Direct, Springer containing keywords like AD, Alzheimer's Disease, deep learning, machine learning, and image processing. These all online databases were selected as they are well-known for their authenticity and they offered the most significant peer-reviewed articles covering the field of image processing, machine learning, and deep learning. While going through these databases, the terms used for searching was expecting to cover most of the effort including image processing, neural networks, and machine learning approaches for the identification of Alzheimer's Disease. The scope of the paper is to make a survey and analyze the studies and research of different groups in image processing-based approaches as well as machine learning-based techniques. This paper also helps the new researcher who is starting to explore the computer-aided methods for AD diagnosis. Fig. 2 shows a general overview of this paper.

III. Imaging Modalities

Different enhanced imaging techniques are being used to identify the signals that are leading toward more accurate AD detection. The quantitative analysis of brain degeneration identification is being more comprehensively applied. Diverse neuroimaging-based signals like CT scan, PET, sMRI, fMRI, and DTI are utilized to produce a more conclusive prediction.

A. Computerized Tomography (CT) Scan

CT scan is a cross-sectional illustration of the brain region that is produced with the help of an x-ray with a constant bombardment of radioactive rays. These images are 100% more transparent than normal x-rays. CT scans cannot be considered as a benchmark for the early detection of AD as other advanced methods are providing precise and accurate results. Some studies tried to endorse its efficiency in affected AD detection by emphasizing its coherence and cost-effective conclusions in comparison with other techniques like PET or MRI. It is safe to conclude that it does not play any substantial role in the early detection of AD.

B. Positron Emission Tomography (PET)

PET scan is a volumetric subatomic illustration method that is used to obtain a 3D brain scan on anatomical and sub-anatomical levels.

PET scanning is done by administering or inhaling a radioactive isotope as a tracing agent also known as a radiotracer. That works as a positron-emitting spec for the subject. Then this radiotracer is detected by a scanning machine. Afterward, the scanner provides a digital image (illustration) of the radiotracer spread in the subject body. The nature of the PET scan depends on a different kind of radiotracer being used. The cost of PET scanning has been raised due to the use of cyclotron agents that play an essential role in the preparation of radiotracer. As the brain functionality depends on the consumption of blood sugar, so the illustration can be deduced that glucose consumption and neural functioning are directly proportional. The PET scan is very peculiar in the prediction of AD even with mild symptoms. The working of PET scanning is quite effective, but the reasons mentioned above show that this is not a healthy diagnostic method.

C. Structural Magnetic Resonance Imaging (sMRI)

MRI is a non-invasive (non-anatomical) imaging (illustration) method that is used for structural evaluation of the concerned brain region. MRI scanners are used in brain scanning. The MRI procedure involves the bombardment of the magnetic rays on the area of which imaging is required. Different kinds of areas are identified due to tissue movement. MRI is widely utilized for the early diagnosis of AD. The number of studies based on MRI has been increased for years due to the availability of open-source databases like ADNI and OASIS. This has also impacted the studies driven around these databases to help detect disease progression monitoring and improved analytical studies. With the help of MRI, the disease impact on the subject's spatial domain and temporal region creating patterns can also be observed. MRI-based studies helped to verify results that the brain tissue degeneration in patients affected with AD and transitioning from MCI to AD can be a speedier process than that of a healthy individual. The concerned domain in this regard could be to detect subjects with MCI which can be done with early detection of AD. The accuracy for MRI based model is 93.18% with 93% precision, 92% recall, and 92% f1–score. Various studies show that early detection of AD with the help of MRI is not 100% accurate. The degeneration in the hippocampal region can easily be identified in a patient suffering from AD in comparison with a non-AD patient. As in many cases, the brain damage is solely not due to AD so determining these meek differences can be critical. But over time MRI based studies (automatic MRI) showed promising improvement. Currently, MRI-based techniques are widely used for the early detection of AD because MRI devices are easily available now.

D. Functional Magnetic Resonance Imaging (fMRI)

The process of functional MRI is also a non-invasive procedure that aids the diagnosis of malfunction caused by AD. fMRI also allows observing the absorption of oxygen during resting and active state to form an activity pattern. So, brain activity during different states can be evaluated. fMRI extracts data from each region of the brain to help diagnose AD. Studies over the period show that patients suffering from AD have reduced activity in the limbic region especially in the hippocampus due to brain damage, and plaque abnormalities, and cerebral cortex region due to vascular damage. But these exceptions are less notable in patients suffering from MCI that indicate the less evident use of fMRI in early detection of MCI. Some studies have also shown some incongruous conclusions for hippocampus regions. In this regard, a resultant U-curve shape is formed. One of the most beneficial uses of fMRI is that it does not involve the use of radioactive substances or radiations due to this fMRI can be used as many times as needed. As patients advance levels of disease, suffering from extreme cognitive impairment, they cannot have adequate motor control. So, to get good results patients need to be steady while going through the scanning process. The prompt changes of AD are due to the neuro-degeneration process of intentional availability between different brain regions [13]. Different researchers working on rs-fMRI have talked about the presence of common changes inside the gut concerning resting-state systems. Oxygen level-subordinate that are initiated as local patterns are generally concerned for the most stressed part of cerebrum capacities like tangible, and non-appearance color innovation [14] [15] [16] [17] [18] [19].

E. Diffusion Tensor (DTI)

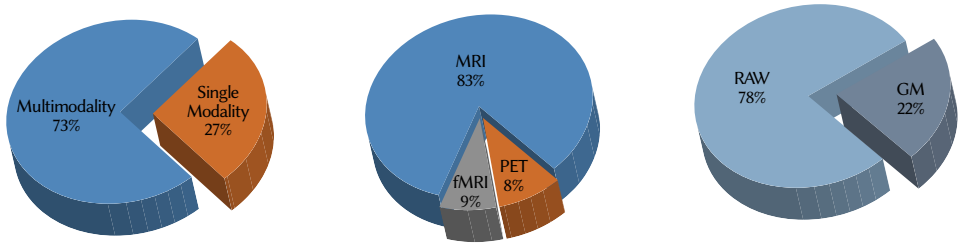
DTI is an MRI-based imaging technique that illustrates minute structural cross-sectional details of brain regions. These samples are also procured non-invasively with MRI scanners. DTI is roughly based on the Browning motion sensation of water molecule activity of human tissues. So, this phenomenon can be described as the microscopic dimension that measures the size, dimensionality, and orientation of the tissue that helps identify the last stage of microscopic degeneration. Studies show that DTI can be an implicit mechanism to aid early-stage AD identification. The Look into utilizing DTI-based features, might lie in expansion isolated under three categories, contingent upon how characteristic would be extracted: i) tractography, ii) integrated network measure process, then iii) distinctive voxel-preference approach.

In the respective section, different types of images utilized in the detection and classification of AD along with their modalities were discussed. Their utilization in the existing literature was also referenced and elaborated. Fig. 3 summarizes the modality of images through the modality chart of images utilized in the above-mentioned techniques.

IV. State of the Art Alzheimer's Detection and Classification Methodology

A. Image Preprocessing

Multiple approaches to image processing were deployed for the diverse domain of studies. In [8] the image processing process involved visual inspection for essential irregularities distinct to Alzheimer's or FTLD and artifacts. The method included: the use of SPM5 (http://www.fil.ion.ucl.ac.uk/spm); The combined images of groups I and III and patients from II and IV using in implementation diffeomorphic



a b c

Fig. 3. (a) This chart demonstrates the multimodality and single modality images ratio (b) This chart demonstrates the MRI, fMRI, and PET images ratio (c) This chart demonstrates the RAW and Gray-Matrix (GM) ratio.

TABLE I. Feature Extraction Techniques Using Machine Learning

Type	Description	Benefits & Limitations
Local Binary Pattern (LBP):	The working of LBP is by limiting a certain gap of image pixels utilizing the middle image pixel value and understands output as a dual array i.e. binary pattern. These image pixels are further labeled by utilizing these patterns. The histogram of these arrays signifies the feature description. To create the system, a well-organized bigger window is suitable, as the values of pixels differ slowly.	**Benefits**: LBP features are computationally simple and fast. It also takes a shorter training time. **Limitations**: LBP features are less accurate because of the high false-positive rate.
Scale Invariant Feature Transform (SIFT):	This approach joins identification and explanation of the features [40]. Its working is by increasing the regions for interest point assortment and further pinpointing important facts in that picture. The bearings info is formerly given to certain important points, that in conclusion defines the Scale Invariant Feature Transform (SIFT).	**Benefits**: The implementation of SIFT is fast and has a robust high rate matching. **Limitations**: SIFT feature extractions are too sensitive to input type and smoothing.
Gray-Level-Cooccurrence-Matrix (GLCM)	It was initially applied for texture examination and it exceeded the best-in-class methods. The working is stipulating the co-occurrence of different levels of gray at a precise position in the image in a precise section of the usual image of gray level, the pixels are generally of related gray level in a precise section, and are greatly related.	**Benefits**: The computational time of GLCM features is low and that is why memory consumption is also very low. **Limitations**: It only works with grayscale images.

registration algorithm; The preprocessing steps for [20] for baseline 1.5 t and T1 weighted MRI dataset from the ADNI repository. These images were normalized and segmented with Statistical Parametric Mapping (SPM) software package. The preprocessing of [21] was conducted with the statistical software package SPM2 (Welcome Trust Centre for Neuroimaging, London, https://www.fil.ion.ucl.ac.uk/spm/). This domain dataset area is of binary class type as they were of Alzheimer's disease (AD) and normal class (NC). Each MRI was 3D, with the tensor intensity value equal to 110×110×110. As one subject can have more than one MRI in the data repository, to avoid the redundancy between the training and testing dataset option,s the best approach of only using the best and earlier acquired MRI's for individual subjects was considered. For this purpose, multiple diverse amounts of images were selected for each training and testing batch for individual studies. In each study preprocessing method was quite like the MRI's or PET scans however, dimensions were altered according to the dimension requirements of the said algorithms.

B. Image Processing Based Techniques for Alzheimer Detection

Image processing-based approaches are utilized to extract the features from the images. The following are the different Image processing-based approaches. Table I demonstrates the features that are mostly extracted by utilizing these approaches from neuroimages.

1. Thresholding

For image segmentation and object detection, thresholding is one of the most widely used techniques. Thresholding is used to separate and elaborate the concerned terrain of the foreground from the background aptly for analytical purposes. The building block of this method depends on the pixel's intensity values of the image under analysis. Threshold values are specified/classified as an intensity histogram for the background (Erosion) and foreground (Dilation) and then both distinct values are analyzed to separate them.

2. Region-Based Methods

In this method, the concerning area is selected based on pre-defined rules according to the nature of the algorithm and dataset. The focus in this regard is the intensity values of the object boundaries in an image. In this process, a seed point is selected and pixels with similar values are used based on properties such as intensity, texture, and color. With these values, concerned regions are separated from the rest of the image. Using the region growing method solely for object detection is not enough as it is usually used with additional techniques like Region splitting, split and merge, and Edge detection.

3. Clustering

Grouping together similar data-points is known as clustering. Clustering is an unsupervised learning approach as it is not based on advanced learning of training data.

But the iterative nature of the clustering method takes care of the segmentation method. The process of clustering continues till each similar object is not assigned to a cluster with the same attributes, for this purpose a similarity measure is defined in advance.

4. Atlas Guided Approaches

This map or atlas-based approach is used to form an image analysis. The image is designating a specific region to form a shape. The primary motivation behind this methodology is to assist radiologists in the revelation and ID of illnesses. The working progress of the approach is enhanced by distinguishing noteworthy life systems in the clinical images.

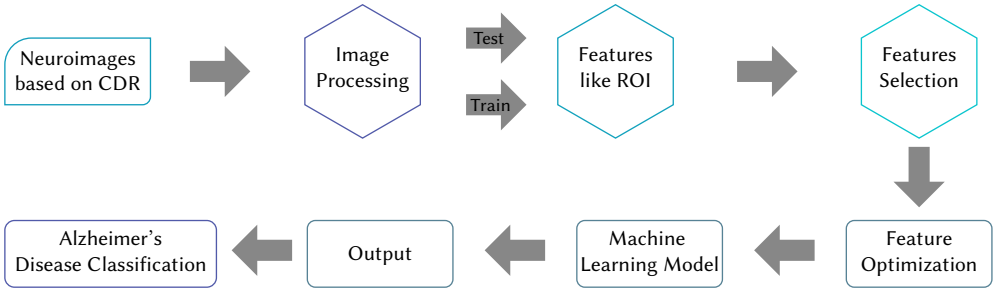


Fig. 4. The fundamental framework for the ML approach for the classification of AD.

5. Edge-based Methods

These approaches are used to detect object boundaries. It is also used to find cracks in detected boundaries. This is one of the most widely used methods to detect object boundaries with the same pixel intensities. The distinction between the pixels, in this case, is carried out by estimating the intensity gradient. These methods are mainly used as a base or central technique for other segmentation approaches.

C. Machine Learning Techniques for Alzheimer Detection

These techniques are extensively used in the clinical application[22] and also established noteworthy attention in the past eras [23]. It is well-thought-out as a division of AI as it allows the removal of an expressive model or pattern from samples. These approaches are mainly categorized into clustering/unsupervised and classification/supervised. Below is a brief explanation of Classification i.e. Supervised Learning and clustering i.e. Unsupervised Learning. The main idea behind this machine learning approach is centered on a trainer which trains label data with the label group by utilizing a training set. Different types of biomarkers are called features that must be learned for the difficulty at hand. Fig. 4 shows the fundamental framework for the ML approach for the classification of AD.

1. Bayesian Classifiers

Bayesian classifiers are the simple statistical classifier centered on Bayes theorem with robust (naïve) independence expectations amongst the feature [24]. Instead of undergoing training by a single algorithmic approach, Naïve Bayes learns by utilizing algorithms centered on universal rules. In [25] Seixas et al. give a Bayesian classifier i.e. Bayesian network decision model for helping detection of Alzheimer disease, Mild Cognitive Impairment (MCI), and Normal Control (NC), while they achieved higher results as compared to numerous famous classifiers like simple Naïve Bayes, Logistic Regression Classification (LRC), multilayer perception Artificial Neural Network, decision table, choice base enhanced through Adaboost approach and J48 Decision Tree. In [26] Liu et al. give the multifold Bayesian kernelization approach that could discriminate Alzheimer's disease and normal control with higher accuracy value nonetheless attained bad output in identifying MCI-converter (MCI) and MCI non-converter (MCI). In [27] plant et al. join a particular feature with grouping by utilizing a Bayesian classifier for the judgment amongst Alzheimer's Disease and Normal Control on Magnetic Resonance Images data and stated up to 92.0% accuracy. In [28] Lopez et al. use the multivariate approach, for example, PCA and Linear Discriminate Analysis for extracting feature, after that applied the Bayesian frame for automatic detection of Alzheimer's Disease and Normal Control by utilizing Positron Emission Tomography i.e. PET and Single Photon Emission Computed Tomography i.e. SPECT.

2. Support Vector Machine

SVM i.e. Support Vector Machine is a very common machine learning algorithm used for both regression and classification-based problems. For non-linear separable data, sample points are mapped over a high dimensional plane exploiting the boundary among points in a high dimensional space [29], [30]. This transformation of data-points from one dimension to another dimensional space is known as the kernel trick [31]. This optimal algorithm is established over the 'Training Set' where the training data is utilized for maturing the procedure that can distinguish among previously defined clusters and a 'Testing Set' where the procedure is utilized to forecast the clusters where the new observations go. To plot the training set from input space to a high-level feature space, Kernel functions can be utilized [32]. The data points that are positioned nearest to the decision surface are support vectors. Support vector machine have different variants, for example hierarchal SVM [33], radial basis function-based SVM [34] and AdaBoostSVM [35]. In the past decades, the interest in SVM in Alzheimer's Disease studies has been increasing [36]. For the classification of patients with Alzheimer's Disease, the early applications of support vector machines to neuroimaging data were mainly meant to translating the mental state of Normal or healthy subjects [37]. Effectively established that it was achievable to differentiate amongst subjects giving correct and incorrect responses with the accuracy of 99.3% centered exclusively on discriminative forms of brain actions. Recently in 2019, Afzal. S et al. [38] utilized machine learning, the Random Forest approach for the classification of multi-class Alzheimer's Disease, and 92.4% accuracy was attained in their findings. MicroRNAs or MiRNA are single-stranded non-coding RNA particles that exhibit distinctive expression to varying pathological and physiological conditions. To differentiate between AD and other neurological ailments that are centered on such genome-based biomarkers, SVM was utilized by Smith-Vikos and Slack [39]. By utilizing three blood-based biomarkers [41], oxidized LDL, antibodies [42], Laske et al. [43] also utilized Support Vector Machine to distinguished patients with Alzheimer's Disease from other normal subjects and attained 81.7%. Higher accuracy was reported by comparing the output of the thresholding-based approach with the SVM-based segmentation. Accuracy of 92.31% plus 96.67% was reported for Alzheimer's disease (AD) diagnosis by utilizing Support Vector Machine in PET and SPECT respectively by Lopez et al. [44]. Juergen Dukart et al. utilized SVM to combine magnetic resonance images (MRI) and FDG-PET for enhancing the identification of AD. They casually removed FDG_PET and Magnetic resonance Images from two different databases i.e. ADNI and Leipzing Cohorts. They attained accuracy of 88% for ADNI datasets and up to 100% for Leipzing datasets [45]. Y. Zhang et al. distinguish among elderly subjects with Alzheimer's Disease, MCI, and Normal control (NC). They utilized 5-fold cross-validation for KSVM-DT and attained 80% classification accuracy [46]. P. Padilla et al. proposed a new Computer-Aided Technique (CAD) centered on non-negative matrix factorization and support vector machine for initial AD analysis. They utilized two databases (PET and SPECT), both databases containing AD patients and healthy control. They proposed NMF-SVM and yields 91% accuracy [47]. In [48], researchers used SVM on MRI and SPECT images and attained much higher accuracy for MRI instead

TABLE II. SVM Based Techniques

Author	Targets	Methods	Imaging Modality	Accuracy
Afzal. S et al. [38]	Multiclass AD classification	SVM	MRI	92.4%
Smith Vikos et al.[42]	AD vs NC	SVM	MRI	90.3%
Laske et al.[43]	AD vs NC	SVM	MRI	81.7%
Lopez et al.[44]	AD vs NC	SVM	MRI	96%
Y.Zhang et al.[46]	MCI vs NC	SVM	MRI	80%
P.Padilla et al.[47]	AD vs NC	NMF-SVM	PET and SPECT	91%
Luiz K Feriera et al.[48]	FDG-PERT vs MRI	SVM	PET	68-71%
Gerardin [51]	AD vs NC	Hippocampi + SVM	MRI	94.0%
	MCI vs NC			83.0%
Hackman [52]	MC vs NC	Wavelet Transform + SVM	MRI	80.44%
Dukart [45]	AD vs NC	Meta-Analysis + SVM	MRI, PET	85.7%
				100.0%
Ortiz[53]	AD vs NC	LVQ + SVM	MRI	91.0%
		PCA + SVM		81.0%
		VAF + SVM		71.0%
Nir[60]	AD vs NC	Diffusion Weighted method + SVM	MRI	86.2%
	MCI vs NC			82.0%
Schmitter [54]	AD vs NC	FreeSurfer + SVM	MRI	NAN

of SPECT that resulted in 74%. In [49] Vemuri et al. utilized SVM as the classification algorithm as well as feature selection technique and attained a sensitivity of 86.0% whereas 92.0% specificity in Alzheimer's disease on magnetic resonance images data. In [45] Dukart et al. utilized a meta-analysis centered on support vector machine for diagnosis of Alzheimer's disease and normal control and attain accuracy of 90%, specificity of 87.8%, and sensitivity of 91.8% on both magnetic resonance images and PET. Magnin et al.[20] utilized Histogram and Support Vector Machine on MRI to distinguish patients with Alzheimer's Disease from other normal subjects and attained accuracy, sensitivity, and specificity of 94.6%, 91.5%, and 96.6% respectively. Fan et al. [50] utilized VBM and non-linear Support Vector Machine on MRI to distinguish patients with SC from other normal subjects and attained an overall accuracy of 90.2%, and with linear SVM attained 88.5% accuracy. Geradin et al.[51] utilized Hippocampi and Support Vector Machine on MRI scans to distinguish patients with Alzheimer's Disease from other normal subjects and attained accuracy, sensitivity, and specificity of 94.6%, 96%, and 92% respectively. Hackman et al. [52] utilized Wavelet Transform and Support Vector Machine on MRI to classify patients with MC from other normal subjects and attained accuracy, sensitivity, and specificity of 80.44%, 87.80%, and 73.08% respectively. Ortiz et al. [53] utilized VAF and Support Vector Machine on MRI to distinguish patients with MC from other normal subjects and attained accuracy, sensitivity, and specificity of 71.0%, 76.0.80%, and 66.08% respectively and after VAF.

Schmitter et al.[54] utilized FreeSurfer and Support Vector Machine on MRI to distinguish patients with MC from other normal subjects and attained sensitivity and specificity, 82.80% and 88.08% respectively. Horn et al.[55] utilized SVM to distinguished patients with Alzheimer's Disease from other Frontotemporal Dementia (FTD) and attained accuracy, sensitivity, and specificity, 87.0%, 88.0%, and 87.08% respectively. Table II shows the Performance comparison of SVM over the above-mentioned techniques.

3. Logistic Regression

Logistic Regression Classification classifies the input samples to their respective classes based on the probabilistic value returned through the logistic sigmoid function. To differentiate Alzheimer's Disease from other sorts of dementia, Logistic Regression Classification is utilized in various Alzheimer's Disease analyses [56]-[58]. Logistic Regression performs classification for Alzheimer's disease MRI in a similar way to SVM [59]. To design a prediction model for the timely detection and progression of Alzheimer's Diseases, Johnson et al. [61] utilized Logistic Regression Classification (LRC). In the large feature space for finding the optimum features like neuropsychological tests, a genetic algorithm (GA) is utilized [62] [63]. These Optimal features from the Genetic Algorithm are maintained as the inputs of the Logistic Regression Classification (LRC). It appears that the Genetic Algorithm can enhance the detection of Alzheimer's disease. For the detection of different analyses of dementia, a two-level prediction model was submitted by Mazzocco and Hussain [64].

4. Linear Discriminant Analysis

Linear Discriminating Analysis (LDA) is thoroughly connected to studies of variance and regression studies which show one dependent variable as a linear combination of other features or dimensions [65]. Linear Discriminating Analysis (LDA) is also called 'Fisher Linear Discriminant ' which is the most common size reduction approach [66]. Linear Discriminating Analysis (LDA) develops a linear discriminant function resultant in minimum errors [13], [67], [68].

Zhao et al. [69] suggested an enhanced iterative trace ration (iITR) procedure to resolve the trace ratio linear discriminate analysis (TR-LDA) problematic for dementia analysis and attained improved results as compared to the principal component analysis (PCA), locality preserving projections (LLP), and maximum margin criterion (MMC). In [55] horn et al. utilized the image features reduced by the partial least square (PLS) to LDA for distinguishing AD from FTD and attained an accuracy of 84%, a sensitivity of 83%, and a specificity of 86% on perfusion SPECT images. Horn et al. [55] utilized PLS and LDA to distinguished patients with Alzheimer's Disease from other FTD by utilizing SPECT images and attained accuracy, sensitivity, and specificity, 80.4%, 83.0%, and 86.08% respectively. Zhao et al.[69] utilized KPCA and TR-LDA to distinguish patients with Alzheimer's Disease from other Normal control and attained an accuracy of 90.01%.

5. K-means Clustering (Hard Clustering)

The Clustering (grouping) approach is also known as unsupervised learning [80] [81], as the classification approaches categorized instances in dissimilar groups, but in the clustering approach (unsupervised learning) there is no training dataset to practice [82]. In

TABLE III. Machine Learning-based Approaches

Author	Targets	Methods	Imaging Modality	Accuracy
J. Akhila et al.[70]	Classification of AD	Feedforward NN	MRI	97.5%
C.V Dolph et al. [71]	Classification	SAE DNN	MRI	56%
Faturrahman et al. [72]	AD vs NC	DBN	MRI	91.7%
H.I.Suk et al. [73]	Features extraction and classification	DESRN	MRI	90.28%
E.M. Alkabawi et al. [74, 75]	Features extraction and classification	CNN + LR	MRI	74.93%
J. Akhila et al.[70]	Features extraction and classification	Feedforward NN	MRI	97.5%
Cui et al.[76]	AD diagnosis	RNN	MRI	89.7%
Wang Yen et al.[77]	AD vs NC vs MCI	CNN	MRI	92.06%
Gunawerdena et al.[78]	AD	CNN	MRI	96%
Seixas et al.[25]	MCI vs NC	Bayesian Network	MRI	NAN
Horn [55]	AD vs FTD	PLS + LDA	SPECT	84.0
		PLS + K-NN	SPECT	88.0
		SVM	SPECT	87.0
Zhao[69]	Dementia vs NC	KPCA + TR-LDA	-	90.01
Plant [21]	AD vs NC	Data Mining + SVM	MRI	90.0
	MCI vs NC	Data Mining + Bayes	MRI	85.71
Papakostas [79]	AD vs NC	VBM V LC-KNN	MRI	80.0
Liu [26]	AD vs NC	Multifold Bayesian Kernelization	MRI, PET	84.74
	MCI vs MCInc		MRI, PET	63.79

medical imaging issues and the detection of Alzheimer's disease, this clustering approach is widely utilized e.g. segmentation of brain tissue, hippocampal division, and entire cerebrum division i.e. segmentation. In [83] authors provided efficiency for various clustering procedures aimed at describing the physiognomies of cerebrum muscles for assessment of Alzheimer's disease in various phases. K-means is an unsupervised learning approach that is mainly utilized when having data without label, which means data having no definite clusters and categories. It is a famous method utilized to assemble in pre-characterized digits of K groups i.e. K-means grouping which is a hard-grouping approach. This procedure aims to find out clusters in the data set. For hippocampal division, cerebrum muscle division, and tumor affected area division (segmentation) K-means is used [79]. To differentiate subjects into pathologic groups Escudero et al. [80] used KM i.e. K-Means by distinguishing EEG temporal measures. Rodrigues et al. [81] isolated cases having AD and normal subjects by utilizing K-means clustering. They observed K-means as the best method for an unsupervised diagnosis of EEG temporal arrangements. To choose the centroid of groups in KM for cerebrum segmentation in Magnetic Resonance Images, a new approach recommended by Liu and Guo [82] is based on shifting average filtering. Papakostas et al. [83] utilized VBM and KNN on MRI to distinguished patients with MC from other normal subjects and attained accuracy, sensitivity, and specificity of 80.44%, 80.80%, and 79.08% respectively. Horn et al. [55] utilized PLS and KNN on SPECT to distinguished patients with Alzheimer's Disease from other FTD and attained accuracy, sensitivity, and specificity of 88.0%, 93.0%, and 85.08% respectively. Table III shows the performance comparison of distinct machine learning-based existing techniques.

D. Deep Learning-Based Approaches

1. Transfer Learning

Transfer Learning (TL) is a well-known Machine Learning approach in which a model that is trained can be reutilized on another related task. Because of not having enough data, numerous studies had anticipated methods to upsurge openly existing data, which are summarized in Table IV. K. Aderghal et al. in [84] proposed transfer learning from transferring the info from sMRI data to DTI images. They trained the model on sMRI with extensive distinct augmentation techniques and then transferred the info to the DTI dataset by utilizing the ADNI dataset repository for Normal subject classification, AD, and MCI. The first subject of images with just structural magnetic resonance images sMRI was chosen from the ADNI-1 repository and the second subject of images was taken from ADNI GO and ADNI. This second subject consisted of structural magnetic resonance images along with the diffusion tensor images. The second subset of images included the group of subjects with both sMRI and the DTI modalities. In this study, the hippocampus region was in focus which is a regular functional structure of the cerebrum containing two sections. To acquire just region of interest (RoI) for given AD individuals, they measured the average of two sections when floating the right region of the hippocampus to the left region. Data augmentation was used as a preprocessing step and all the experiments were done by utilizing the Café Deep learning framework. They attained classification accuracies of 92.5% for AD vs NC, 85.0% for AD vs MCI, and 80.0% for Mild Cognitive Impairment vs NC.

In [85], Afzal et al. proposed a new approach by utilizing image augmentation-based techniques to classify the AD utilizing the OASIS dataset. They performed all the experiments using transfer learning and attained 98.2% performance accuracy. In [86] N.M et al. proposed a transfer learning approach for the prediction of the binary class Alzheimer's Disease and Maqsood et al. [87] proposed a transfer learning approach by utilizing pre-trained AlexNet for multiclass classification of Alzheimer's Disease.

T. D. Phong et al. [88] established the efficacy of utilizing pre-trained models as a starting point for other networks. The other two models proposed in research, i.e. GoogNet and ResNet, are reinforced by python's TensorFlow library2 and are pre-trained with ImageNet, having great expertise to distinguish numerous types of real images. Later, those models utilized in this research were only trained for the fully connected layers of the network. S. Wang, et al. [89] utilize the TL approaches and Augmentation to resolve the inadequate samples of the data to identify Mild Cognitive Impairment on Magnetic Resonance Images. They utilize OASIS2 and their performance accuracy was 90.6% for MCI vs Normal Control. The Diffusion Tensor Images maps are frequently understood as upright modularity for AD identification. So M. A. Nowrangi et al. in [90] have compared the NC, AD, and MCI. The outputs showed MD is a good indicator. B. Cheng et al. [91] use MCI vs structural Mild Cognitive Impairment as the Supplementary domain

TABLE IV. Transfer Learning-based Approaches

Author	Method	Dataset	Images	Accuracy
Afzal et al. [85]	Transfer Learning	OASIS	MRI	98.41%
Muazzam et al. [87]	Transfer Learning	OASIS	MRI	92.85%
N. M et al [86]	Transfer Learning	OASIS	MRI	99.4%
K. Aderghal et al. [84]	Transfer Learning sMRI to DTI	ADNI	MRI	92.5% AD vs NC
				85.0% AD vs MCI
				80.0% MCI vs NC
T. D. Phong et al. [88]	Transfer Learning	115 different hospitals	CT scans	NAN
S. Wang, et al. [89]	Transfer Learning	OASIS	MRI	90.6%
B. Cheng et al. [91]	Multi-Domain Transfer Learning	ADNI	MRI	94.7%
T. Glozman et al. [92]	ImageNet Transfer Learning	ADNI	MRI	83.5%
M. Dyrba et al. [93]	FA and MD	ADNI	MRI	83%
Li et al. [94]	AD classification	ADNI	MRI	84% with adaptation

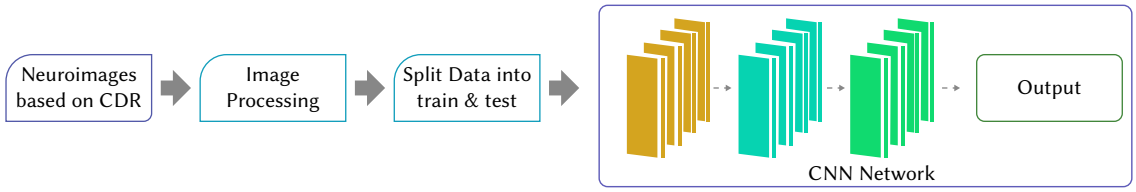


Fig. 5. The fundamental framework for the deep learning approach for the classification of AD.

for classification of Normal Control vs Alzheimer's Disease The main idea of our multi-domain transfer learning-based method is to exploit the multi-auxiliary domain data to enhance the classification of the targeted domain. They then linked the multi_domain and the target domain to enhance the selection and classification of features phase. The accuracy for AD vs NC was 94% and for MCI vs NC, they attained 82.1% classification accuracy. T. Glozman et al. in [92] utilized the pre-trained ImageNet as the source domain, by utilizing the full brain with joint classification. Only initial layers were transferred. They attained an accuracy of 83.5% for AD vs NC. M. Dyrba et al. [93] utilized FA and MD with the SVM, utilizing the info gain measure approach for AD classification. For FA, the accuracy was 80% whereas the accuracy was 83% in the case of MD. Li et al. [94] also proposed the approach for the classification of this disease for small datasets on a knowledge transfer perspective and attained accuracies of 49.0% with Tongi, 61.5% using naïve combination, 55.3% using insane weighting, and 84% using with adaptation. Table IV shows the Performance comparison of Transfer Learning for deep learning-based existing techniques. LDA was used for AD classification using different features. The results are evaluated on ADNI and 63.7% accuracy was achieved. Another study combined the features from CSF, GM, and WM and performed classification [95]. These features were used for AD stage classification over the ADNI dataset with 79.8% accuracy [96]. D. Chitradevi et al. [97] proposed a method using multiple sub-regions of the cerebral and these subregions were WM, GM, and hippocampus. Different classification methods including PSA and gray wolf optimization were used for AD classification with 98% accuracy. Hao et al. [98] performed AD stage detection using thresholding with 95% accuracy. Chihun Park et al. [99] worked on a similar algorithm for AD stage classification and achieved 82.3% accuracy. Arifa et al. [100] proposed a system using CNN and hybrid features for AD and achieved very good results.

2. Convolutional Neural Network

In [76], researchers present an MLP combinational framework, multilayer perception, and RNN identification of this disease by using MRI. Initially, MLP is used and then authors use the 2-level RNN formed on the MLP. They achieved an accuracy of 89.7% for the AD classification. For all these experiments, they use ADNI data sets.

Wang Yan et al. [77] use three-class subjects with balanced sample sizes and achieved greater precision by combining multimode magnetic resonance with the CNN core. They achieved 92.06% accuracy for this experiment. Gunawardena et al. [101] used a total of 1615 scans and reached 84.4% accuracy by utilizing the vector carrier and then proposed the CNN approach and attained an accuracy of 96%. This section discussed the application of pre-processing techniques over brain images utilized in the existing literature. Each of the techniques considering AD detection is discussed and their representative articles are also briefly overviewed. The section in its second half also discussed the utilized machine learning algorithms for efficient and accurate segmentation and classification results. All the techniques were comparatively analyzed, and their performance was represented in tabular form. A generic CNN framework is presented in Fig. 5.

V. Datasets and Softwares

Diverse freely online-available datasets and programming bundles are accessible to help. Mind picture investigation bundles, for example, FreeSurfer (surfer.nmr.mgh), FSL (https://fsl.fmrib.ox.ac.uk/), MIPAV(mipav.cit.nih.gov), and SPM, which give integral assets to various robotized pre-preparing systems [102]. Additionally, programming bundles, for example, MATLAB (mathworks.com), Keras (keras.io), Tensorflow (tensorflow.org), Theano (deeplearning.net/programming/ theano) , Caffe (caffe.berkeleyvision.org), and Torch (torch.ch) are utilized to execute profound frameworks [103]-[105] Also, freely online-available, for example, ADNI [106], AIBL [107], and OASIS [108] are exceptionally useful too. These datasets make freely functional attractive reverberation images of the brain. Fig. 6 represents the overall utilization of the tools for the given problem.

A. ADNI Datasets

The data of Alzheimer's disease Neuroimaging Initiative (ADNI), a database available at (adni.loni.ucla.edu), were used throughout in every Alzheimer's detection-related study. The dataset of ADNI was launched in 2003 by the National Institute on Aging (NIA), the Food and Drug Administration (FDA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB) private non-profits, and pharmaceutical companies of the private sector. The main role of the ADNI repository has been to analyze its images whether serial

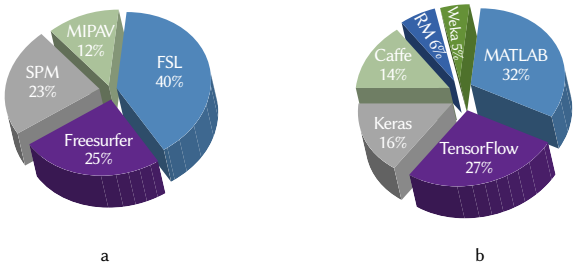


Fig. 6. (a) Pie Diagram showing the utilization ratio of different software for preprocessing (b) Pie Diagram showing the utilization ratio of different software for classification purposes.

magnetic resonance imaging (MRI), positron emission tomography (PET), other clinical and neuropsychological, and biological markers.

B. OASIS Datasets

The dataset of the Open Access Series of Imaging Studies (OASIS) is an open-source repository for the scientific community for Alzheimer's disease classification. The basic reason for the compilation of this dataset is to facilitate future discoveries in neurodegenerative disease, which is based on a multi-model-based dataset. Formally released data for OASIS-Cross-sectional and OASIS-Longitudinal have been used for hypothesis-driven data analyses, development of neuroanatomical atlases, and development of segmentation algorithms. OASIS is a neuroimaging-based data of longitudinal dimension, clinical dimension data, cognitive-based data, and biomarker dataset. This dataset is formed representing progressive stages of the disease from normal to mild, moderate to severe stages of the dataset. These stages can be distinguished in their clinical dementia rating (CDR) representing the overall classes. The OASIS dataset is hosted on which provides open access to the substantial database of processed MRI images and neuroimaging for the community. This dataset is a broad demographic, genetic spectrum, and cognitive-based dataset. This can be used for neuroimaging-based clinical and cognitive research purposes. This is a multi-staged dataset range from normal aging to cognitive decline. A broad horizon of studies in this domain is based on the OASIS dataset. The dataset can be accessed via https://www.oasis-brains.org/.

C. Other Neuroimaging Dataset

Apart from open-access datasets, the remaining studies dataset was collected from diverse open source scanners based on over the diverse sources across the spectrum for diverse Alzheimer's disease detection and other health infections [109]-[111]. The details of each respective study dataset are comprised of multiple T weighted magnetic resonance imaging (MRI), positron emission tomography (PET) scans.

Each study uses and selects a dataset according to the need and requirement of the algorithm to fully scale the problem statement under consideration. In this study two main data repositories, ADNI, and OASIS were used broadly as they are available publicly and can be used with ease. OASIS and ADNI dataset images are collected and processed with a more domain-centric approach. As the OASIS and ADNI datasets have millions of MRIs and PET images with dedicated domains, these datasets are used widely for experimental studies. This section discusses the datasets utilized in our discussed techniques in existing literature regarding the detection and classification of AD. The section also provides links to the data repositories for access to publicly available datasets.

VI. Discussion and Future Direction

In this section, we have discussed the most important features for the identification of AD. Alzheimer's disease is neurodegenerative dementia and has an impact on the fitness of an individual. It affects the parts of the brain related to memory and sensing such as the hippocampus, amygdala, parietal lobe, temporal lobe, etc. This paper efficiently presents procedures for improving AD diagnosis by using various approaches to image processing, machine learning, and deep learning. Firstly, classifying the early Mild-Cognitive-Impairments individuals from normal controls and its conversion to Alzheimer's are most significant. This timely identification at the initial stage of this disease would propose so many possible advantages to the individual. Over the last decade, numerous contributions have been done by utilizing different machine learning and deep learning approaches, but still, early detection remains an issue for AD diagnosis, instead of considering the whole brain part, some studies focus only on the Region of Interest part. This RoI-based has low feature dimension and can be easily interpreted. The performance of AD diagnosis is greatly based on the quality of the neuroimages. Neuroimaging modalities including MRI, fMRI, and PET are essential to Alzheimer's detection. This neuro-imaging identification for Alzheimer's is being a complex procedure involving countless influences that rely on these image modalities. Various other aspects, for example, age in years, gender, and education level are also useful. The most influential transfer learning models are mainly AlexNet, LeNet, and GoogleNet. There are a lot of open-source libraries that can be deployed to explore these models, as these are fast and efficient. Currently, neural network transfer learning methods are extensively used but the unavailability of enough data samples is a problem for the AD classification.

This approach is the most prominent and emerging technique in the machine learning field. The Convolutional Neural Network is another most prominent and state-of-the-art technology in AD diagnosis. These CNN approaches are widely used in many other computer-vision tasks including classification. The Convolutional Neural Network-based approaches are the most efficient and significant method for large scale issues including classification of more than 1000 classes, without handcrafted features, CNN works automatically to classify the distinct classes but the main issue in deep learning is that it requires huge training data to train the network, sometimes it is difficult to tune the CNN because of the hyperparameters. If the training data has not enough data samples then the overfitting issue may occur. The number of Alzheimer's individuals, as well as MCI, could be very limited in each dataset, which is inadequate for testing a deep learning framework. This circumstance is more terrible for multi-modality studies. Therefore, some research combines the datasets to avoid this class imbalance issue. Another way to resolve this class imbalance issue is to apply distinct augmentation approaches. Data augmentation is an efficient approach that increases the number of samples in the training set without gathering new data.

The future directions for AD classification include more localized pattern recognition to identify the changes in the brain. For this purpose, the researchers can divide the images into multiple equal parts for feature extraction and AD classification. The AD classification can be further improved by combining multiple modalities and by using reinforcement learning.

Acknowledgment

"This work was supported by the Soonchunhyang University Research Fund."

References

[1] M. A. Binnewijzend, M. M. Schoonheim, E. Sanz-Arigita, A. M. Wink, W. M. van der Flier, N. Tolboom, et al., "Resting-state fMRI changes in Alzheimer's disease and mild cognitive impairment," Neurobiology

of aging, vol. 33, pp. 2018-2028, 2012. https://doi.org/10.1016/j.neurobiolaging.2011.07.003

[2] H. Braak and E. Braak, "Neuropathological stageing of Alzheimer-related changes," Acta neuropathologica, vol. 82, pp. 239-259, 1991. https://doi.org/10.1007/BF00308809

[3] E. E. Bron, M. Smits, W. M. Van Der Flier, H. Vrenken, F. Barkhof, P. Scheltens, et al., "Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge," NeuroImage, vol. 111, pp. 562-579, 2015 https://doi.org/10.1016/j.neuroimage.2015.01.048

[4] V. Camus, P. Payoux, L. Barré, B. Desgranges, T. Voisin, C. Tauber, et al., "Using PET with 18 F-AV-45 (florbetapir) to quantify brain amyloid load in a clinical environment," European journal of nuclear medicine and molecular imaging, vol. 39, pp. 621-631, 2012. https://doi.org/10.1007/s00259-011-2021-8

[5] R. C. Petersen, G. E. Smith, S. C. Waring, R. J. Ivnik, E. G. Tangalos, and E. Kokmen, "Mild cognitive impairment: clinical characterization and outcome," Archives of neurology, vol. 56, pp. 303-308, 1999. doi:10.1001/archneur.56.3.303

[6] A. s. Disease and R. D. Association, 2017 Alzheimer's Disease Facts and Figures: Includes a Special Report on the Next Frontier of Alzheimer's Research: Alzheimer's Association, 2017. https://doi.org/10.1016/j.jalz.2017.02.006

[7] J. P. Lerch, J. Pruessner, A. P. Zijdenbos, D. L. Collins, S. J. Teipel, H. Hampel, et al., "Automated cortical thickness measurements from MRI can accurately separate Alzheimer's patients from normal elderly controls," Neurobiology of aging, vol. 29, pp. 23-30, 2008. https://doi.org/10.1016/j.neurobiolaging.2006.09.013

[8] S. Klöppel, C. M. Stonnington, C. Chu, B. Draganski, R. I. Scahill, J. D. Rohrer, et al., "Automatic classification of MR scans in Alzheimer's disease," Brain, vol. 131, pp. 681-689, 2008. https://doi.org/10.1093/brain/awm319

[9] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi, "Forecasting the global burden of Alzheimer's disease," Alzheimer's & dementia, vol. 3, pp. 186-191, 2007. https://doi.org/10.1016/j.jalz.2007.04.381

[10] R. L. Buckner, "Memory and executive function in aging and AD: multiple factors that cause decline and reserve factors that compensate," Neuron, vol. 44, pp. 195-208, 2004. https://doi.org/10.1016/j.neuron.2004.09.006

[11] G. F. Busatto, G. E. Garrido, O. P. Almeida, C. C. Castro, C. H. Camargo, C. G. Cid, et al., "A voxel-based morphometry study of temporal lobe gray matter reductions in Alzheimer's disease," Neurobiology of aging, vol. 24, pp. 221-231, 2003. https://doi.org/10.1016/S0197-4580(02)00084-2

[12] C. Cabral, P. M. Morgado, D. C. Costa, M. Silveira, and A. s. D. N. Initiative, "Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages," Computers in biology and medicine, vol. 58, pp. 101-109, 2015. https://doi.org/10.1016/j.compbiomed.2015.01.003

[13] H. Wang, S. Yan, D. Xu, X. Tang, and T. Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, 2007, pp. 1-8. DOI: 10.1109/CVPR.2007.382983

[14] M. D. Fox and M. E. Raichle, "Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging," Nature reviews neuroscience, vol. 8, p. 700, 2007. https://doi.org/10.1038/nrn2201

[15] M. D. Greicius, B. Krasnow, A. L. Reiss, and V. Menon, "Functional connectivity in the resting brain: a network analysis of the default mode hypothesis," Proceedings of the National Academy of Sciences, vol. 100, pp. 253-258, 2003. https://doi.org/10.1073/pnas.0135058100

[16] P. T. Fox, A. R. Laird, and J. L. Lancaster, "Coordinate-based voxel-wise meta-analysis: Dividends of spatial normalization. Report of a virtual workshop," Human brain mapping, vol. 25, pp. 1-5, 2005. DOI: 10.1002/hbm.20139

[17] Y. He, L. Wang, Y. Zang, L. Tian, X. Zhang, K. Li, et al., "Regional coherence changes in the early stages of Alzheimer's disease: a combined structural and resting-state functional MRI study," Neuroimage, vol. 35, pp. 488-500, 2007. https://doi.org/10.1016/j.neuroimage.2006.11.042

[18] S. Teipel, A. Drzezga, M. J. Grothe, H. Barthel, G. Chételat, N. Schuff, et al., "Multimodal imaging in Alzheimer's disease: validity and usefulness for early detection," The Lancet Neurology, vol. 14, pp. 1037-1053, 2015. https://doi.org/10.1016/S1474-4422(15)00093-9

[19] Y. Chen, D. Wolk, J. Reddin, M. Korczykowski, P. Martinez, E. Musiek, et al., "Voxel-level comparison of arterial spin-labeled perfusion MRI and FDG-PET in Alzheimer disease," Neurology, p. WNL. 0b013e31823a0ef7, 2011. DOI: https://doi.org/10.1212/WNL.0b013e31823a0ef7

[20] B. Magnin, L. Mesrob, S. Kinkingnéhun, M. Pélégrini-Issac, O. Colliot, M. Sarazin, et al., "Support vector machine-based classification of Alzheimer's disease from whole-brain anatomical MRI," Neuroradiology, vol. 51, pp. 73-83, 2009. DOIhttps://doi.org/10.1007/s00234-008-0463-x

[21] C. Plant, S. J. Teipel, A. Oswald, C. Böhm, T. Meindl, J. Mourao-Miranda, et al., "Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease," Neuroimage, vol. 50, pp. 162-174, 2010. https://doi.org/10.1016/j.neuroimage.2009.11.046

[22] J. Friedrich, R. Urbanczik, and W. Senn, "Code-specific learning rules improve action selection by populations of spiking neurons," International journal of neural systems, vol. 24, p. 1450002, 2014. https://doi.org/10.1142/S0129065714500026

[23] G. Lee, M. Kwon, S. Kavuri, and M. Lee, "Action-perception cycle learning for incremental emotion recognition in a movie clip using 3D fuzzy GIST based on visual and EEG signals," Integrated Computer-Aided Engineering, vol. 21, pp. 295-310, 2014. DOI: 10.3233/ICA-140464

[24] U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley, "A brief survey of machine learning methods and their sensor and IoT applications," in Information, Intelligence, Systems & Applications (IISA), 2017 8th International Conference on, 2017, pp. 1-8. DOI: 10.1109/IISA.2017.8316459

[25] F. L. Seixas, B. Zadrozny, J. Laks, A. Conci, and D. C. M. Saade, "A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer′ s disease and mild cognitive impairment," Computers in biology and medicine, vol. 51, pp. 140-158, 2014. https://doi.org/10.1016/j.compbiomed.2014.04.010

[26] S. Liu, Y. Song, W. Cai, S. Pujol, R. Kikinis, X. Wang, et al., "Multifold Bayesian kernelization in Alzheimer's diagnosis," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2013, pp. 303-310. DOIhttps://doi.org/10.1007/978-3-642-40763-5_38

[27] M. Catá Villá, "Feature selection methods for predicting pre-clinical stage in Alzheirmer's Disease," Universitat Politècnica de Catalunya, 2014.

[28] M. López, J. Ramírez, J. Górriz, D. Salas-Gonzalez, I. Alvarez, F. Segovia, et al., "Automatic tool for Alzheimer's disease diagnosis using PCA and Bayesian classification rules," Electronics Letters, vol. 45, pp. 389-391, 2009. DOI: 10.1049/el.2009.0176 ,

[29] F. Pereira, T. Mitchell, and M. Botvinick, "Machine learning classifiers and fMRI: a tutorial overview," Neuroimage, vol. 45, pp. S199-S209, 2009. https://doi.org/10.1016/j.neuroimage.2008.11.007

[30] V. Vapnik, The nature of statistical learning theory: Springer science & business media, 2013.

[31] G. Mirzaei, A. Adeli, and H. Adeli, "Imaging and machine learning techniques for diagnosis of Alzheimer's disease," Reviews in the Neurosciences, vol. 27, pp. 857-870, 2016. DOI: https://doi.org/10.1515/revneuro-2016-0029

[32] H. Furuta, K. Maeda, and E. Watanabe, "Application of genetic algorithm to aesthetic design of bridge structures," Computer-Aided Civil and Infrastructure Engineering, vol. 10, pp. 415-421, 1995. https://doi.org/10.1111/j.1467-8667.1995.tb00301.x

[33] J. S. Chou and A. D. Pham, "Smart artificial firefly colony algorithm-based support vector regression for enhanced forecasting in civil engineering," Computer-Aided Civil and Infrastructure Engineering, vol. 30, pp. 715-732, 2015.

[34] H. Adeli, Advances in design optimization: CRC press, 1994.

[35] O. Chapelle, V. Sindhwani, and S. S. Keerthi, "Optimization techniques for semi-supervised support vector machines," Journal of Machine Learning Research, vol. 9, pp. 203-233, 2008.

[36] V. Vural and J. G. Dy, "A hierarchical method for multi-class support vector machines," in Proceedings of the twenty-first international conference on Machine learning, 2004, p. 105. https://doi.org/10.1145/1015330.1015427

[37] J. D. Haynes and G. Rees, "Neuroimaging: decoding mental states from brain activity in humans," Nature Reviews Neuroscience, vol. 7, p. 523, 2006. DOIhttps://doi.org/10.1038/nrn1931

[38] S. Afzal, M. Javed, M. Maqsood, F. Aadil, S. Rho, and I. Mehmood, "A Segmentation-Less Efficient Alzheimer Detection Approach Using

Hybrid Image Features," in Handbook of Multimedia Information Security: Techniques and Applications, ed: Springer, 2019, pp. 421-429. DOI: https://doi.org/10.1007/978-3-030-15887-3_20

[39] S. Bauer, L.-P. Nolte, and M. Reyes, "Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization," in International Conference on Medical Image Computing and Computer-Assisted Intervention, 2011, pp. 354-361. https://doi.org/10.1007/978-3-642-23626-6_44

[40] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, pp. 91-110, 2004. https://doi.org/10.1023/B:VISI.0000029664.99615.94

[41] G. Orru, W. Pettersson-Yeo, A. F. Marquand, G. Sartori, and A. Mechelli, "Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review," Neuroscience & Biobehavioral Reviews, vol. 36, pp. 1140-1152, 2012. https://doi.org/10.1016/j.neubiorev.2012.01.004

[42] T. Smith-Vikos and F. J. Slack, "MicroRNAs circulate around Alzheimer's disease," Genome biology, vol. 14, p. 125, 2013. https://doi.org/10.1186/gb-2013-14-7-125

[43] C. Laske, T. Leyhe, E. Stransky, N. Hoffmann, A. J. Fallgatter, and J. Dietzsch, "Identification of a blood-based biomarker panel for classification of Alzheimer's disease," International Journal of Neuropsychopharmacology, vol. 14, pp. 1147-1155, 2011. https://doi.org/10.1017/S1461145711000459

[44] M. Lopez, J. Ramirez, J. Gorriz, D. Salas-Gonzalez, I. Á. Lvarez, F. Segovia, et al., "Neurological image classification for the Alzheimer's Disease diagnosis using Kernel PCA and Support Vector Machines," in Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE, 2009, pp. 2486-2489. 10.1109/NSSMIC.2009.5402069

[45] J. Dukart, K. Mueller, H. Barthel, A. Villringer, O. Sabri, M. L. Schroeter, et al., "Meta-analysis based SVM classification enables accurate detection of Alzheimer's disease across different clinical centers using FDG-PET and MRI," Psychiatry Research: Neuroimaging, vol. 212, pp. 230-236, 2013. https://doi.org/10.1016/j.pscychresns.2012.04.007

[46] Y. Zhang, S. Wang, and Z. Dong, "Classification of Alzheimer disease based on structural magnetic resonance imaging by kernel support vector machine decision tree," Progress In Electromagnetics Research, vol. 144, pp. 171-184, 2014.

[47] P. Padilla, M. López, J. M. Górriz, J. Ramirez, D. Salas-Gonzalez, and I. Álvarez, "NMF-SVM based CAD tool applied to functional brain images for the diagnosis of Alzheimer's disease," IEEE Transactions on medical imaging, vol. 31, pp. 207-216, 2012. DOI: 10.1109/TMI.2011.2167628

[48] L. K. Ferreira, J. M. Rondina, R. Kubo, C. R. Ono, C. C. Leite, J. Smid, et al., "Support vector machine-based classification of neuroimages in Alzheimer's disease: direct comparison of FDG-PET, rCBF-SPECT and MRI data acquired from the same individuals," Revista Brasileira de Psiquiatria, pp. 0-0, 2017. https://doi.org/10.1590/1516-4446-2016-2083

[49] P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, et al., "Alzheimer's disease diagnosis in individual subjects using structural MR images: validation studies," Neuroimage, vol. 39, pp. 1186-1197, 2008. https://doi.org/10.1016/j.neuroimage.2007.09.073

[50] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos, "COMPARE: classification of morphological patterns using adaptive regional elements," IEEE transactions on medical imaging, vol. 26, pp. 93-105, 2007. DOI: 10.1109/TMI.2006.886812

[51] E. Gerardin, G. Chételat, M. Chupin, R. Cuingnet, B. Desgranges, H.-S. Kim, et al., "Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging," Neuroimage, vol. 47, pp. 1476-1486, 2009. https://doi.org/10.1016/j.neuroimage.2009.05.036

[52] K. Hackmack, F. Paul, M. Weygandt, C. Allefeld, J.-D. Haynes, and A. s. D. N. Initiative, "Multi-scale classification of disease using structural MRI and wavelet transform," Neuroimage, vol. 62, pp. 48-58, 2012. https://doi.org/10.1016/j.neuroimage.2012.05.022

[53] A. Ortiz, J. M. Górriz, J. Ramírez, F. J. Martínez-Murcia, and A. s. D. N. Initiative, "LVQ-SVM based CAD tool applied to structural MRI for the diagnosis of the Alzheimer's disease," Pattern Recognition Letters, vol. 34, pp. 1725-1733, 2013. https://doi.org/10.1016/j.patrec.2013.04.014

[54] D. Schmitter, A. Roche, B. Maréchal, D. Ribes, A. Abdulkadir, M. Bach-Cuadra, et al., "An evaluation of volume-based morphometry for prediction of mild cognitive impairment and Alzheimer's disease," NeuroImage: Clinical, vol. 7, pp. 7-17, 2015. https://doi.org/10.1016/j.nicl.2014.11.001

[55] J.-F. Horn, M.-O. Habert, A. Kas, Z. Malek, P. Maksud, L. Lacomblez, et al., "Differential automatic diagnosis between Alzheimer's disease and frontotemporal dementia based on perfusion SPECT images," Artificial intelligence in medicine, vol. 47, pp. 147-158, 2009. https://doi.org/10.1016/j.artmed.2009.05.001

[56] A. Rao, Y. Lee, A. Gass, and A. Monsch, "Classification of Alzheimer's Disease from structural MRI using sparse logistic regression with optional spatial regularization," in Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, 2011, pp. 4499-4502. DOI: 10.1109/IEMBS.2011.6091115

[57] S. Kato, A. Homma, T. Sakuma, and M. Nakamura, "Detection of mild Alzheimer's disease and mild cognitive impairment from elderly speech: Binary discrimination using logistic regression," in Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, 2015, pp. 5569-5572. DOI: 10.1109/EMBC.2015.7319654

[58] X. Zhang, B. Hu, X. Ma, and L. Xu, "Resting-state whole-brain functional connectivity networks for mci classification using l2-regularized logistic regression," IEEE transactions on nanobioscience, vol. 14, pp. 237-247, 2015. DOI: 10.1109/TNB.2015.2403274

[59] R. Casanova, F.-C. Hsu, M. A. Espeland, and A. s. D. N. Initiative, "Classification of structural MRI images in Alzheimer's disease from the perspective of ill-posed problems," PloS one, vol. 7, p. e44877, 2012. https://doi.org/10.1371/journal.pone.0044877

[60] T. M. Nir, J. E. Villalon-Reina, G. Prasad, N. Jahanshad, S. H. Joshi, A. W. Toga, et al., "Diffusion weighted imaging-based maximum density path analysis and classification of Alzheimer's disease," Neurobiology of aging, vol. 36, pp. S132-S140, 2015. https://doi.org/10.1016/j.neurobiolaging.2014.05.037

[61] P. Johnson, L. Vandewater, W. Wilson, P. Maruff, G. Savage, P. Graham, et al., "Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease," BMC bioinformatics, vol. 15, p. S11, 2014. https://doi.org/10.1186/1471-2105-15-S16-S11

[62] H. G. Lee, C. Y. Yi, D. E. Lee, and D. Arditi, "An advanced stochastic time-cost tradeoff analysis based on a CPM-guided genetic algorithm," Computer-Aided Civil and Infrastructure Engineering, vol. 30, pp. 824-842, 2015. https://doi.org/10.1111/mice.12148

[63] M. Martínez-Ballesteros, J. Bacardit, A. Troncoso, and J. C. Riquelme, "Enhancing the scalability of a genetic algorithm to discover quantitative association rules in large-scale datasets," Integrated Computer-Aided Engineering, vol. 22, pp. 21-39, 2015. DOI: 10.3233/ICA-180580

[64] T. Mazzocco and A. Hussain, "Novel logistic regression models to aid the diagnosis of dementia," Expert Systems with Applications, vol. 39, pp. 3356-3361, 2012. https://doi.org/10.1016/j.eswa.2011.09.023

[65] R. A. Fisher, "The use of multiple measurements in taxonomic problems," Annals of eugenics, vol. 7, pp. 179-188, 1936.

[66] K. Fukunaga, "Introduction to statistical pattern classification," ed: Academic Press USA:, 1990. https://doi.org/10.1117/12.737157

[67] F. Nie, S. Xiang, and C. Zhang, "Neighborhood MinMax Projections," in IJCAI, 2007, pp. 993-998.

[68] S. Xiang, F. Nie, and C. Zhang, "Learning a Mahalanobis distance metric for data clustering and classification," Pattern Recognition, vol. 41, pp. 3600-3612, 2008. https://doi.org/10.1016/j.patcog.2008.05.018

[69] M. Zhao, R. H. Chan, P. Tang, T. W. Chow, and S. W. Wong, "Trace ratio linear discriminant analysis for medical diagnosis: a case study of dementia," IEEE signal processing letters, vol. 20, p. 431, 2013. DOI: 10.1109/LSP.2013.2250281

[70] J. Akhila, C. Markose, and R. Aneesh, "Feature extraction and classification of Dementia with neural network," in 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2017, pp. 1446-1450. DOI: 10.1109/RBME.2018.2886237

[71] C. V. Dolph, M. Alam, Z. Shboul, M. D. Samad, and K. M. Iftekharuddin, "Deep learning of texture and structural features for multiclass Alzheimer's disease classification," in 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2259-2266. DOI: 10.1109/IJCNN.2017.7966129

[72] M. Faturrahman, I. Wasito, N. Hanifah, and R. Mufidah, "Structural

MRI classification for Alzheimer's disease detection using deep belief network," in 2017 11th International Conference on Information & Communication Technology and System (ICTS), 2017, pp. 37-42. DOI: 10.1109/ICTS.2017.8265643

[73] H.-I. Suk, S.-W. Lee, D. Shen, and A. s. D. N. Initiative, "Deep ensemble learning of sparse regression models for brain disease diagnosis," Medical image analysis, vol. 37, pp. 101-113, 2017. https://doi.org/10.1016/j.media.2017.01.008

[74] E. M. Alkabawi, A. R. Hilal, and O. A. Basir, "Feature abstraction for early detection of multi-type of dementia with sparse auto-encoder," in 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2017, pp. 3471-3476. DOI: 10.1109/SMC.2017.8123168

[75] E. M. Alkabawi, A. R. Hilal, and O. A. Basir, "Computer-aided classification of multi-types of dementia via convolutional neural networks," in 2017 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2017, pp. 45-50. DOI: 10.1109/MeMeA.2017.7985847

[76] R. Cui, M. Liu, and G. Li, "Longitudinal analysis for Alzheimer's disease diagnosis using RNN," in Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, 2018, pp. 1398-1401. https://doi.org/10.1016/j.media.2020.101694

[77] Y. Wang, Y. Yang, X. Guo, C. Ye, N. Gao, Y. Fang, et al., "A Novel Multimodal MRI Analysis for Alzheimer's Disease Based on Convolutional Neural Network," in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 754-757. DOI: 10.1109/EMBC.2018.8512372

[78] K. Gunawardena, R. Rajapakse, and N. Kodikara, "Applying convolutional neural networks for pre-detection of alzheimer's disease from structural MRI data," in Mechatronics and Machine Vision in Practice (M2VIP), 2017 24th International Conference on, 2017, pp. 1-7. DOI: 10.1109/M2VIP.2017.8211486

[79] G. A. Papakostas, A. Savio, M. Graña, and V. G. Kaburlasos, "A lattice computing approach to Alzheimer's disease computer assisted diagnosis based on MRI data," Neurocomputing, vol. 150, pp. 37-42, 2015. https://doi.org/10.1016/j.neucom.2014.02.076

[80] F. Peng and Y. Ouyang, "Optimal clustering of railroad track maintenance jobs," Computer-Aided Civil and Infrastructure Engineering, vol. 29, pp. 235-247, 2014. https://doi.org/10.1111/mice.12036

[81] H. Wang, A. Yajima, R. Y. Liang*, and H. Castaneda, "Bayesian modeling of external corrosion in underground pipelines based on the integration of Markov chain Monte Carlo techniques and clustered inspection data," Computer-Aided Civil and Infrastructure Engineering, vol. 30, pp. 300-316, 2015. https://doi.org/10.1111/mice.12096

[82] J. Huo, Y. Gao, W. Yang, and H. Yin, "Multi-instance dictionary learning for detecting abnormal events in surveillance videos," International journal of neural systems, vol. 24, p. 1430010, 2014. https://doi.org/10.1142/S0129065714300101

[83] T. Varghese, K. R. Sheela, P. Mathuranath, and A. Singh, "Evaluation of different stages of Alzheimer's disease using unsupervised clustering techniques and voxel based morphometry," in Information and Communication Technologies (WICT), 2012 World Congress on, 2012, pp. 953-958. DOI: 10.1109/WICT.2012.6409212

[84] K. Aderghal, A. Khvostikov, A. Krylov, J. Benois-Pineau, K. Afdel, and G. Catheline, "Classification of Alzheimer Disease on Imaging Modalities with Deep CNNs Using Cross-Modal Transfer Learning," in 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS), 2018, pp. 345-350. DOI: 10.1109/CBMS.2018.00067

[85] S. Afzal, M. Maqsood, F. Nazir, U. Khan, F. Aadil, K. M. Awan, et al., "A Data Augmentation-Based Framework to Handle Class Imbalance Problem for Alzheimer's Stage Detection," IEEE Access, vol. 7, pp. 115528-115539, 2019. DOI: 10.1109/ACCESS.2019.2932786

[86] N. M. Khan, N. Abraham, and M. Hon, "Transfer Learning With Intelligent Training Data Selection for Prediction of Alzheimer's Disease," IEEE Access, vol. 7, 2019. DOI: 10.1109/ACCESS.2019.2920448

[87] M. Maqsood, F. Nazir, U. Khan, F. Aadil, H. Jamal, I. Mehmood, et al., "Transfer Learning Assisted Classification and Detection of Alzheimer's Disease Stages Using 3D MRI Scans," Sensors, vol. 19, 2019. https://doi.org/10.3390/s19112645

[88] T. D. Phong, H. N. Duong, H. T. Nguyen, N. T. Trong, V. H. Nguyen, T. Van Hoa, et al., "Brain hemorrhage diagnosis by using deep learning," in Proceedings of the 2017 International Conference on Machine Learning and Soft Computing, 2017, pp. 34-39. https://doi.org/10.1145/3036290.3036326

[89] S. Wang, Y. Shen, W. Chen, T. Xiao, and J. Hu, "Automatic recognition of mild cognitive impairment from mri images using expedited convolutional neural networks," in International Conference on Artificial Neural Networks, 2017, pp. 373-380. https://doi.org/10.1007/978-3-319-68600-4_43.

[90] M. A. Nowrangi, C. G. Lyketsos, J.-M. S. Leoutsakos, K. Oishi, M. Albert, S. Mori, et al., "Longitudinal, region-specific course of diffusion tensor imaging measures in mild cognitive impairment and Alzheimer's disease," Alzheimer's & Dementia, vol. 9, pp. 519-528, 2013. https://doi.org/10.1093/cercor/bhy031

[91] B. Cheng, M. Liu, D. Shen, Z. Li, D. Zhang, and A. s. D. N. Initiative, "Multi-domain transfer learning for early diagnosis of Alzheimer's disease," Neuroinformatics, vol. 15, pp. 115-132, 2017. https://doi.org/10.1007/s12021-016-9318-5

[92] T. Glozman, J. Solomon, F. Pestilli, and L. Guibas, "Shape-Attributes of Brain Structures as Biomarkers for Alzheimer's Disease," Journal of Alzheimer's Disease, vol. 56, pp. 287-295, 2017. DOI: 10.3233/JAD-160900

[93] M. Dyrba, M. Ewers, M. Wegrzyn, I. Kilimann, C. Plant, A. Oswald, et al., "Robust automated detection of microstructural white matter degeneration in Alzheimer's disease using machine learning classification of multicenter DTI data," PloS one, vol. 8, p. e64925, 2013. https://doi.org/10.1371/journal.pone.0064925

[94] W. Li, Y. Zhao, X. Chen, Y. Xiao, and Y. Qin, "Detecting Alzheimer's Disease on Small Dataset: A Knowledge Transfer Perspective," IEEE journal of biomedical and health informatics, 2018. DOI: 10.1109/JBHI.2018.2839771

[95] T. Altaf, S. M. Anwar, N. Gul, M. N. Majeed, and M. Majid, "Multi-class Alzheimer's disease classification using image and clinical features," Biomedical Signal Processing and Control, vol. 43, pp. 64-74, 2018. https://doi.org/10.1016/j.bspc.2018.02.019

[96] O. B. Ahmed, M. Mizotin, J. Benois-Pineau, M. Allard, G. Catheline, C. B. Amar, et al., "Alzheimer's disease diagnosis on structural MR images using circular harmonic functions descriptors on hippocampus and posterior cingulate cortex," Computerized Medical Imaging and Graphics, vol. 44, pp. 13-25, 2015. https://doi.org/10.1016/j.eswa.2016.04.029

[97] D. Chitradevi and S. Prabha, "Analysis of brain sub regions using optimization techniques and deep learning method in Alzheimer disease," Applied Soft Computing, vol. 86, p. 105857, 2020. https://doi.org/10.1016/j.asoc.2019.105857

[98] X. Hao, Y. Bao, Y. Guo, M. Yu, D. Zhang, S. L. Risacher, et al., "Multi-modal neuroimaging feature selection with consistent metric constraint for diagnosis of Alzheimer's disease," Medical Image Analysis, vol. 60, p. 101625, 2020. https://doi.org/10.1016/j.media.2019.101625

[99] C. Park, J. Ha, and S. Park, "Prediction of Alzheimer's disease based on deep neural network by integrating gene expression and DNA methylation dataset," Expert Systems with Applications, vol. 140, p. 112873, 2020. https://doi.org/10.1016/j.eswa.2019.112873

[100] A. Shikalgar and S. Sonavane, "Hybrid Deep Learning Approach for Classifying Alzheimer Disease Based on Multimodal Data," in Computing in Engineering and Technology, ed: Springer, 2020, pp. 511-520. https://doi.org/10.1007/978-981-32-9515-5_49

[101] A. Giersch and J. T. Coull, "TRF1: It Was the Best of Time (s)...," Timing & Time Perception, vol. 6, pp. 231-414, 2018. https://doi.org/10.1163/22134468-00603001

[102] S. Leandrou, S. Petroudi, P. A. Kyriacou, C. C. Reyes-Aldasoro, and C. S. Pattichis, "Quantitative MRI brain studies in mild cognitive impairment and Alzheimer's disease: a methodological review," IEEE reviews in biomedical engineering, vol. 11, pp. 97-111, 2018. DOI: 10.1109/RBME.2018.2796598

[103] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, et al., "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60-88, 2017. https://doi.org/10.1016/j.media.2017.07.005

[104] D. Ravì, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, et al., "Deep learning for health informatics," IEEE journal of biomedical and health informatics, vol. 21, pp. 4-21, 2016. DOI: 10.1109/ICICI.2017.8365301

[105] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect

recognition: insights and new developments," IEEE Transactions on Affective Computing, 2019. DOI: 10.1109/TAFFC.2018.2890471

[106] C. R. Jack Jr, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, et al., "The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods," Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol. 27, pp. 685-691, 2008. https://doi.org/10.1002/jmri.21049

[107] K. A. Ellis, A. I. Bush, D. Darby, D. De Fazio, J. Foster, P. Hudson, et al., "The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease," International psychogeriatrics, vol. 21, pp. 672-687, 2009. https://doi.org/10.1017/S1041610209009405

[108] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults," Journal of cognitive neuroscience, vol. 19, pp. 1498-1507, 2007. https://doi.org/10.1162/jocn.2007.19.9.1498

[109] A. Sedik, A. M. Iliyasu, A. El-Rahiem, M. E. Abdel Samea, A. Abdel-Raheem, M. Hammad, et al., "Deploying Machine and Deep Learning Models for Efficient Data-Augmented Detection of COVID-19 Infections," Viruses, vol. 12, p. 769, 2020, https://doi.org/10.3390/v12070769

[110] A. Alghamdi, M. Hammad, H. Ugail, A. Abdel-Raheem, K. Muhammad, H. S. Khalifa, et al., "Detection of myocardial infarction based on novel deep transfer learning methods for urban healthcare in smart cities," Multimedia Tools and Applications, pp. 1-22, 2020, https://doi.org/10.1007/s11042-020-08769-x

[111] S. Toraman, T. B. Alakus, and I. Turkoglu, "Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks," Chaos, Solitons & Fractals, vol. 140, p. 110122, 2020, https://doi.org/10.1016/j.chaos.2020.110122

Sitara Afzal

Sitara Afzal has completed her MS from COMSATS University Islamabad, Attock campus. Her research interests are machine learning and image processing.

Muazzam Maqsood

Muazzam Maqsood is serving as an Assistant Professor at COMSATS University Islamabad, Attock Campus, Pakistan. He has earned his Ph.D. from UET Taxila in 2017. His research Interests includes machine learning, recommender systems, and image processing

Umair Khan

Umair Khan is pursuing his Ph.D. from Italy. He has done his M.S. in Computer System Engineering from UET Peshawar, Pakistan. He is also a lecturer with COMSATS University Islamabad at Attock. His research concentrates on advanced concepts of image processing and machine learning.

Irfan Mehmood

Irfan Mehmood is Assistant Professor of Applied Artificial Intelligence, Faculty of Engineering & Informatics, School of Media, Design, and Technology, University of Bradford, UK. His areas of interest are multimedia analytics, information mining, and summarization. Specifically, he has made a significant contribution in the areas of visual surveillance, information mining, and data encryption.

Hina Nawaz

Hina Nawaz is currently pursuing a master's degree with COMSATS University Islamabad, Attock campus. Her research interests are machine learning and image processing.

Farhan Aadil

Farhan Aadil received his B.S. degree in Computer Science from Allama Iqbal Open University, Pakistan in 2005. He pursued a career in computer science for 4 years (2005 to 2009). He received his M.S. degree in Software Engineering and the Ph.D. degree in Computer Engineering from the University of Engineering and Technology, Taxila, Pakistan in 2011 and 2016 respectively. He is currently working as an assistant professor at COMSATS University Islamabad, Attock Campus, Pakistan. His research interests include Machine Learning, Pattern recognition, and bio-inspired algorithms.

Oh-Young Song

Oh-Young Song is serving as an associate professor in the Department of Software at Sejong University in Korea. His research is in the area of computer graphics. He is particularly interested in physics-based animation, character animation, numerical algorithms, VR/AR, HCI, and machine learning.

Yunyoung Nam

Yunyoung Nam received the B.S., M.S., and Ph.D. degrees in computer engineering from Ajou University, Korea in 2001, 2003, and 2007 respectively. He was a Senior Researcher with the Center of Excellence in Ubiquitous System, Stony Brook University, Stony Brook, NY, USA, from 2007 to 2010, where he was a Postdoctoral Researcher, from 2009 to 2013. He was a Research Professor with Ajou University, from 2010 to 2011. He was a Postdoctoral Fellow with the Worcester Polytechnic Institute, Worcester, MA, USA, from 2013 to 2014. He was the Director of the ICT Convergence Rehabilitation Engineering Research Center, Soonchunhyang University, from 2017 to 2020. He has been the Director of the ICT Convergence Research Center, Soonchunhyang University, since 2020, where he is currently an Assistant Professor with the Department of Computer Science and Engineering. His research interests include multimedia databases, ubiquitous computing, image processing, pattern recognition, context-awareness, conflict resolution, wearable computing, intelligent video surveillance, cloud computing, biomedical signal processing, rehabilitation, and healthcare systems.

Imputation of Rainfall Data Using the Sine Cosine Function Fitting Neural Network

Po Chan Chiu[1,2*], Ali Selamat[1,3,4*], Ondrej Krejcar[4], King Kuok Kuok[5], Enrique Herrera-Viedma[6], Giuseppe Fenza[7]

[1] School of Computing, Faculty of Engineering & MagicX (Media and Games Center of Excellence), Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor (Malaysia)
[2] Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak (Malaysia)
[3] Malaysia Japan International Institute of Technology (MJIIT), Universiti Teknologi Malaysia Kuala Lumpur, Jalan Sultan Yahya Petra, 54100 Kuala Lumpur (Malaysia)
[4] Faculty of Informatics and Management, University of Hradec Kralove, Rokitanského 62, 500 03 Hradec Kralove (Czech Republic)
[5] Faculty of Engineering, Computing and Science, Swinburne University of Technology Sarawak Campus, 93350 Kuching, Sarawak (Malaysia)
[6] Andalusian Research Institute Data Science and Computational Intelligence, University of Granada, 18071 Granada (Spain)
[7] Dipartimento di Scienze Aziendali-Management & Innovation Systems (DISA-MIS), University of Salerno, 84084 Fisciano (Italy)

Received 13 September 2020 | Accepted 2 July 2021 | Published 12 August 2021

Abstract

Missing rainfall data have reduced the quality of hydrological data analysis because they are the essential input for hydrological modeling. Much research has focused on rainfall data imputation. However, the compatibility of precipitation (rainfall) and non-precipitation (meteorology) as input data has received less attention. First, we propose a novel pre-processing mechanism for non-precipitation data by using principal component analysis (PCA). Before the imputation, PCA is used to extract the most relevant features from the meteorological data. The final output of the PCA is combined with the rainfall data from the nearest neighbor gauging stations and then used as the input to the neural network for missing data imputation. Second, a sine cosine algorithm is presented to optimize neural network for infilling the missing rainfall data. The proposed sine cosine function fitting neural network (SC-FITNET) was compared with the sine cosine feedforward neural network (SC-FFNN), feedforward neural network (FFNN) and long short-term memory (LSTM) approaches. The results showed that the proposed SC-FITNET outperformed LSTM, SC-FFNN and FFNN imputation in terms of mean absolute error (MAE), root mean square error (RMSE) and correlation coefficient (R), with an average accuracy of 90.9%. This study revealed that as the percentage of missingness increased, the precision of the four imputation methods reduced. In addition, this study also revealed that PCA has potential in pre-processing meteorological data into an understandable format for the missing data imputation.

Keywords

Imputation, Missing Rainfall Data, Principal Component Analysis (PCA), Sine Cosine Neural Network, Deep Learning.

DOI: 10.9781/ijimai.2021.08.013

I. Introduction

RAINFALL is a critical component of the hydrological cycle. Numerous hydrological research areas, such as flood forecasting [1], flood risk assessment [2], rainfall forecasting [3], climate variability analysis [4], and water resources modeling [5], require reliable and complete rainfall data series. However, hydrological data analysis is challenging due to the presence of missing rainfall data. For this reason, data imputation has attracted a great deal of attention from researchers to fill in the missing values with approximations. The traditional imputation approaches include listwise deletion [6], arithmetic mean and median imputation [7], and multiple imputations [8]. However, these methods are time-consuming and less accurate [9].

In recent years, numerous artificial neural network (ANN) studies have used historical rainfall data series from nearest neighbor stations to treat the problems of missing data [10]-[12]. More efficient algorithms, such as the Levenberg-Marquardt backpropagation algorithm [13], the Gaussian mixture model-based K-nearest neighbor (GMM-KNN) algorithm [14], and the Bayesian principal component analysis (BPCA) [15] have been applied to impute the missing values in water resource engineering.

* Corresponding author.
E-mail addresses: pcchiu@unimas.my (P. C. Chiu), aselamat@utm.my (A.Selamat).

Although ANNs have been applied to treat the problem of missing data, ANNs tend to be trapped in local optima as it smoothly converges towards local minima rather than global minima. To overcome this, several novel approaches have been combined with ANNs to improve the performance of the estimation results. The sine cosine algorithm (SCA) is a metaheuristic technique developed by Mirjalili [16] to solve optimization problems using the sine and cosine trigonometric functions. SCA has been successfully applied in modal dimensional [17], short-term hydrothermal scheduling [18], support vector regression [19], and the traveling salesman problem [20]. To the best of the authors' knowledge, there is no existing sine cosine neural network that focuses on missing rainfall data imputation.

Furthermore, the use of raw hourly rainfall data from nearest neighbor stations could be unreliable for the prediction of the missing data of the target station. The long dry periods contain long sequences of zero values at the beginning, middle, or end of the records, in which rain does not usually fall every hour. Modeling long dry rainfall periods poses challenges such as underestimation or overestimation of the length of long dry periods [21] and [22]. As a result, a neural network is not able to estimate the missing rainfall value based on hourly rainfall datasets accurately. Hence, the hourly rainfall dataset needs to be combined with other non-precipitation data for the estimation of missing rainfall data.

According to Kashiwao et al. [23], rainfall is caused by a variety of meteorological conditions, and the mathematical model for it is non-linear. The meteorological data have different units of measurement and accuracy. Thus, the meteorological data need to be pre-processed prior to imputation. Normalization is the most commonly used approach. Yen [24] applied a *mapminmax* approach to normalizing the meteorological parameters in the study, while Chhetri et al. [25] normalized the weather parameters using a min-max scaler. In addition, Grange [26] proposed using a random forest machine learning algorithm for meteorological normalization to detect interventions in an air quality time series. According to Kashiwao et al. [23], the investigation into the method used to choose meteorological data is needed because suitable data can vary among prediction points due to the difference in the effect of conditions, such as altitude, ocean current, and airflow. For this reason, this paper proposes using principal component analysis (PCA) as a novel pre-processing mechanism to extract the core relationships in the meteorological data. PCA is used to identify patterns in data and express the similarities and differences of the data [27]. PCA has been used in many studies to isolate independent factors (principal components) that significantly explain the variation of a dependent variable [28]-[32]. However, the compatibility of both non-precipitation and precipitation as input has been given less attention in previous studies. Therefore, we propose using PCA as a novel pre-processing tool for meteorological data and introduce the combination of significant principal components (PCs) and rainfall data from nearest neighbor gauging stations as the input for the estimation of missing rainfall values.

The contributions of this paper are the following:

- To introduce a pre-processing mechanism for non-precipitation data by using principal component analysis (PCA).
- To propose a sine cosine function fitting neural network (SC-FITNET) imputation that focuses on missing time series data.
- To evaluate the performances of sine cosine function fitting neural network (SC-FITNET) imputation with the state-of-art models for infilling missing rainfall values at different percentages of missingness.

II. Methodology

The proposed methodology employed in this study consists of two main phases, as shown in Fig. 1. The phases are the data preparation phase and the missing data imputation phase.

A. Phase 1: Data Preparation

In this study, the data preparation phase attempts to transform raw data into an understandable format prior to the missing data imputation. The data preparation phase involves data pre-processing and data integration. Due to the variety of measurement units, the raw meteorological data must be pre-processed. For example, the values of mean surface wind (direction) are stored at 00°, 010°, ..., 058°. These characters are considered noise in the data because the neural network could not understand and interpret those characters accurately.

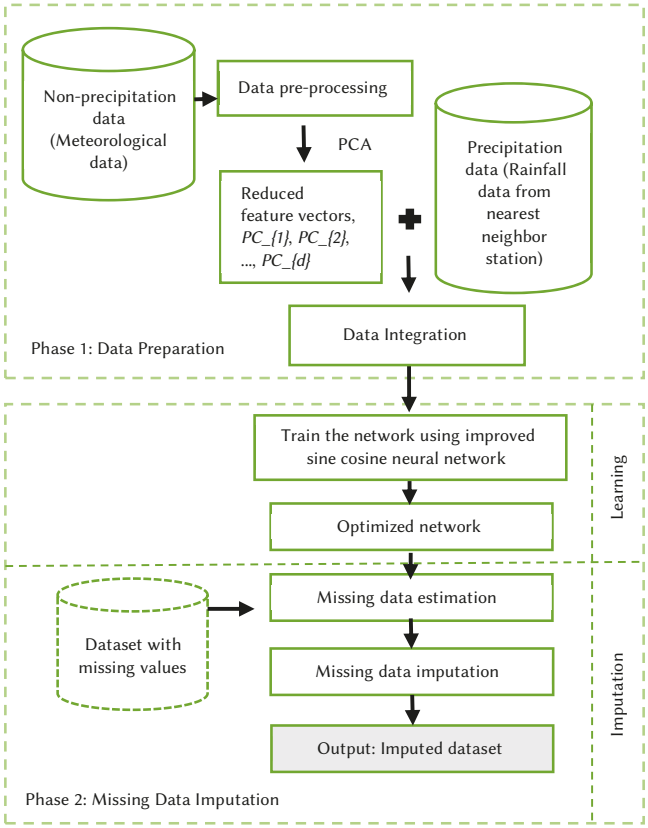


Fig. 1. The proposed methodology of missing data imputation.

In the related literature, the advantages of PCA are able to reveal hidden structure in the dataset, detect outliners, and filter out the noise in data [33]. In addition, PCA is one of the most used approaches to pre-process the weather [28] and meteorological data [30]. Therefore, PCA was used to pre-process the raw meteorological data.

PCA was proposed by Pearson [34] and formalized by Hotelling [35]. Using PCA, these meteorological data were transformed into a smaller number of variables. PCA reduces the number of meteorological features by constructing a new and smaller number of variables that capture a significant portion of the original meteorological features. The pre-process of meteorology data starts with normalizing the variables by subtracting the mean from each data point. Next, the covariance and correlation between every pair of variables (meteorological features) were calculated based on the following equations [27] and [36]:

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} \quad (1)$$

where, $cov(x, y)$ is the covariance of the variables x and y, x_i and y_i are the independent variable of observations, $\bar{x}$ and $\bar{y}$ are the mean values of the variables x_i and y_i, respectively and n is the number of data points in the observations.

$$r(x,y) = \frac{cov(x,y)}{s_x s_y} \tag{2}$$

where, $r(x, y)$ is the correlation of the variables x and y, s_x is the sample standard deviation of the random variable x, and s_y is the number of data points in the observations.

Then eigenvector and eigenvalue of the matrix are obtained as follows:

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,n} \end{bmatrix} \tag{3}$$

$$Av = \lambda v \tag{4}$$

The eigenvector v of each variable can be obtained by identifying the determinant of its characteristic polynomial as follows:

$$(A - \lambda I)v = 0 \tag{5}$$

The eigenvalue can be formulated using the following Equation:

$$p(\lambda) = |A - \lambda I| \tag{6}$$

After these steps, the principal components (*PC_{1}, PC_{2}, …, PC_{d}*) can be determined. The first principal component accounts for the highest variance in the meteorological dataset, followed by the second principal component for the next highest variance. This continues until the total of the principal components is equal to the number of features in the meteorological dataset.

The last step is to compute the feature vector. A matrix M of dimensions $n \times d$ is represented as

$$M = \begin{bmatrix} f_{1,1} & f_{1,2} & f_{1,3} & \cdots & f_{1,d} \\ f_{2,1} & f_{2,2} & f_{2,3} & \cdots & f_{2,d} \\ f_{3,1} & f_{3,2} & f_{3,3} & \cdots & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{n,1} & f_{n,2} & f_{n,3} & \cdots & f_{n,d} \end{bmatrix} \tag{7}$$

where, f_{ij} is a reduced feature vector from $n \times n$ original data to size $n \times d$, n is the number of data points in the observations, and d is the number of principal components.

The final output of the PCA is combined with the raw rainfall data from the nearest neighbor gauging stations and then used as the input to the neural network for missing data imputation.

B. Phase 2: Missing Data Imputation

The missing data imputation phase consists of two sub-phases, namely learning and imputation. In the learning sub-phase, the combined dataset from phase 1 will be used as an input to the neural network training. By using the ANN approach, the neural network is trained and optimized to learn the complex and non-linear relationships between the features in the dataset. The output of the learning sub-phase is an optimized network with a set of optimal network weights and biases. Next, the imputation sub-phase involves missing data estimation using the optimized network. During the missing data imputation, the estimated missing data are imputed into the missing values in the dataset. Hence, the final output of this phase is the imputed database.

III. Imputation Methods

Artificial neural networks (ANNs) based rainfall and runoff (R-R) modeling were first applied in the early 1990s. ANNs learn complex and non-linear relationships that are difficult to model using statistical approaches. Hence, in this study, four ANN models are employed to estimate the missing time series values.

A. Feedforward Neural Network (FFNN)

The feedforward neural network (FFNN) model is the simplest type of ANNs [37]. The architecture of the FFNN network consists of p -many inputs (input neurons), a single hidden layer with q-many hidden neurons, and a single output. A simulation for estimation of the missing rainfall data using FFNN was carried out with ten neurons in the hidden layer. The activation functions for the hidden layer and output layer are tan-sigmoid and purelin, respectively.

B. Sine Cosine Function Fitting Neural Network (SC-FITNET)

The function-fitting neural network (FITNET) is a feedforward network that forms a generalization of the input and output relationship. FITNET produces an associated set of target outputs, with tan-sigmoid transfer function in the hidden layers and linear transfer function in the output layer. The FITNET model was trained with two hidden layers; a first hidden layer with 15 neurons and a second layer with three neurons.

To improve the performance of missing data prediction, the FITNET model is optimized by the sine cosine algorithm (SCA). The improved neural network is therefore named as sine cosine function fitting neural network, abbreviated as SC-FITNET. The sine cosine algorithm (SCA) is a metaheuristic optimization technique introduced by Mirjalili [16] to solve continuous optimization problems. One of the most significant advantages of SCA is its simplicity, as reported by Qu et al. [17]. SCA has fewer parameters that need to be fine-tuned compared to other algorithms. The capability of SCA in missing rainfall data imputation has not yet been explored. Hence, the SCA is employed to train the FITNET model for missing data prediction.

First, the network is trained using a function-fitting neural network to identify and learn the relationships between features in the dataset. Then, the SCA is employed to optimize the search solutions by determining the optimal network weights and biases.

SCA starts the optimization process with a set of search solutions, X. The set of search solutions is initialized randomly and repeatedly evaluated by an objective function. The objective of the training is to minimize the prediction error. The evaluation of the training is measured by the mean square error (MSE) as follows [38]:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y - \tilde{y})^2 \tag{8}$$

where N is the number of observations, y is the actual value, and $\tilde{y}$ is the predicted value.

Next, the search solution is improved by the position-updating function in Equation (9)[16]. The SCA updates the best solutions obtained and denotes it as a destination point, P.

$$X_i^{t+1} = \begin{cases} X_i^t + r1 * \sin(r_2) * \left| r_3 P_i^t - X_{i,}^t \right|, & r_4 < 0.5 \\ X_i^t + r1 * \cos(r_2) * \left| r_3 P_i^t - X_{i,}^t \right|, & r_4 \geq 0.5 \end{cases} \tag{9}$$

where, X_i is the position vector of the current solution in the i^{th} dimension, t is the current iteration, P_i is the destination solution, r_1, r_2, r_3, r_4 are random variables, and the r_4 value is between 0 and 1.

As seen in (9), there are four parameters in SCA, namely r_1, r_2, r_3 and r_4. The parameter r_1 is the movement direction parameter that determines the region of the next solution, which is updated using (10). The parameter r_2 identifies the movement of forwards or outwards P_i

within the value of 0 and 2π. Next, the parameter r_3 is the random weights of P_i with a value either less than 1 or greater than 1. The parameter r_4 is used to switch between the sine and cosine functions.

$$r_1(t) = a * \left(1 - \frac{t}{t_{max}}\right) \tag{10}$$

where, t is the current iteration, t_{max} is the maximum iteration of SCA, and a is a constant.

As the iteration of SCA increases, the ranges of sine and cosine in the position-updating functions are updated to optimize the local search, as shown in Line 7 of Algorithm 1. Then, the best network weights and biases are updated to improve the network model. The execution of the search solution will be halted if the network has achieved the minimum error or reached the maximum network epochs. Next, given the optimized network, the network model is tested with another dataset of the same format to predict the missing rainfall data. Then, the estimated missing rainfall data are imputed into the missing dataset. The proposed SC-FITNET imputation is presented in Algorithm 1.

Algorithm 1: The proposed sine cosine function fitting neural network (SC-FITNET) imputation

Input: Pre-processed meteorology and nearest neighbor rainfall

1. **Do**
2. Select random search agents (solutions) (X) and SCA parameters (r_1, r_2, r_3 and r_4)
3. **Do**
4. Evaluate each of the search agents by the objective function
5. Update the best solution obtained so far (P)
6. Update the parameters r_1, r_2, r_3 and r_4
7. Update the position of search agents using Equation (9)
8. **While** (t < maximum number of iterations)
9. **Return** the best solution (P) obtained as the global optimum solution
10. Track the best network
11. Update training state
12. **While** (MSE > the minimum error) or (epoch < maximum number of epochs)
13. Use the optimized network
14. Train the optimized net for another dataset of the same format
15. **Do**
16. Impute the estimated values into the missing value
17. **While** (there is missing value)

Output: Imputed rainfall dataset

Note: The algorithm in the dotted line box was adapted from Mirjalili [16]

In addition, different values of the parameters are introduced to the SC-FITNET. The parameters are tuned based on the try and error method. The parameter settings are outlined in Table I.

TABLE I. The SC-FITNET Parameters

Parameters for SC-FITNET	Value
a	2
Search agents	30
Max number of epochs	1000
Max iteration of SCA	500

C. Sine Cosine Feedforward Neural Network (SC-FFNN)

The third model evaluated was a sine cosine feedforward neural network (SC-FFNN). The adaptation of the sine cosine algorithm into the feedforward neural network is employed to improve the accuracy of missing rainfall data imputation. The model was trained with ten neurons in the hidden layer. The SC-FFNN applied the same parameters setting, as in Table I.

D. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) is a recurrent deep neural network model [39]. Recent studies have successfully applied LSTM based deep learning models for time series forecasting [40], data augmentation [41] and sequence labeling [42]. Hence, we developed a LSTM multivariate time series model to predict the missing values. The LSTM model consists of five layers; an input layer, two layers of LSTM, a fully connected dense layer, and an output layer, as illustrated in Fig. 2. The two LSTM layers are employed to model the time series relationship, while the fully connected layer takes the output of the LSTM layers to a final missing data prediction.

After data pre-processing, the data are reshaped into a multivariate format for the LSTM models. The activation function used in this model was the default tanh, Adam optimizer, 20 epochs of training with a batch size of 32, GPU execution environment, two hidden layers of 120 neurons each and one-time delay handling the prediction of missing time series.

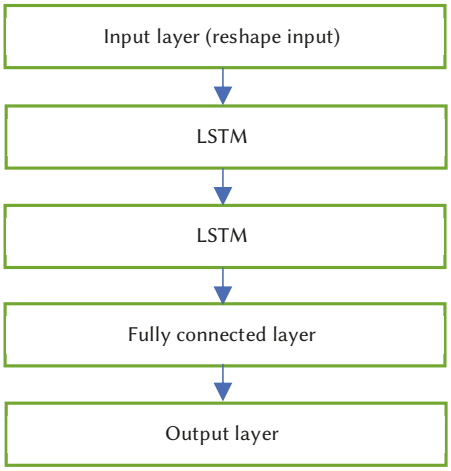


Fig. 2. The architecture of a LSTM network for multivariate time series prediction.

IV. Materials and Methods

A. Study Area

The selected study area for this study is Sungai Merang, or the Merang River gauge station, approximately 80 km from Kuching City, Sarawak, Malaysia. Sungai Merang is one of the five rainfall gauge stations in the Bedup River catchment, as shown in Fig. 3. Its nearest neighbor gauge stations over the basin are Bukit Matuh (BM), Semuja Nonok (SN), Sungai Busit (SB) and Sungai Teb (ST). The surface areas of the five rainfall gauge stations are SM: 8.550 km^2; BM: 8.075 km^2; SN: 7.600 km^2, SB: 8.075 km^2 and ST: 15.320 km^2.

The primary vegetation in this area is paddy and fruit plantation. The area is mostly covered with clayey soils and partly covered with coarse loamy soil. The soil texture enhances the infiltration rate but reduces the surface runoff. Hence, the water supply plan for paddy irrigation is crucial and extremely important for the village. However, the water supply plan and hydrological data analysis are challenging due to the presence of missing rainfall values at the Sungai Merang gauge station. Therefore, this study focuses on the missing rainfall data imputation at that gauge station.

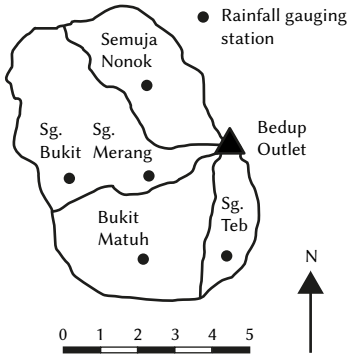


Fig. 3. Sungai Merang and its nearest neighbor gauging stations [3].

B. Meteorological Data

The meteorological data for Kuching station was acquired from the Malaysian Meteorological Department, as shown in Table II [43]. In this study, ten types of meteorological data were collected: date, time, the pressure at mean sea-level (MSL), dry-bulb temperature, relative humidity, mean surface wind (direction), mean surface wind (speed), rainfall duration, rainfall amount and cloud cover.

TABLE II. Meteorological Data From Kuching Station, Sarawak

Meteorological data	Measurement Unit
Date	YYMMDD
Time	MST
Pressure MSL	Hpa
Dry Bulb Temperature	° C
Relative Humidity	%
Mean Surface Wind (direction)	°
Mean Surface Wind (speed)	m/s
Rainfall Duration	min
Rainfall Amount	mm
Cloud Cover (cloud amount)	Oktas

C. Rainfall From Nearest Neighbor Stations

The rainfall data from the Sungai Merang gauging station and its nearest neighbor gauging stations were collected from the Department of Irrigation and Drainage, Sarawak, as shown in Table III [44]. Overall, the correlation coefficients between the Sungai Merang station and each of the neighbor stations are greater than 0.8 and located within a radius range of 5 km. Since the Sungai Merang gauging station exhibits a high correlation coefficient with its nearest neighbor stations, the complete rainfall data series from the four neighbor stations of the corresponding hour, day, month and year are used to predict the missing values of Sungai Merang's rainfall data. Based on the availability of continuous and complete data (without missing values) for the five gauging stations, this study analyzed the observed hourly rainfall data from the year 2002 until 2003. With a sample size of 11,680 complete records, the neural networks were trained with a training length of 8180 and tested with datasets of 3500 records. In [45]-[48], the data were randomly deleted and removed from the testing datasets. Hence, for the preparation of missing values in rainfall data, this study employed a rate-based approach [49] in which 10%, 20%, 30%, 40%, and 50% were randomly removed from the testing datasets. In total, two sets of testing data were prepared for each percentage of the missingness. In this study, the missing data were categorized as missing completely at random (MCAR) [50] because the presence of missing rainfall data at the Sungai Merang gauge station is not affected by the data in that area or any nearby area.

TABLE III. The Sungai Merang Gauging Station and Its Nearest Neighbor Gauging Stations

Station Name	Latitude	Longitude	Distance from Sg Merang (km)	Correlation Coefficient
Sungai Merang	001 05 40	110 36 25	-	-
Bukit Matuh	001 03 50	110 35 35	3.88	0.8558
Semuja Nonok	001 06 25	110 35 50	2.10	0.8647
Sungai Busit	001 05 25	110 34 40	3.44	0.8676
Sungai Teb	001 03 15	110 37 00	4.37	0.8046

D. Data Input Description

The number of data inputs, p, to the missing data imputation model was based on the number of cumulative principal components (*PC_{1}*, *PC_{2}*, …, *PC_{d}*) and raw rainfall data from the nearest neighbor stations.

$$input\ p\{d\} = cummulative\ of\ PC\_\{d\}, NNS1, NNS2, NNS3, NNS4 \quad (11)$$

where, PC is the principal component (s), $\{d\}$ is the number of principal components, and $NNS1$, $NNS2$, $NNS3$, $NNS4$ are the complete rainfall from the four nearest neighbor stations (NNS).

E. Performance Measures

The performances of the two imputation methods are measured by the mean absolute error (MAE), root mean square error (RMSE), and correlation coefficient (R).

- Mean absolute error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|O_i - T_i| \quad (12)$$

- Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(O_i - T_i)^2}{N}} \quad (13)$$

- The correlation coefficient (R)

$$R = \frac{\sum_{i=1}^{N}(T_i - \bar{T})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N}(T_i - \bar{T})^2 (O_i - \bar{O})^2}} \quad (14)$$

where N is the total number of observations, O_i is the actual values of observations, $\bar{O}$ is the mean values of the actual observations, T_i is the imputed values, and $\bar{T}$ is the mean of the imputed values.

V. Experiment

The proposed SC-FITNET missing data imputation was compared with the FFNN imputation, the SC-FFNN imputation and LSTM multivariate time series imputation using a combination input p of the meteorological data series (cumulative PC) and rainfall data series from nearest neighbor stations. A different number of inputs p was introduced, from $p1$ to $p10$, to determine the significant input p to the neural network. The average result gave the minimum MAE and RMSE measures, but the highest measure of R was chosen as the significant input p. For better evaluation of the proposed algorithm, we tested the imputation algorithms on two missing datasets. For each missing dataset, all the imputation algorithms were executed with 30 independent runs over each input p at different missing data rates (10%, 20%, 30%, 40%, and 50%). The average values of the performance measures for FFNN, SC-FFNN, SC-FITNET, and LSTM imputation, respectively, over two missing datasets, are presented in the following sub-sections.

TABLE IV. Comparison of MAE, RMSE, and R Values for FFNN Imputation at Different Percentages of Missingness

Input	MAE (mm)						RMSE (mm)						R					
P	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG
P1	0.113	0.217	0.275	0.398	0.497	0.300	1.029	1.454	1.338	1.741	1.976	1.508	0.920	0.834	0.863	0.789	0.711	0.823
P2	0.120	0.237	0.316	0.445	0.572	0.338	0.994	1.587	1.495	1.810	2.517	1.681	0.929	0.861	0.882	0.806	0.761	0.848
P3	0.134	0.271	0.362	0.514	0.647	0.386	1.043	1.629	1.515	2.017	2.350	1.711	0.925	0.855	0.872	0.803	0.754	0.842
P4	**0.108**	**0.212**	**0.280**	**0.397**	**0.497**	**0.299**	**0.871**	**1.210**	**1.103**	**1.435**	**1.638**	**1.251**	**0.947**	**0.896**	**0.914**	**0.856**	**0.808**	**0.884**
P5	0.148	0.286	0.387	0.533	0.686	0.408	1.186	1.757	1.678	1.887	2.427	1.787	0.919	0.865	0.884	0.818	0.774	0.852
P6	0.153	0.316	0.424	0.573	0.762	0.446	1.160	2.658	2.585	2.349	3.949	2.540	0.920	0.862	0.877	0.812	0.764	0.847
P7	0.150	0.298	0.406	0.571	0.708	0.426	1.003	1.424	1.365	1.724	1.920	1.487	0.927	0.855	0.871	0.808	0.759	0.844
P8	0.148	0.290	0.391	0.549	0.690	0.413	1.070	1.513	1.393	1.773	2.061	1.562	0.917	0.841	0.868	0.792	0.723	0.828
P9	0.164	0.345	0.452	0.610	0.786	0.472	1.219	2.891	2.783	2.285	3.547	2.545	0.897	0.795	0.820	0.752	0.680	0.789
P10	0.166	0.313	0.424	0.588	0.742	0.447	1.056	1.484	1.381	1.767	2.024	1.542	0.919	0.847	0.868	0.798	0.727	0.832

TABLE V. Comparison of MAE, RMSE, and R Values for SC-FFNN Imputation at Different Percentages of Missingness

Input	MAE (mm)						RMSE (mm)						R					
P	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG
P1	**0.072**	**0.135**	**0.163**	**0.244**	**0.301**	**0.183**	0.831	**1.108**	**0.952**	**1.344**	**1.473**	**1.142**	0.951	0.912	**0.936**	**0.874**	**0.842**	**0.903**
P2	0.087	0.165	0.214	0.310	0.388	0.233	**0.830**	1.108	1.014	1.375	1.534	1.172	**0.952**	**0.914**	0.930	0.870	0.837	0.901
P3	0.106	0.210	0.280	0.400	0.503	0.300	0.887	1.214	1.126	1.514	1.725	1.293	0.944	0.897	0.913	0.849	0.812	0.883
P4	0.116	0.228	0.296	0.427	0.529	0.319	0.957	1.305	1.140	1.546	1.739	1.337	0.933	0.875	0.905	0.832	0.785	0.866
P5	0.115	0.221	0.298	0.420	0.528	0.316	0.877	1.170	1.064	1.438	1.617	1.233	0.946	0.902	0.920	0.855	0.812	0.887
P6	0.113	0.220	0.296	0.420	0.522	0.314	0.869	1.158	1.038	1.425	1.591	1.216	0.947	0.904	0.925	0.857	0.815	0.889
P7	0.122	0.241	0.320	0.445	0.562	0.338	0.972	1.661	1.552	1.582	2.113	1.576	0.930	0.863	0.882	0.826	0.767	0.854
P8	0.142	0.279	0.381	0.537	0.665	0.401	1.019	1.412	1.324	1.733	1.920	1.481	0.929	0.876	0.897	0.829	0.786	0.863
P9	0.128	0.250	0.334	0.468	0.589	0.354	0.933	1.240	1.127	1.502	1.695	1.299	0.939	0.889	0.909	0.841	0.788	0.873
P10	0.123	0.242	0.322	0.454	0.568	0.342	0.909	1.179	1.088	1.469	1.640	1.257	0.942	0.897	0.917	0.849	0.809	0.883

TABLE VI. Comparison of MAE, RMSE, and R Values for SC-FITNET Imputation at Different Percentages of Missingness

Input	MAE (mm)						RMSE (mm)						R					
P	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG
P1	**0.072**	**0.133**	**0.159**	**0.238**	**0.299**	**0.180**	**0.812**	**1.074**	**0.918**	**1.262**	**1.409**	**1.095**	**0.953**	**0.917**	**0.940**	**0.885**	**0.851**	**0.909**
P2	0.081	0.153	0.189	0.277	0.347	0.209	0.873	1.187	1.034	1.362	1.529	1.197	0.946	0.896	0.921	0.864	0.826	0.890
P3	0.087	0.169	0.209	0.302	0.382	0.230	0.921	1.278	1.116	1.424	1.644	1.277	0.939	0.876	0.906	0.848	0.786	0.871
P4	0.093	0.176	0.219	0.318	0.401	0.241	0.956	1.302	1.128	1.463	1.686	1.307	0.935	0.873	0.906	0.840	0.774	0.866
P5	0.103	0.195	0.245	0.348	0.443	0.267	1.027	1.432	1.250	1.560	1.824	1.419	0.923	0.844	0.882	0.816	0.732	0.839
P6	0.100	0.192	0.240	0.342	0.432	0.261	1.002	1.397	1.220	1.537	1.771	1.385	0.928	0.851	0.887	0.822	0.747	0.847
P7	0.101	0.191	0.238	0.340	0.434	0.261	1.012	1.404	1.234	1.541	1.807	1.400	0.927	0.853	0.888	0.823	0.741	0.846
P8	0.155	0.301	0.402	0.559	0.707	0.425	1.175	1.656	1.541	1.871	2.174	1.683	0.915	0.834	0.871	0.811	0.729	0.832
P9	0.101	0.193	0.238	0.341	0.435	0.262	1.028	1.430	1.253	1.556	1.829	1.419	0.924	0.845	0.882	0.818	0.725	0.839
P10	0.085	0.176	0.207	0.304	0.396	0.234	0.927	1.371	1.182	1.598	1.878	1.391	0.938	0.860	0.897	0.807	0.712	0.843

TABLE VII. Comparison of MAE, RMSE, and R Values for LSTM Imputation at Different Percentages of Missingness

Input	MAE (mm)						RMSE (mm)						R					
P	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG	10%	20%	30%	40%	50%	AVG
P1	0.081	0.162	0.191	0.255	0.338	0.205	1.015	1.585	1.401	1.538	1.907	1.489	0.928	0.812	0.857	0.825	0.704	0.825
P2	0.080	0.160	0.188	0.251	0.335	0.203	1.016	1.584	1.399	1.539	1.903	1.488	0.928	0.813	0.857	0.825	0.706	0.826
P3	0.082	0.163	0.194	0.258	0.345	0.209	1.017	1.584	1.401	1.536	1.909	1.489	0.927	0.813	0.857	0.825	0.704	0.825
P4	**0.080**	**0.159**	0.187	**0.248**	0.333	**0.201**	**1.010**	**1.581**	**1.398**	**1.532**	**1.900**	**1.484**	**0.928**	**0.813**	**0.857**	**0.827**	**0.708**	**0.827**
P5	0.083	0.165	0.195	0.261	0.346	0.210	1.018	1.587	1.401	1.544	1.909	1.492	0.927	0.812	0.857	0.823	0.704	0.825
P6	0.083	0.165	0.195	0.262	0.348	0.210	1.017	1.586	1.399	1.541	1.909	1.491	0.927	0.812	0.857	0.824	0.705	0.825
P7	0.084	0.169	0.203	0.270	0.359	0.217	1.014	1.582	1.399	1.538	1.907	1.488	0.928	0.813	0.857	0.825	0.705	0.826
P8	0.080	0.159	**0.185**	0.249	**0.332**	**0.201**	1.025	1.590	1.405	1.548	1.919	1.497	0.926	0.811	0.856	0.822	0.700	0.823
P9	0.082	0.164	0.195	0.261	0.347	0.210	1.018	1.587	1.401	1.547	1.913	1.493	0.927	0.812	0.857	0.823	0.703	0.824
P10	0.081	0.162	0.192	0.257	0.341	0.207	1.013	1.584	1.401	1.538	1.906	1.488	0.928	0.813	0.857	0.825	0.706	0.826

Note: The best results obtained are made bold.

A. Effect of Different Imputation Methods on Rainfall Data Series at Different Input P and Percentages of Missingness

Table IV, V, VI, and VII show the effects of different imputation methods on rainfall data series at different input p and missing rates. As seen in Table IV, the performances of FFNN increased as the input p decreased. Performance measures such as MAE, RMSE, and R show that FFNN achieved the best accuracy in total when input $p4$ was applied to the network. The average values of MAE, RMSE, and R measures for FFNN imputation were 0.299 mm, 1.251 mm, and 0.884 at $p4$, respectively.

From Table V, among the input p values, the first and second input ($p1$, $p2$) demonstrated good performances for predicting missing rainfall data. In particular, the SC-FFNN imputation for $p1$ showed excellent performance in estimating the various percentages of missingness in terms of MAE, RMSE, and R. The SC-FFNN imputation achieved an average accuracy of 90 %. The average MAE and RMSE measures of SC-FFNN were 0.183 mm and 1.142 mm at $p1$, respectively.

Meanwhile, the SC-FITNET imputation achieved optimal performance when the input $p1$ was used with an average accuracy of 90.9 %, as shown in Table VI. The average MAE and RMSE values are 0.180 mm 1.095 mm, respectively. On the other hand, performance measure such as MAE indicates the LSTM imputation achieved the lowest average error at the $p4$ and $p8$ as in Table VII. For the RMSE and R measures, the LSTM imputation obtained the best performances when input $p4$ was used, with an average value of 1.484 mm and 0.827, respectively. Overall, input $p1$ is the significant input for SC-FITNET and SC-FFNN imputation, while input $p4$ is the significant input for FFNN and LSTM imputation to achieve optimal imputation performances.

Furthermore, the study indicates the different missing rates would impact the accuracy of the missing data imputation. For example, when the missing rates increased from 10% to 50% at input $p1$, the MAE and RMSE measures increased from [0.072 mm, 0.831 mm] to [0.301 mm, 1.473 mm] respectively, but R decreased from 0.951 to 0.842 when using SC-FFNN imputation. Overall, the same input p that achieved the lowest mean absolute error (MAE) might also achieve the highest correlation coefficient (R). However, at 10% and 20% missingness, this study revealed that the same input p with the lowest value of MAE achieved the second-highest value of R instead of the highest value. This happens when the SC-FFNN imputation is able to measure the error between the predicted and the eventual outcomes accurately, but the R measures of correlation and dependence between the predicted and observed rainfall were statistically not the strongest.

A closer inspection revealed that the values of MAE for the four imputation methods linearly increased when the proportions of missing values increased. However, the values of RMSE linearly increased when the dataset had more than 30% missing values. This study supports the previous findings of Gill [51], Lee and Huber [52], Shang [53], Kim [54], and Ayilara [55] that the performance of imputation decreased when the proportion of missingness increased. According to Gill [51], the effect of missing data in information becomes very significant for hydrologic predictions as the percentage of missing data increases. Hence, this study concluded that more missing rainfall data in the dataset results in a poorer model performance, which is consistent with previous research [51]-[55].

B. Effect of Data Pre-processing Methods on Missing Data Prediction Performance

To investigate the effect of pre-processing data on the precision of missing rainfall imputation, a min-max normalization was used as a benchmark pre-processing data. The best performances obtained from the four models tabulated in Table IV, V, VI and VII were compared with the min-max normalization approach. For the min-max normalization approach, the raw meteorological and rainfall data were normalized as follows:

$$input\ p = \frac{X - min(X)}{max(X) - min(X)} \tag{15}$$

where $min(x)$ is the minimum value, $max(x)$ is the maximum value, and X is the data point.

Table VIII, IX and X show the effect of two different data pre-processing methods on the missing data estimation performance in terms of MAE, RMSE and R. For the min-max normalization approach, this study revealed that the LSTM imputation outperformed the other three models due to its capability to correlate the features in data. The performance measures such as MAE, RMSE and R show that the LSTM imputation achieved the lowest MAE and RMSE but highest R, as in Table VII, IX and X. The SC-FITNET was the second place with an average accuracy of 59%, followed by SC-FFNN, and FFNN imputation at an average accuracy of 55% and 49%, respectively. However, the performances of SC-FITNET, SC-FFNN and FFNN became unreliable as the percentage of missing data increases. The min-max normalization approach leads to inaccurate prediction due to the presence of zeros during the long dry periods. As a result, the three neural network models were not able to estimate the missing rainfall accurately. Hence, the min-max normalization approach is not suitable to be used for the long dry periods because it does not handle outliers very well.

TABLE VIII. Result on Imputation Process - Mean Absolute Error (MAE)

Missing rates	min-max				proposed work - PCA			
	FFNN	SC-FFNN	SC-FITNET	LSTM	FFNN	SC-FFNN	SC-FITNET	LSTM
10%	0.707	0.480	0.413	**0.107**	0.108	0.072	**0.072**	0.080
20%	1.418	0.968	0.833	**0.203**	0.212	0.134	**0.133**	0.159
30%	2.125	1.442	1.204	**0.248**	0.280	0.163	**0.159**	0.187
40%	2.846	1.938	1.617	**0.348**	0.397	0.244	**0.238**	0.248
50%	3.551	2.412	2.026	**0.450**	0.497	0.301	**0.299**	0.333
Avg	2.129	1.448	1.219	**0.271**	0.299	0.183	**0.180**	0.201

TABLE IX. Result on Imputation Process - Root Mean Square Error (RMSE)

Missing rates	min-max				proposed work - PCA			
	FFNN	SC-FFNN	SC-FITNET	LSTM	FFNN	SC-FFNN	SC-FITNET	LSTM
10%	3.101	2.198	1.838	**1.127**	0.871	0.831	**0.812**	1.010
20%	4.471	3.158	2.661	**1.659**	1.210	1.108	**1.074**	1.581
30%	5.301	3.724	2.956	**1.476**	1.103	0.952	**0.918**	1.398
40%	6.158	4.351	3.421	**1.716**	1.435	1.344	**1.262**	1.532
50%	6.894	4.866	3.850	**2.063**	1.638	1.473	**1.409**	1.900
Avg	5.185	3.659	2.945	**1.608**	1.251	1.142	**1.095**	1.484

TABLE X. Result on Imputation Process - Correlation Coefficient, R

Missing rates	min-max				proposed work - PCA			
	FFNN	SC-FFNN	SC-FITNET	LSTM	FFNN	SC-FFNN	SC-FITNET	LSTM
10%	0.693	0.754	0.794	**0.910**	0.947	0.951	**0.953**	0.928
20%	0.522	0.593	0.613	**0.792**	0.896	0.912	**0.917**	0.813
30%	0.495	0.558	0.612	**0.840**	0.914	0.936	**0.940**	0.857
40%	0.416	0.474	0.525	**0.776**	0.856	0.874	**0.885**	0.827
50%	0.343	0.409	0.409	**0.637**	0.808	0.842	**0.851**	0.708
Avg	0.494	0.558	0.591	**0.791**	0.884	0.903	**0.909**	0.827

Note: The best results obtained are made bold.

On the other hand, the four models achieved higher performances when the proposed PCA data pre-processing approach was used. It shows that the proposed significant input was able to help the four models to estimate the missing time series at higher accuracy compared to the min-max approach. Performance measures such as MAE and RMSE show that the SC-FITNET has the lowest error rates among the other three models, while the SC-FFNN imputation was in second place. Furthermore, the correlation and coefficient, R-value indicates the SC-FITNET scored the highest average accuracy of 90.9%, followed by SC-FFNN, FFNN and LSTM imputation. It shows that the adaptation of sine cosine algorithm into the existing neural network (SC-FITNET and SC-FFNN) was able to optimize the neural network and achieved higher accuracy but lower MAE and RMSE values compared to the FFNN imputation. The possible reason is that the position-updating function of SC-FITNET and SC-FFNN could positively optimize the entire search space for the best weights and biases of the neural networks and consequently increase the accuracy of the imputation.

Furthermore, the LSTM performed slightly better when the proposed PCA data pre-processing approach was used than the min-max approach. However, this study revealed that SC-FITNET and SC-FFNN slightly indicate a better prediction performance compared to the LSTM model. In addition to that, recent studies have shown that temporal convolutional networks (TCN) [56], and multilayer perceptron (MLP) [57] can outperform recurrent models such as LSTM. The LSTM model may require a large amount of data to perform better than the other methods. In terms of computational time, the LSTM model required more time to perform the missing data estimation process than the three models, FFNN, SC-FFNN and SC-FITNET (results not shown here). Hence, the FFNN, SC-FFNN and SC-FITNET models have the advantage of being computationally less costly compared to the LSTM model. In particular, there is a reduction of the average training time in the three models, approximately four times less than the LSTM model.

Overall, the SC-FITNET imputation has proven to be the top performer when the proposed PCA data pre-processing approach was used, while the LSTM imputation demonstrated the top performer for the min-max normalization approach.

VI. Conclusion

We investigated the potential of using meteorological and rainfall data from nearest neighbor gauging stations for infilling missing rainfall data. Before the imputation, this study introduced PCA to extract significant features from the meteorological data. The comparison of different combination input in imputation was presented and evaluated using four imputation methods, SC-FITNET, SC-FFNN, LSTM and FFNN. With medium size data of 11,680 real-life records, the four methods were trained and compared at five different percentages of missingness under MCAR conditions (10%, 20%, 30%, 40%, and 50%). The study concluded that the proposed SC-FITNET imputation has a higher capability in treating missing values for the PCA pre-processed dataset than than the LSTM, SC-FFNN and FFNN imputation in terms of MAE, RMSE, and R. By adopting the position-updating function, the proposed SC-FITNET imputation successfully achieved better accuracy in missing data estimation as compared to the other three approaches. Hence, the results of the proposed SC-FITNET imputation in this work support its use for infilling real-life missing rainfall data. In addition, the study revealed that the meteorological data (non-precipitation) and rainfall data (precipitation) from nearest neighbor stations are compatible and can be used as input for missing data imputation. The performances of the proposed PCA as data pre-processing have an obvious advantage over the benchmark.

For future work, considering a longer period of data, investigating other data pre-processing techniques and further testing the effectiveness of the proposed algorithm on different types of datasets are recommended. In addition, the imputed rainfall dataset could be used as an input in the hydrological data analysis. The imputed data could be employed to estimate the river flow and the occurrence of floods during the rainy season, to determine the severity and frequency of drought during the dry season, to design water supply, and other hydrological data analyses.

Acknowledgment

The authors would like to acknowledge the Malaysian Meteorological Department and Department of Irrigation and Drainage (DID), Sarawak, Malaysia, for providing the meteorological and rainfall data in this study. This work was supported/funded by the Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS/1/2018/ICT04/UTM/01/1). The authors sincerely thank Universiti Teknologi Malaysia (UTM) under Research University Grant Vot-20H04, Malaysia Research University Network (MRUN) Vot 4L876, and SLAI supported under Ministry of Higher Education Malaysia for the completion of the research. The work is partially supported by the SPEV project (ID: 2102-2021), Faculty of Informatics and Management, University of Hradec Kralove. We are also grateful for the support of Ph.D. students Michal Dobrovolny and Sebastien Mambou in consultations regarding application aspects from Hradec Kralove University, Czech Republic. The APC was funded by the SPEV project 2102/2021, Faculty of Informatics and Management, University of Hradec Kralove.

References

[1] P. Muñoz, J. Orellana-Alvear, P. Willems, and R. Célleri. "Flash-flood forecasting in an Andean mountain catchment—Development of a step-wise methodology based on the random forest algorithm," *Water*, vol. 10, no. 11, 2018, pp. 1519.

[2] S. Szewrański, J. Chruściński, J. Kazak, M. Świąder, K. Tokarczyk-Dorociak, and R. Żmuda, "Pluvial flood risk assessment tool (PFRA) for rainwater management and adaptation to climate change in newly urbanised areas," *Water*, vol. 10, no. 4, 2018, pp. 386.

[3] K.K. Kuok, S. Harun, S.M. Shamsuddin, and P.C. Chiu, "Evaluation of daily rainfall-runoff model using multilayer perceptron and particle swarm optimization feedforward neural networks," *Journal of Environmental Hydrology*, vol. 18, no. 10, 2010, pp. 1-16.

[4] N. Yang, B.H. Men, and C.K. Lin, "Impact analysis of climate change on water resources," *Procedia Engineering*, vol. 24, 2011, pp. 643-648.

[5] K.K. Kuok, S. Harun, and P.C. Chiu, "Hourly runoff forecast at different leadtime for a small watershed using artificial neural networks," *International Journal of Advances in Soft Computing and its Application*, vol. 3, 2011, pp. 68-86.

[6] R.A. Mcdonald, P.W. Thurston, and M.R. Nelson, "A Monte Carlo study of missing item methods," *Organizational Research Methods*, vol. 3, no. 1, 2000, pp. 71-92.

[7] P.E. McKnight, K.M. McKnight, S. Sidani, and A.J. Figueredo, "Missing data: A gentle introduction," Guilford Press. 2007.

[8] K.J. Lee and J.B. Carlin, "Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation," *American Journal of Epidemiology*, vol. 171, no. 5, 2010, pp. 624-632.

[9] Y. Gao, C. Merz, G. Lischeid, and M. Schneider, "A review on missing hydrological data processing," *Environmental earth sciences*, vol. 77, no. 2, 2018, pp. 47.

[10] S. Londhe, P. Dixit, S. Shah, and S. Narkhede, "Infilling of missing daily rainfall records using artificial neural network," *ISH Journal of Hydraulic Engineering*, vol. 21, no. 3, 2015, pp. 255-264.

[11] T. Canchala-Nastar, Y. Carvajal-Escobar, W. Alfonso-Morales, W.L. Cerón and E. Caicedo, "Estimation of missing data of monthly rainfall in southwestern Colombia using artificial neural networks, " *Data in brief*,

vol. 26, 2019, pp. 104517.
[12] P.C. Chiu, A. Selamat, O. Krejcar, and K.K. Kuok, "Missing rainfall data estimation using artificial neural network and nearest neighbor imputation," In Advancing Technology Industrialization Through Intelligent Software Methodologies, Tools and Techniques: Proceedings of the 18th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_19), IOS Press, vol. 318, 2019, pp. 132.
[13] M.R. Mispan, N.F.A. Rahman, M.F. Ali, K. Khalid, M.H.A. Bakar and S.H. Haron, "Missing river discharge data imputation approach using artificial neural network," Methodology, vol. 25, 2015, pp. 20.
[14] P.C. Chiu, A. Selamat and O. Krejcar, "Infilling missing rainfall and runoff data for Sarawak, Malaysia using gaussian mixture model based K-nearest neighbor Imputation," In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, Springer, 2019, pp. 27-38.
[15] W.Y. Lai, and K.K. Kuok, "A study on bayesian principal component analysis for addressing missing rainfall data," *Water Resources Management*, 2019, pp.1-14.
[16] S. Mirjalili, "SCA: A sine cosine algorithm for solving optimization problems," *Knowledge-Based Systems*, vol. 96, 2016, pp. 120-133.
[17] C. Qu, Z. Zeng, J. Dai, Z. Yi, and W. He, "A modified sine-cosine algorithm based on neighborhood search and greedy levy mutation," *Computational Intelligence and Neuroscience*, 2018.
[18] S. Das, A. Bhattacharya and A.K. Chakraborty, "Solution of short-term hydrothermal scheduling using sine cosine algorithm," *Soft Computing*, vol. 22, no. 19, 2018, pp. 6409-6427.
[19] S. Li, H. Fang, and X. Liu, "Parameter optimization of support vector regression based on sine cosine algorithm," *Expert Systems with Applications*, vol. 91, 2018, pp. 63-77.
[20] M.A. Tawhid, and P. Savsani, "Discrete sine-cosine algorithm (DSCA) with local search for solving traveling salesman problem," *Arabian Journal for Science and Engineering*, 2018, pp. 1-11.
[21] R.E. Chandler, V.S. Isham, N.A. Leith, P.J. Northrop, C.J. Onof, and H.S. Wheater, "Uncertainty in rainfall inputs," World Scientific/Imperial College Press, 2011.
[22] O. Stoner, and T. Economou, "An advanced hidden markov model for hourly rainfall time series," arXiv:1906.03846. 2019.
[23] T. Kashiwao, K. Nakayama, S. Ando, K. Ikeda, M. Lee and A. Bahadori, "A neural network-based local rainfall prediction system using meteorological data on the Internet: A case study using data from the Japan Meteorological Agency," *Applied Soft Computing*, vol. 56, 2017, pp. 317-330.
[24] M.H. Yen, D.W. Liu, Y.C. Hsin, C.E. Lin, and C.C. Chen, "Application of the deep learning for the prediction of rainfall in Southern Taiwan," *Scientific Reports*, vol. 9, no. 1, 2019, pp. 1-9.
[25] M. Chhetri, S. Kumar, P.P. Roy, and B.G. Kim, "Deep BLSTM-GRU model for monthly rainfall prediction: A case study of Simtokha, Bhutan," *Remote Sensing*, vol. 12, no. 19, 2020, pp.3174.
[26] S.K. Grange, and D.C. Carslaw, "Using meteorological normalisation to detect interventions in air quality time series," *Science of the Total Environment*, vol. 653, 2019, pp.578-588.
[27] L.I. Smith, "A tutorial on principal components analysis", 2002. Accessed: Jan. 3, 2020. [Online]. Available: http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
[28] C. Skittides, and W.G. Früh, "Wind forecasting using principal component analysis," *Renewable Energy*, vol. 69, 2014, pp. 365-374.
[29] M. Hubert, P.J. Rousseeuw, and W. Van den Bossche, "MacroPCA: An all-in-one PCA method allowing for missing values as well as cellwise and rowwise outliers," *Technometrics*, vol. 61, no.4, 2019, pp. 459-473.
[30] Z. Zuśka, J. Kopcińska, E. Dacewicz, B. Skowera, J. Wojkowski, and A. Ziernicka–Wojtaszek, "Application of the principal component analysis (PCA) method to assess the impact of meteorological elements on concentrations of particulate matter (PM10): A case study of the mountain valley (the Sącz Basin, Poland)," *Sustainability*, vol. 11, no. 23, 2019, pp. 6740.
[31] Y.Y Choi, H. Shon, Y.J. Byon, D.K. Kim, S. Kang, "Enhanced application of principal component analysis in machine learning for imputation of missing traffic data," *Applied Science*, vol. 9, no. 10, 2019, pp. 2149.
[32] B.S. Harish, and S.V.A Kumar. "Anomaly based intrusion detection using modified fuzzy clustering," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol 4, no. 6, 2017, pp. 54-59, doi: 10.9781/ijimai.2017.05.002.
[33] T. Kurita, "Principal component analysis (PCA)," In: Ikeuchi K. (eds) *Computer Vision: A Reference Guide*, Springer, 2014.
[34] K. Pearson, "Principal components analysis," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 6, no.2, 1901, pp. 559.
[35] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, 1933, pp. 417-441.
[36] R. Khattree, and D.N. Naik "Multivariate data reduction and discrimination with SAS software," SAS Institute, 2000.
[37] G. Bebis, and M. Georgiopoulos, "Feedforward neural networks," *IEEE Potentials*, vol. 13, no. 4, 1994, pp. 27-31.
[38] S. Mirjalili, "How effective is the Grey Wolf optimizer in training multi-layer perceptrons," *Applied Intelligence*, vol. 43, no.1, 2015, pp.150-161.
[39] S. Hochreiter, and J. Schmidhuber, "Long short-term memory," *Neural Computing*, vol. 9, no. 8, 1997, pp. 1735-1780.
[40] E. Mussumeci, and F.C. Coelho, "Large-scale multivariate forecasting models for Dengue-LSTM versus random forest regression," *Spatial and Spatio-temporal Epidemiology*, vol. 35, 2020, pp. 100372.
[41] S. Maya, and U. Ken, "DADIL: Data augmentation for domain-invariant learning," *Data Science and Pattern Recognition*, vol. 4, no. 2, 2020, pp. 33–49.
[42] J.C.W. Lin, Y. Shao, Y. Djenouri and U. Yun, "ASRNN: A recurrent neural network with an attention model for sequence labeling," *Knowledge-Based Systems*, vol. 212, 2020, pp. 106548.
[43] Hourly meteorological dataset for Kuching station: 2002 to 2003, Malaysian Meteorological Department, Selangor, Malaysia, October 2019.
[44] Hourly rainfall datasets for Sungai Merang station and nearest neighbor stations: 2002 to 2003, Department of Irrigation and Drainage (DID), Sarawak, Malaysia, October 2019.
[45] A.J. Henry, N.D. Hevelone, S. Lipsitz, and L.L. Nguyen, "Comparative methods for handling missing data in large databases," *Journal of Vascular Surgery*, vol. 58, no. 5, 2013, pp. 1353-1359.
[46] J.R. Cheema, "Some general guidelines for choosing missing data handling methods in educational research," *Journal of Modern Applied Statistical Methods*, vol. 13, no. 2, 2014, pp. 3.
[47] P. Zhu, Q. Xu, Q. Hu, C. Zhang, and H. Zhao, "Multi-label feature selection with missing labels," *Pattern Recognition*, vol. 74, 2018, pp. 488-502.
[48] H. Hassani, M. Kalantari, and Z. Ghodsi, "Evaluating the performance of multiple imputation methods for handling missing values in time series data: A study focused on East Africa, soil-carbonate-stable isotope data," *Stats*, vol. 2, no. 4, 2019, pp. 457-467.
[49] S. Oba, M.A. Sato, I. Takemasa, M. Monden, K.I. Matsubara, and S. Ishii, "A Bayesian missing value estimation method for gene expression profile data," *Bioinformatics*, vol. 19, no.16, 2003, pp. 2088-2096.
[50] R.J. Little, and D.B. Rubin, "Statistical analysis with missing data," John Wiley & Sons, 2014.
[51] M.K. Gill, T. Asefa, Y. Kaheil, and M. McKee, "Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique," *Water Resources Research*, vol. 43, no.7, 2007.
[52] J. H. Lee, and Jr.J. Huber, "Multiple imputation with large proportions of missing data: How much is too much?," In United Kingdom Stata Users' Group Meetings 2011 (No. 23). Stata Users Group, 2011.
[53] Q. Shang, Z. Yang, S. Gao, and D. Tan, "An imputation method for missing traffic data based on FCM optimized by PSO-SVR," *Journal of Advanced Transportation*, 2018.
[54] T. Kim, W. Ko, and J. Kim, "Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting," *Applied Sciences*, vol. 9, no. 1, 2019, pp. 204.
[55] O.F. Ayilara, L. Zhang, T.T. Sajobi, R. Sawatzky, E. Bohm, and L.M. Lix, "Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry," *Health and Quality of Life Outcomes*, vol. 17, no. 1, 2019, pp. 106.
[56] S. Bai, J.Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," arXiv preprint arXiv:1803.01271, 2018.

[57] A. Cecaj, M. Lippi, M. Mamei, and F. Zambonelli, "Comparing deep learning and statistical methods in forecasting crowd distribution from Aggregated Mobile Phone Data," *Applied Sciences*, vol. 10, no. 18, 2020, pp. 6580.

Po Chan Chiu

Po Chan Chiu is currently pursuing Ph.D degree in Computer Science from Universiti Teknologi Malaysia (UTM). She received the M.Sc. in information technology from the Universiti Malaysia Sarawak (UNIMAS), in 2010. She was the Software Engineer at private companies for 3 years. Her research interests include artificial intelligence, optimization, data analytics and neural networks.

Ali Selamat

Ali Selamat is currently a Full Professor with Universiti Teknologi Malaysia (UTM), Malaysia. He has also been the Dean of the Malaysia Japan International Institute of Technology (MJIIT), UTM, since 2018. An academic institution established under the cooperation of the Japanese International Cooperation Agency (JICA) and the Ministry of Education Malaysia (MOE) to provide the Japanese style of education in Malaysia. He is also a Professor with the Software Engineering Department, School of Computing, UTM and the Chair of the IEEE Computer Society Malaysia Section. He has published more than 120 research articles with IF JCR, with more than 2400 citations received in the Web of Science and h-index 26. His research interests include software engineering, software process improvement, software agents, web engineering, information retrievals, pattern recognition, genetic algorithms, neural networks, soft computing, collective computational intelligence, strategic management, key performance indicator, and knowledge management. He is on the Editorial Board of the journal Knowledge-Based Systems (Elsevier).

Ondrej Krejcar

Ondrej Krejcar is currently a Full Professor of systems engineering and informatics with the University of Hradec Kralove, Czech Republic. He is also the Vice-Dean for science and research at the Faculty of Informatics and Management, UHK. He is also the Director of the Center for Basic and Applied Research, University of Hradec Kralove. At the University of Hradec Kralove, he is a Guarantee of the Doctoral Study Programme in Applied Informatics, where he is focusing on lecturing on smart approaches to the development of information systems and applications in ubiquitous computing environments. His h-index is 20 (according Web of Science), with more than 1500 citations received in the Web of Science. He has published more than 110 research articls with IF JCR. He has a number of collaborations throughout the world (e.g., Malaysia, Spain, U.K., Ireland, Ethiopia, Latvia, and Brazil). His research interests include control systems, smart sensors, ubiquitous computing, manufacturing, wireless technology, portable devices, biomedicine, image segmentation and recognition, biometrics, technical cybernetics, and ubiquitous computing. His second area of interest is in biomedicine (image analysis), as well as biotelemetric system architecture (portable device architecture and wireless biosensors), and the development of applications for mobile devices with use of remote or embedded biomedical sensors. Dr. Krejcar has also been a Management Committee Member substitute of the project COST CA16226, since 2017. In 2018, he was the 14th Top-Peer Reviewer in Multidisciplinary in the World according to Publons. He is on the Editorial Board of Sensors (MDPI) with JCR Index and several other ESCI indexed journals. He has been the Vice-Leader and a Management Committee Member at WG4 of the project COST CA17136, since 2018. Since 2019, he has been the Chairman of the Program Committee of the KAPPA Program, Technological Agency of the Czech Republic, as a Regulator of the EEA/Norwegian Financial Mechanism in the Czech Republic (2019–2024). Since 2014, he has been the Deputy Chairman of the Panel 7 (Processing Industry, Robotics and Electrical Engineering) of the Epsilon Program, Technological Agency of the Czech Republic.

King Kuok Kuok

King Kuok Kuok is a senior lecturer at Swinburne University of Technology Sarawak Campus. He received his MEng from the UNIMAS in 2004 and Ph.D. from the UTM in 2010. He was the Field Engineer for Hydrological and Water Resources Branch, Department of Irrigation and Drainage, State of Sarawak, Malaysia from 2002 to 2009 and the Road, Civil and Structural Design Engineer at private companies for more than 10 years. His research interests include water resources, water supply, hydrology, artificial intelligence and building information modeling.

Enrique Herrera-Viedma

Enrique Herrera-Viedma is a Professor of Computer Science and the Vice-President of Research and Knowledge Transfer with the University of Granada. His H-index is 85 with more than 25000 citations received in Web of Science and 97 in Google Scholar with more than 38500 citations received. His current research interests include group decision-making, consensus models, linguistic modeling, aggregation of information, information retrieval, bibliometric, digital libraries, Web quality evaluation, recommender systems, and social media. He has been identified as one of the World's Most Influential Researchers by Shanghai Center and Thomson Reuters/Clarivate Analytics in both the computer science and scientific engineering categories in 2014–2020. Prof. Herrera-Viedma was the 2019–2020 Vice-President of Publications with the IEEE SMC Society and an Associate Editor of several journals, such as the IEEE Transactions on Fuzzy Systems, the IEEE Transactions on Systems, Man, and Cybernetics: Systems, Information Sciences, Applied Soft Computing, Soft Computing, Fuzzy Optimization and Decision Making, and Knowledge-Based Systems.

Giuseppe Fenza

Giuseppe Fenza received the Ph.D. degree in Computer Sciences at the University of Salerno, Italy, in 2009. From 2009 until now, he collaborates to several research initiatives mainly focused on Knowledge Extraction from unstructured resources defining intelligent systems based on the combination of techniques from Soft Computing, Semantic Web, areas in which he has many publications. He has been deeply involved in several EU and Italian Research and Development projects on ICT and, in particular, on Situation Awareness, Service Discovery, Enterprise Information Management and e-Commerce. He serves as Associate Editor in international journals, such as: Neurocomputing, International Journal of Grid and Utility Computing, International Journal of Engineering Business Management. He has published extensively about: Fuzzy Decision Making, Ontology Elicitation, Situation and Context-Awareness, Semantic Information Retrieval. Recently, he is working in the field of Big Data, Social Media Analytics, and Web Intelligence by proposing novel methods for instance, to support microblog summarization, time-aware information retrieval and recommendation extraction. He is currently an Assistant Professor in Computer Science at the Department of Management and Innovation Systems, University of Salerno, Italy.

Design and Development of an Energy Efficient Multimedia Cloud Data Center with Minimal SLA Violation

Nirmal Kr. Biswas[1], Sourav Banerjee[2]*, Utpal Biswas[3]

[1] Global Institute of Management & Technology, Krishnagar, 741102 (India)
[2] Kalyani Government Engineering College, Kalyani, 741235 (India)
[3] University of Kalyani, Kalyani, 741235 (India)

Received 21 October 2020 | Accepted 13 January 2021 | Published 20 April 2021

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Multimedia computing (MC) is rising as a nascent computing paradigm to process multimedia applications and provide efficient multimedia cloud services with optimal Quality of Service (QoS) to the multimedia cloud users. But, the growing popularity of MC is affecting the climate. Because multimedia cloud data centers consume an enormous amount of energy to provide services, it harms the environment due to carbon dioxide emissions. Virtual machine (VM) migration can effectively address this issue; it reduces the energy consumption of multimedia cloud data centers. Due to the reduction of Energy Consumption (EC), the Service Level Agreement violation (SLAV) may increase. An efficient VM selection plays a crucial role in maintaining the stability between EC and SLAV. This work highlights a novel VM selection policy based on identifying the Maximum value among the differences of the Sum of Squares Utilization Rate (*MdSSUR*) parameter to reduce the EC of multimedia cloud data centers with minimal SLAV. The proposed *MdSSUR* VM selection policy has been evaluated using real workload traces in CloudSim. The simulation result of the proposed *MdSSUR* VM selection policy demonstrates the rate of improvements of the EC, the number of VM migrations, and the SLAV by 28.37%, 89.47%, and 79.14%, respectively.

Keywords

Energy Consumption (EC), VM Migrations, Multimedia Cloud (MC), SLA Violation (SLAV), VM Selection.

DOI: 10.9781/ijimai.2021.04.004

I. Introduction

Nowadays, multimedia is emerging as a service over the Internet. Multimedia applications [1], like image searching, sharing, editing, video conferencing, multimedia content delivery, multimedia streaming, video retrieval, etc. required a massive amount of storage and computation power. Thus Cloud Computing [2] technology can provide multimedia application services to the users on demand. In Multimedia Cloud (MC) [3], cloud service providers deploy the cloud resources to process multimedia demands and provide the necessary service to the users. The user can store and process the requisite multimedia application data in the cloud in a distributed manner in the modern paradigm of Multimedia Cloud (MC). The need for a full installation of user's media applications in the user's computer is over. The biggest challenge is to optimally allocate the resources to maintain the Quality of Service (QoS) [4] of the various multimedia applications.

The demand for multimedia services has increased rapidly day by day. The MC data centers required a substantial amount of energy to provide services to the users. However, it's a challenge to the researchers to give the MC services to the users with satisfactory Quality of Service (QoS). A large-scale MC data centers consist of millions of servers. It consumes a considerable amount of energy and emitting a massive amount of carbon dioxide into the environment. The electricity consumed by global data centers in 2018 was an estimated 198 terawatt-hours (TWh), which is almost 1% of the demand for global electricity [5]. Because of the energy consumption of global data centers, the average electricity emission rate at each data center is about 4.4 kilogram of carbon dioxide per kilowatt-hour [6]. The Yale School of the Environment estimates that global data centers have a gross emission of carbon dioxide compared to the aviation industry [7] around the world, amounting to around 900 billion kilograms of carbon dioxide [8]. So, global data centers' energy consumption reduction with satisfactory QoS to the MC users becomes a key concern to the researchers. The virtualization approach is used to address these issues. In the MC environment, Virtual machines (VMs) are created in physical machines (PM) using virtualization [9]–[11] technology, depending on the user's request. PMs are encapsulated different applications in the form of VMs by separating with each VM. Each VM required some resources like CPU, Memory, Storage, Band Width, etc. To run the VM, the sum of the required resources must always be lesser than the host capacity. VM Consolidation (VMC) [12]–[14] is an approach which can efficiently utilize the resources with satisfactory QoS. The Service Level Agreement (SLA) [15] between MC users and service providers define QoS. VMC can help to reduce the energy

* Corresponding author.
E-mail address: mr.sourav.banerjee@ieee.org

consumption of MC data centers with optimal SLA violations. It has four main steps. Firstly, detect the overloaded hosts and then detect the underloaded hosts. Now select some VMs from overloaded hosts and all VMs from underloaded hosts. The selected VMs must eventually migrate to the medium loaded hosts, and all of the underloaded hosts are shut down. MC data center's required energy can be minimized by using VMC and VM migration [16], [17]. In live VM migration, an entire running VM can move from one host to another host without any interruption of the user's service. But, too much VM migration may increase the SLA violation and cost of the operation.

VM selection is an essential part of VMC. It creates a bunch of VMs from an overloaded host, which should migrate to moderately loaded hosts. The selection of proper VMs from an overloaded host may control the number of migrations. It can reduce the operational cost with reduced energy consumption and SLA violation. Herein, we have proposed a novel VM selection policy based on Maximum value among the differences in the Sum of Squares Utilization Rate (*MdSSUR*). The proposed *MdSSUR* VM selection policy selects VMs from overload hosts and performs the VM migration with reduced EC and SLA violation. It also reduced the number of migrations. The proposed *MdSSUR* VM selection policy has been executed and evaluated in CloudSim 3.0 [18]. The performance of the proposed *MdSSUR* VM selection policy has been assessed with some established VM selection approaches [19]–[21]. The proposed *MdSSUR* VM selection policy significantly improved the EC, SLAV, and VM migration.

The remaining section of the paper is organized as follows. Section II gives a brief description of the literature survey and state of the art. The proposed *MdSSUR* VM selection policy is shown in section III. The VM consolidation based on *MdSSUR* VM selection policy is shown in section IV. The detailed analysis of the proposed *MdSSUR* VM selection policy is shown in section V. Finally, Section VI concludes the research work.

II. Literature Survey

MC's rapid growth has got more attention from MC providers for the MC data center's cost and efficiency. MC providers of-fer high-quality services at the lowest price to mesmerize the MC users. The number of MC users are increasing exponentially. So, efficient approaches must be adopted by the service providers to satisfy the user's requirements with minimum energy consump-tion and SLA violation. Virtualization [11] is an approach where VMs are into the servers to provide the services to the MC users.

In [22]–[24], the authors successfully tried to address the EC issue with a DVFS approach in cloud computing. DVFS is an approach in which the server load is balanced dynamically with the CPU's voltage and frequency. The energy consumption has reduced for lower voltage and frequency. But, lower CPU frequency may decrease the CPU performance. Beloglazov et al. [19], proposed a system based on VM consolidation and VM migration to improve energy consumption. They developed the following policies for VM selection: 1) Minimum migration time (MMT), 2) Maximum correlation (MC), 3) Minimum utilization (MU), and 4) Random selection (RS). The MMT prefers VMs whose migration time is minimum, and MC selects VMs with full correlation. In the MU selection policy, underutilized VMs are selected, and RS selects the VMs randomly.

Yadav et al. [20] implemented a proposal referred to as the maximum utilization minimum size (MuMs). MuMs is based on CPU utilization and VM's RAM size. It selects highly utilized VMs with minimal RAM, and as an essential parameter, selects the ratio between CPU and VM's RAM size. Akhter et al. [21] proposed a policy to reduce the EC of cloud data centers. Their proposed VM selection policy is known as Maximum migration time (MxMT). The MxMT select VMs with maximum migration time. It reduces EC by 19%. But, MxMT had undoubtedly experienced a severe effect of SLA violation. Lin et al. [25] proposed a model for the task of sequence labeling. Natural language processing (NLP) is used for managing text and speech. In NLP, one of the essential tasks is sequence labeling to define and allocate category label to each unit in the particular entry. These traditional models' efficiency is heavily dependent on manufactured features and task-specific intelligence, which are very time-consuming. The authors developed an attention segmental recurrent neural network (ASRNN) for the task of sequence labeling. The model depends on an ordered recognition neural semi-Markov condition random fields. A hierarchical structure uses to incorporate character level and word level information. The proposed model takes advantage of the hierarchical structure, with many data that achieve competitive efficiency.

R. Mandal et al. [26] developed a VM selection policy known as the Power-Aware VM selection policy. It selects the maximum utilized VM and adds it to the migration list. The utilization of a VM is computed by the ratio of current VM utilization and VM's allocated resources. R. Yadav et al. [27] proposed a Bandwidth-Aware VM selection policy. This policy chooses a VM from a host that is overloaded with a minimum current utilization and total migration time. The Bandwidth Transfer Component (BTC) has been computed by dividing the VM size and its current utilization by available bandwidth. Now, the migration time is calculated for all VMs using BTC and ping time (PT), and finally, a VM with minimum migration time is selected and added to the migration list. Lin et al. [28] proposed an efficient approach namely HUIM-BPSOsig to mine high-utility itemsets. The proposed approach is based on discrete particle swarm optimization (PSO). C. Zhang et al. [29] developed a VM selection policy to reduce the energy consumption and SLA violation of the cloud dissenters. The VM selection aimed to minimize the number and cost of migration. The authors refer to the VM selection as the Minimize Number and Cost of Migrations (MNCM) policy. In MNCM, a VM with maximum VM resource occupancy (VRO) is selected for migration. H. Toumi et al. [30] develop a cooperative framework between Hybrid Intrusion Detection System (Hy-IDS) based upon Mobile Agents and virtual firewalls. The possibility of intrusion rises in occurrence due to the massive use of the cloud. Security, accountability, and stability in the cloud model are essential for customer satisfaction. The minimization of the effect of any penetration into this area is one of the security concerns. The proposed cooperative framework system makes for quicker and more productive detection and resolution of new distributed threats.

Y. Wen et al. [31] proposed a VM selection policy known as minimum migration (MM) policy. The authors calculate the Euclidean distance between the VMs load pattern and PMs load pattern. The distance has been sorted, and depending on a threshold value, select the VMs which consume more resources. In [32], the authors proposed a policy based on Minimum Utilization Gap (MUG). The authors computed the difference between the utilization of overloaded host and upper utilization threshold value as Δ. The relative utilization of each VM was computed by the ratio between the required MIPS and total capacity. The Utilization Gap is the absolute difference between Δ and relative utilization of each VM. In the migration list, VMs with Minimum Utilization Gap are added. Lin et al. [33] developed a model to secure secret and sensitive information. The 6G networking based on Terahertz offers the absolute highest efficiency and reliability but faces new man-in-the-middle attacks. The main challenge of such extremely vulnerable environments is the security and confidentiality of the data. The authors proposed an ant colony optimization (ACO) method to secure 6G IoT networks. The proposed method has multiple targets and the deletion of a transaction to ensure data security.

Every ant in the population is represented as a set of possible deletion transactions for hiding sensitive information. The authors claim that the proposed method reaches a negligible side effect with a low average computational cost. V.K. Solanki et al. [34] have developed a module that integrates new technical peripherals for simple energy-saving trends and modernizes the module in IoT. Owing to irresponsible officials' most resources like water and electricity have been wasted in different cities. The developed module can significantly save the wastage of these resources.

S. B. Melhem et al. [35] evolved a VM selection policy known as Minimum Migration Time Minimum VM Migrated Count (MmtMiMc). MmtMiMc first selects VMs with a minimal quantity of memory and sort them in growing order. Then, from the selected VMs, the policy finds the VMs with a minimum number of VM migrated count and adds them to the migration list to perform the VM migration. They claimed that MmtMiMc decreases the number of VM migration maximum of up to 52.11%. S.M. Moghaddam et al. [36] proposed a VM selection policy based on predictive maximum CPU usages and minimum migration time. The ratio between the memory of VM and available bandwidth of a host is computed, and its minimum value is multiplied with a maximum predicted CPU usage of the VM. H. Peng et al. [37] developed a gradual gradable neural language learning structure. It can be used in the Continuous Bag-of-Words (CBOW) and skip-gram model. The authors extended the classical hierarchical formation from a human tree to a weighted contextual frequency aggregated tree for a long time. S.A. Makhlouf et al. [38] proposed a novel method for data-intensive workflow scheduling applications. Several optimization methods have been developed to improve the cost and efficiency of data-intensive scientific Workflow Scheduling (DiSWS) in cloud computing. Most of the DiSWS techniques are based on an optimization process using heuristic and metaheuristic approaches. The authors explore the task hierarchy in data data-intensive scientific workflows by their proposed method.

J. S. Pan et al. [39] proposed an approach named Multi-group Grasshopper Optimization Algorithm (MGOA). A modern algorithm that imitates Grassley's actions in nature is the Grasshopper Optimization Algorithm (GOA). The MGOA can be used to address the capacitated vehicle routing problem (CVRP). The authors claim that the efficiency of the MGOA is better than the Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). K.W. Huang et al. [40] developed an image recognition framework using the GoogLeNet model. The proposed framework is a convolutionary-neural-network module focused on deep learning. The image recognition framework can enhance the precision of module recognition effectively for preprocessed images. M. El. Ghazouani et al. [41] proposed a block-chain based solution that will maintain the privacy of cloud data checks by deduplicating data. An excellent alternative to ensure data storage reliability in cloud computing is called the deduplication of data. Cloud technology provides many benefits for storage service and poses security problems, notably concerning data privacy, a core component of any cloud system. The proposed method guarantees that consumer data remain confidential for auditors in the course of audits.

P. Xu. et al. [42] developed a VM allocation policy (VMA-ACO) based on Ant Colony Optimization (ACO) [43]. The main aim of the VMA-ACO policy was to maximize resource utilization by balancing the load of physical machines. The utilization of resources has increased by VMA-ACO with minimal SALV. One of the main disadvantages of ACO is its slow convergent. The authors proposed the "PM selection expectation" parameter to overcome the drawback. Yadav et al. [44] proposed the Minimum Sum of CPU Utilization and Memory Size (MSCM) based VM selection policy. In MSCM, the host utilization and the total number of assigned VMs for that host was computed. Then, VMs are sorted in increasing order using their CPU utilization and memory. Now, the authors calculated the host upper utilization threshold, and all the VMs were selected for migration who can bring down its utilization below the upper threshold.

J. Thaman et al. [45] proposed Variance minimization-based selection (Var_Sel) policy. Var_Sel is based on unifying the utilization across the hosts. It selects a VM, which reduced the mean square deviation of the excess load of hosts. The authors proposed a variance-based heuristics approach that selects VMs for migration.

III. Proposed System Model

Consider a large-scale data centers consist of '*m*' overloaded hosts and 1n1 Virtual Machines (VMs). The proposed *MdSSUR* VM selection policy selects some VMs from the overloaded host to perform VM migration with optimal SLAV and reduced EC. The following factors are the base of the proposed *MdSSUR* VM selection policy:

- Compute the Sum of Squares Utilization Rate (*SSUR*) of an Overloaded Host using RAM, Band Width, and MIPS.
- Compute the Remaining Sum of Squares Utilization Rate (*RSSUR*) of an Overloaded Host by excluding one VM re-sources from that Overloaded Host and find out the difference Sum of Squares Utilization Rate (*dSSUR*).
- Select VM with Maximum value among the differences of the Sum of Squares Utilization Rate (*MdSSUR*) to perform VM migration.

Table I sums up the abbreviations of the terms defined in this section.

TABLE I. Abbreviations and Full Names

Abbreviation	Full Name
urRam	Utilization rate of RAM
urBw	Utilization rate of Band Width
urMIPS	Utilization rate of CPU in MIPS
urAvg.	Average utilization
sdurRam	Squared Difference Utilization Rate of RAM
sdurBw	Squared Difference Utilization Rate of Band Width
sdurMIPS	Squared Difference Utilization Rate of CPU in MIPS
ssur	Sum of Squares Utilization Rate
rurRam	Remaining Utilization rate of RAM
rurBw	Remaining Utilization rate of Band Width
rurMIPS	Remaining Utilization rate of CPU in MIPS
rurAvg.	Remaining Average utilization
rsdurRam	Remaining Squared Difference Utilization Rate of RAM
rsdurBw	Remaining Squared Difference Utilization Rate of Band Width
rsdurMIPS	Remaining Squared Difference Utilization Rate of CPU in MIPS
rssur	Remaining Sum of Squares Utilization Rate
dSSRU	Difference Sum of Squares Utilization Rate
udMax	Maximum difference Utilization

A. Sum of Squares Utilization Rate (SSUR) of an Overloaded Host

The Sum of Squares Utilization Rate (*SSUR*) of an overloaded host has computed using the utilization resources like RAM, Band Width, and MIPS. Algorithm 1 is used to find the *SSUR* of an overloaded

host. The Utilization Rate (*UR*) of Ram, Band Width, and MIPS of an overloaded host computed using line number 2, 3, and 4 of algorithm 1, respectively. The Squared Difference Utilization Rate (*SDUR*) of Ram, Band Width, and MIPS estimated using line number 6, 7, and 8 of algorithm 1, respectively. Finally, The Sum of Squares Utilization Rate (*SSUR*) of that overloaded host is computed using line number 9 of algorithm 1.

Algorithm 1: Sum of Squares Utilization Rate (SSUR) of an Overloaded Host

Input: $host_j = j^{th}$ overloaded Host

Output: *ssur* = Sum of Squares Utilization Rate

1 start
2 $urRam \leftarrow host_j.usedRam \div host_j.totalRam$
3 $urBw \leftarrow host_j.usedBw \div host_j.totalBw$
4 $urMIPS \leftarrow 1 - (host_j.availableMIPS \div host_j.totalMIPS)$
5 $urAvg. \leftarrow (urRam + urBw + urMIPS) \div 3$
6 $sdurRam \leftarrow (urAvg. - urRam)^2$
7 $sdurBw \leftarrow (urAvg. - urBw)^2$
8 $sdurMIPS \leftarrow (urAvg. - urMIPS)^2$
9 $ssur \leftarrow sdurRam + sdurBw + sdurMIPS$
10 **return** *ssur*
11 stop

B. Remaining Sum of Squares Utilization Rate (RSSUR) of an Overloaded Host

The Remaining Sum of Squares Utilization Rate (*RSSUR*) of an Overloaded Host has been computed by excluding one VM's resources from that host. Algorithm 2 is used to find the *RSSUR* of an overloaded host. The utilization rate of Ram, Band Width, and MIPS excluding one VM's resources of the overloaded host is computed by using line number 2, 3, and 4 of algorithm 2, respectively.

Algorithm 2: Remaining Sum of Squares Utilization Rate (*RSSUR*) of an Overloaded Host

Input: $host_j = j^{th}$ overloaded Host, $vm_i = i^{th}$ Virtual Machine of j^{th} overloaded Host

Output: *rssur* = Remaining Sum of Squares Utilization Rate

1 start
2 $rurRam \leftarrow (host_j.usedRam - vm_i.Ram) \div host_j.totalRam$
3 $rurBw \leftarrow (host_j.usedBw - vm_i.Bw) \div host_j.totalBw$
4 $rurMIPS \leftarrow 1 - ((host_j.availableMIPS - vm_i.MIPS) \div host_j.totalMIPS)$
5 $rurAvg. \leftarrow (rurRam + rurBwr + urMIPS) \div 3$
6 $rsdurRam \leftarrow (rurAvg. - rurRam)^2$
7 $rsdurBw \leftarrow (rurAvg. - rurBw)^2$
8 $rsdurMIPS \leftarrow (rurAvg. - rurMIPS)^2$
9 $rssur \leftarrow rubfRam + rubfBw + rubfMIPS$
10 **return** *rssur*
11 stop

The Remaining Squared Difference Utilization Rate (*RSDUR*) of Ram, Band Width, and MIPS of that overloaded host excluding one VM's resources computed using line number 6, 7, and 8 of algorithm 2, respectively. Finally, The Remaining Sum of Squares Utilization Rate (*RSSUR*) of an overloaded host excluding one VM's resources computed using line number 9 of algorithm 2.

C. Maximum Difference Sum of Squares Utilization Rate (MdSSUR) VM Selection Policy

The proposed *MdSSUR* VM selection policy described in Algorithm 3. The set of active hosts are the input in the proposed *MdSSUR* VM selection policy. If any active host is overloaded, find out all allocated VMs of that host and add it to the *migratableV ms* list using line number 4. Initially, the *SelectedV m* is set as *NULL* using line number 5, and the Maximum difference utilization (*udMax*) is set by the minimum value using line number 6. Now, compute the *SSUR* of the overloaded host using line number 7. The line number 7 has called algorithm 1 to compute the *SSUR* of that overloaded. Then, for each VM, estimate the *RSSUR* of the overloaded host by excluding each VM's resources using line number 10. The line number 10 has called algorithm 2 to estimate the *RSSUR* of that overloaded host by excluding the resource of each VM. The difference between *SSUR* and *RSSUR* is computed as a difference Sum of Squares Utilization Rate (*dSSUR*) using line number 11. The value of udMax is set as a minimum. Therefore, if the condition of line number 12 becomes true, then the VM for which the value of *dSSUR* is maximum is assigned in *udMax*, and that VM is assigned in *SelectedV* m by line number 13 and 14, respectively. Finally, *SelectedV* m is added to the migration list using line number 18.

Algorithm 3: Proposed MdSSUR VM Selection Policy

Input: *host_list* = set of Active Hosts

Output: *V MsT oMigrateList* = List of Selected VMs needs to be Migrated

```
1  start
2  for each host, in host_list do
3    if (isHostOverloaded(host)) then
4      migratableV ms ← host.AllocatedV Ms()
5      SelectedV m ← NULL
6      udMax ← Double.MinV alue()
7      ssur ← host.SSUR()
8      for each vm, in migratableV ms do
9        if !(isInMigration(vm) then
10         rssur ← RSSUR(host, vm)
11         dSSUR ← ssur − rssur
12         if (dSSUR > udMax) then
13           udMax ← dSSUR
14           SelectedV m ← vm
15         end
16       end
17     end
18     V MsT oMigrateList.add(SelectedV m)
19   end
20 end
```

The difference Sum of Squares Utilization Rate (*dSSUR*) can also be represented by Eq. 1.

$$dSSUR_i = SSUR_{host_j} - RSSUR_{host_j}^{Vm_i} \quad (1)$$

where, $host_j$ is the j^{th} host from the overloaded *host_list*, *V* m_i is the i^{th} VM from the *V m_list* which has been allocated to the $host_j$, $SSUR_{hostj}$ is the Sum of Squares Utilization Rate of $host_j$, $RSSUR_{host_j}^{Vm_i}$ is the Remaining Sum of Squares Utilization Rate of $host_j$ excluding V_{mi} resources, and $dSSUR_i$ is the i^{th} difference of the Sum of Squares Utilization Rate. The VM which has Maximum *dSSUR* is selected form the *V m_list* using Eq. 2.

$$SelectedVm = \max_{\forall Vm_i \in Vm\_list} \{dSSUR_i\} \quad (2)$$

Now, the *SelectedV m* is added to the migration list, and the process will be continued until all overloaded hosts are examined.

IV. Energy Efficient and SLA Aware VM Consolidation Based on *MdSSUR* Policy

VM consolidation comprises the following steps: A) Detection of Overloaded Host, B) Detection of Underloaded Host, C) VM Selection, and D) VM Placement.

A. Overload Host Detection

Overload host detection is the first step of VM consolidation. Initially, the authors [19] set a threshold value of 0.9, and if the CPU utilization of any host is more than the threshold value, then the host is marked as over-utilized. Then, using Linear Regression Robust (LRR), future CPU utilization is predicted. In [19], the authors proposed the LRR prediction model to overcome the disadvantage of Linear Regression (LR) [19] prediction model. The LR prediction model is based on the Loess method proposed by Cleveland [46]. The LR prediction model is vulnerable to outliers due to heavy-tailed distributions. To overcome the Loess method's disadvantage and make it robust, Cleveland proposed the least-squares method [47]. If the multiplication of predicted CPU utilization and safety parameter is greater than or equal to one, the host is marked as over-utilized. For evaluating the proposed *MdSSUR* VM selection policy, the LRR has been used to detect the over-utilized hosts.

B. Underload Host Detection

In [19], the authors proposed an iterative process to determine the underloaded hosts. After the migration of VMs from overloaded hosts to moderately loaded hosts, the underutilized host detection processes start. The system finds the minimum utilized host by comparing it with the other hosts. All the VMs from an underloaded host migrated to the moderately loaded hosts keeping them as not overloaded.

C. VM Selection

Now, select some VMs from overloaded hosts and all VMs from underloaded hosts in this step. The proposed *MdSSUR* VM selection policy described in Algorithm 3 is used to select the VMs from overloaded hosts and added to the migration list. All the VMs from underloaded hosts are selected and added to the migration list to perform the VM migration.

D. VM Placement

After the migration, overloaded hosts will become moderately loaded hosts, and underloaded hosts will be in sleep mode. All the migratable VMs must be placed in some moderately loaded hosts based on some VM placement policy. In this research work, the chosen VM placement policy is the Power-Aware Best Fit Decreasing (PABFD) [19] placement policy for the evaluation of the proposed *MdSSUR* VM selection policy. In PABFD policy, all migratable VMs are sorted decreasingly based on CPU utilization. Each VM has been allocated into a host that required minimum power consumption due to the allocation.

V. Performance Evaluation

A. Experimental Setup

One of the main aspects of the proposed *MdSSUR* VM selection approach is to reduce the total number of VM migrations. The migration list is prepared based on selecting a VM where the difference Sum of Squares Utilization Rate (*dSSUR*) is maximum in an overloaded host and performs VM migration in moderately loaded hosts. It will keep the overloaded hosts undercontrol, and energy consumption by the host will be reduced. The proposed *MdSSUR* VM selection policy will significantly reduce the number of migrations. As a result, it will substantially minimize SLA violations. CloudSim [18], [48], [49] toolkit is the most popular simulator used for large-scale virtualized cloud applications. It provides a stronger virtualized model of cloud architecture compare to other simulators. It supports dynamic resource management and scalability.

A data center containing 800 heterogeneous hosts is used to evaluate the proposed MdSSUR VM selection policy, 50% of them are re HP ProLiant ML110 G4 servers 245 clocked at 1,860 1860MHz, and the remaining are HP ProLiant ML110 G5 servers clocked at 2,660 MHz. Each one has two cores, 4 GB memory, 1 GB/s network bandwidth. Table II is to show the characteristics of the hosts. The hosts' energy consumption characteristics are given in Table III.

TABLE II. Characteristics of Hosts [19]

Host	Clock Speed	Cores	RAM	Bandwidth
G4	1860 MHz	2	4 GB	1 Gbps
G5	2660 MHz	2	4 GB	1 Gbps

The standard Amazon EC2 [51] has been used for the VM instances. Four different types of VM are available. One of the VM instances is created into the host, depending on the requirement of the workload. Table IV is to show the characteristics of the VMs.

TABLE IV. Characteristics of VMs [50]

VM Instances	Clock Speed	Cores	RAM
Micro Instance	500 MHz	1	613 MB
Small Instance	1000 MHz	1	1740 MB
Extra large Instance	2000 MHz	1	1740 MB
High-CPU Medium Instance	2500 MHz	1	870 MB

1. Workload

The experiment has been run using real-life workload traces to make simulation-based approaches more acceptable. Planet-Lab [52] has collected these workload traces from an infrastructure monitoring framework, called CoMon [53]. These traces consist of the CPU utilization data by more than a thousand VMs from many servers located over 500 different places in the world. After every 300 seconds, the utilization values were recorded. During March and April of 2011, ten random dates were chosen from the workload traces. Between them, four days of data is selected for the evaluation of the proposed *MdSSUR* VM selection policy. Table V is to show the characteristics of each workload.

TABLE III. Energy Consumption of PMs At Different Load [19]

PM	Energy Consumption (in Watts) at Different Load on Hosts										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
HP ProLiant ML110 G4	86	89.6	92.6	96	99.5	102	106	108	121	114	117
HP ProLiant ML110 G5	93.7	97	101	105	110	116	121	125	129	133	135

TABLE V. Characteristics of Workload [19]

Date	No. of VMs	No. of Hosts	Mean (%)	St. dev. (%)
03-03-2011	1052	800	12.31	17.09
06-03-2011	898	800	11.44	16.83
03-04-2011	1463	800	12.39	16.55
20-04-2011	1033	800	10.43	15.21

B. Result & Analysis

The performance evaluation of the proposed *MdSSUR* VM selection policy has been measured and compared with some classical VM selection algorithms like Minimum Migration Time (MMT) [19], Maximum Correlation (MC) [19], Minimum Utilization (MU) [19], Random Selection (RS) [19], Maximum Utilization Minimum Size (MuMs) [20], Maximum Migration Time (MxMT) [21], and Minimize Number and Cost of Migrations (MNCM) [29]. These policies are previously mentioned in Section II. The Beloglazov et. al.'s [19] proposed metrics have been used to measured and compared the effectiveness of the proposed *MdSSUR* VM selection policy.

1. Performance Degradation Due to Migration (PDM)

Performance degradation due to migration (PDM) is an SLA-based metric. It is represented in Eq. 3. Excessive VM migration may degrade performance. Fig. 1 shows the comparative analysis of PDM of the proposed *MdSSUR* VM selection policy.It indicates that the PDM of the proposed *MdSSUR* VM selection policy is very significantly lesser than other VM selection policies.

$$PDM = \frac{1}{M}\sum_{j=1}^{M}\frac{C_{Deg_j}}{C_{CPU_j}} \tag{3}$$

where,

- M is the number of Virtual machines.
- C_{Deg_j} is the performance degradation of VM j due to migration.
- C_{CPU_j} is the total capacity requested by VM j during its life time.

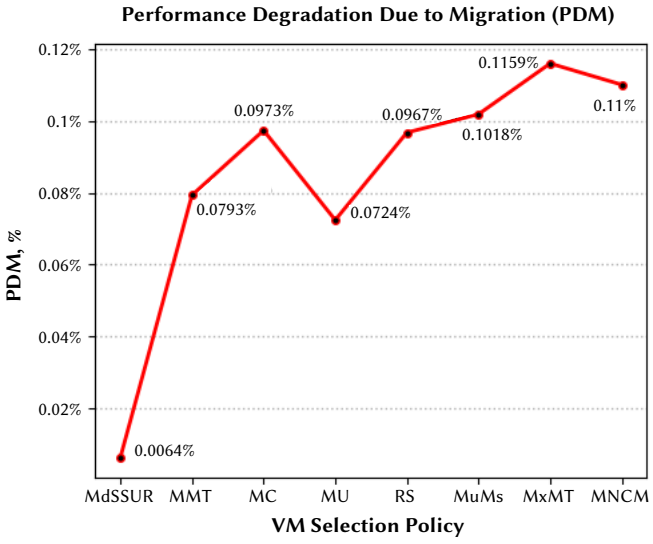


Fig. 1. Performance Degradation due to Migration (PDM) comparison.

2. Service Level Agreement Violation (SLAV)

Service Level Agreement violation (SLAV) is one of the most important metrics. SLAs [54] contains several parameters to satisfy MC users. So, the level of QoS is measured by and reduced by SLA violation. SLAV is calculated by Eq. 4.

$$SLAV = SLATH \times PDM \tag{4}$$

where,

- *SLAV* is a percentage violation of Service Level Agreement.
- *SLATH* is the duration of the 100% CPU use of an active host.
- *PDM* is the performance degradation during VMs migration in percentage.

Fig. 2 is to show the comparative analysis of SLAV, and our proposed VM selection policy has reduced SLAV by 79.14% on an average compare to other approaches.

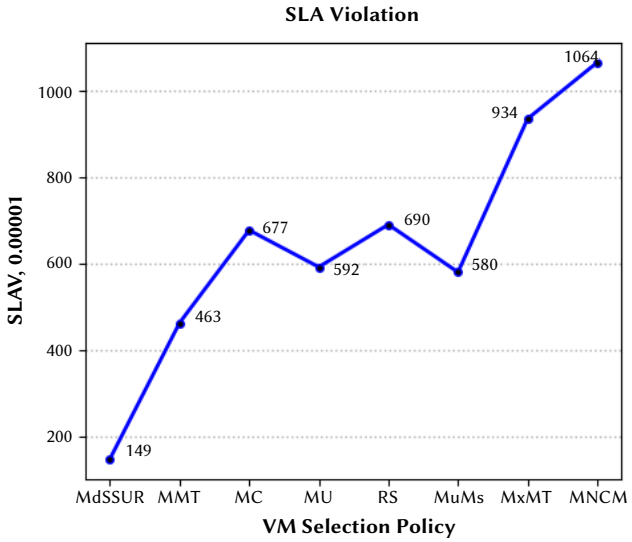


Fig. 2. SLA violation (SLAV) comparison.

3. Number of VM Migrations (NVMG)

Migration of VM is an expensive process. The VM manager initiates VM migration during the VM placement at each time frame. The migrated VMs will possess some CPU time and network bandwidth on both source hosts and targeted hosts. So, VM migration may detrimentally influence the performance of hosts. It may increase the EC and SLA violation of the data center. Therefore, a limited number of VM migration is more desirable. A limited number of VM migration can reduce the total cost of the operation requested by MC users. It can also reduce the total EC and SLA violation of the cloud data centers.

Fig. 3 is to show the comparative analysis of the Number of VM migrations. The total number of migration of the proposed VM selection policy is 2.41. The average number of migration of the other compared policies is 22.89. The proposed VM selection policy has reduced VM migrations by 89.47% on average compared to other approaches mentioned above.

4. Number of Host Shutdowns (HSD)

The number of host shutdowns is a migration based metric. An enormous number of VM migration can increase host shutdowns and the energy consumption of MC data centers. If specific hosts are repeatedly switched on and off, it may increase the MC data centers' EC and operational cost.

Fig. 4 is to show the comparative analysis of the number of host shutdowns of the proposed *MdSSUR* VM selection policy.

From Fig. 4, the proposed VM selection strategy has clearly limited the number of host shutdowns.

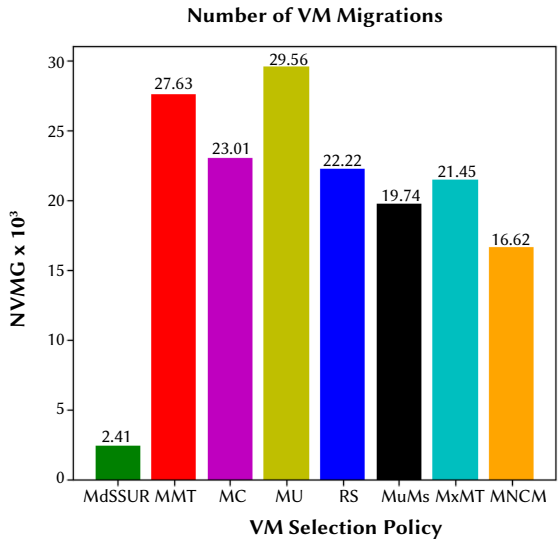


Fig. 3. Number of VM migrations (NVMG) comparison.

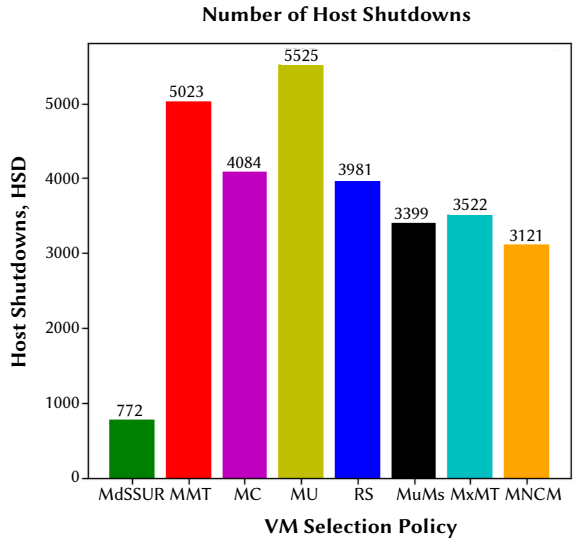


Fig. 4. Number of host shutdowns (HSD) comparison.

5. Total Energy Consumption (EC)

Nowadays, Total energy consumption becomes a key concern to researchers. Reducing MC data centers' energy consumption with optimal SLA violations has become the main objective of the researchers because it has a massive impact on environments.

Fig. 5 is to show the comparative analysis of Total energy consumption. Fig. 5 indicates that our proposed VM selection policy has reduced energy consumption by 28.37% on an average compare to other baseline policies.

6. Energy and SLA Violation (ESV)

The EC and SLA violations of the MC data centers are the essential matrices. However, EC and SLAV are negatively correlated. The EC of MC data centers may typically be minimized by the expense of an increased amount of SLA violations. The resource management system is aimed at reducing EC and the SLA violations of MC data centers. It can be computed by Eq. 5. So, ESV established a relation between two negatively correlated matrices.

$$ESV = EC \times SLAV \tag{5}$$

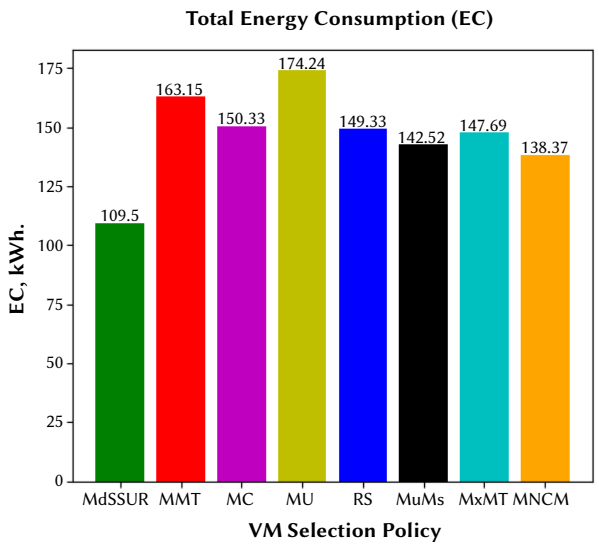


Fig. 5. Total energy consumption (EC) comparison.

Fig. 6 is to show the comparative analysis of ESV. The ESV of the proposed MdSSUR VM selection policy is 1.62, and the average ESV of the other compared policies is 10.75. It indicates that our proposed VM selection has reduced ESV by 84.93% on an average compare to other approaches.

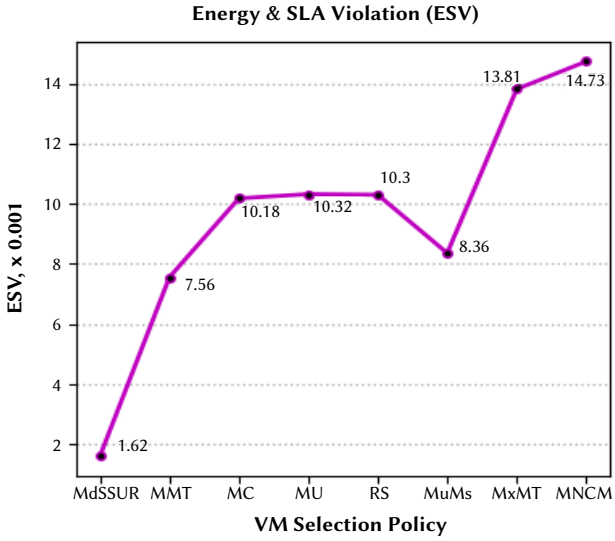


Fig. 6. Energy and SLA Violation (ESV) comparison.

7. Total Execution Time (ET)

Total execution time (ET) is the time to complete an algorithm for a given workload. It determines the efficiency of the algorithms in terms of time. So, the throughput of the MC user's request depends on the total execution time. If the total execution time can be minimized, the throughput of the MC user's request will increase. It can also reduce the operational cost of the MC user's request.

Fig. 7 shows the comparative analysis of total execution time, indicating that the proposed MdSSUR VM selection policy is much faster than other algorithms. The average execution time of the proposed MdSSUR VM selection policy is 50.1 milliseconds, and the average execution time of the other compared policies is 341.1 millisecond. So, the total execution time significantly increased by the proposed MdSSUR VM selection to reduce the MC user's request's operational cost.

TABLE VI. COMPARATIVE ANALYSIS OF *MDSSUR* VM SELECTION POLICY WITH RENOWNED VM SELECTION POLICIES

VM Selection Policy	Energy in kWh.	PDM (%)	SLAV $\times 10^{-5}$	Migration $\times 10^{3}$	Host Shutdw.	ESV $\times 10^{-3}$	ExeTime in Milisec.
MdSSUR	109.5	0.0064	149	2.41	772	1.62	50.01
MMT	163.15	0.0793	463	27.632	5023	7.56	409.76
MC	150.33	0.0973	677	23.004	4084	10.18	350.7
MU	174.24	0.0724	592	29.555	5525	10.32	461.77
RS	149.33	0.0967	690	22.223	3981	10.3	318.91
MuMs	142.52	0.1018	580	19.744	3399	8.63	312.51
MxMT	147.69	0.1159	933.5	21.45	3822	13.81	304.91
MNCM	138.37	0.11	1064	16.62	3121	14.73	229.09

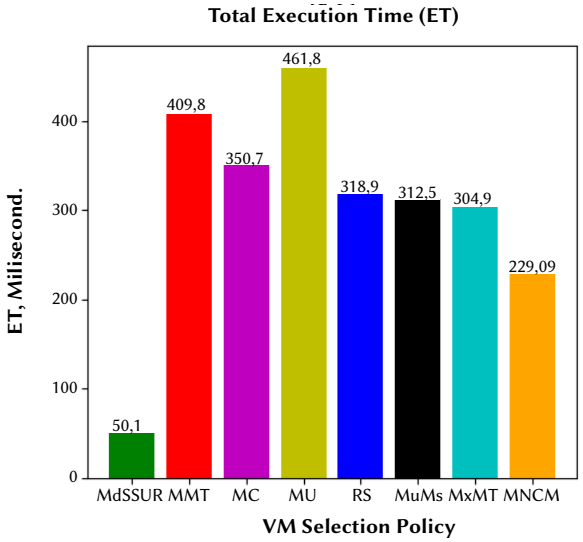


Fig. 7. Total Execution Time (ET) comparison.

Table VI shows a comparative analysis of all the above mention metrics, and Table VII shows the average improvements rate of the proposed MdSSUR VM selection policy based on the above mention metrics.

TABLE VII. AVERAGE IMPROVEMENT RATE OF *MDSSUR* POLICY

Metric	Avg. Improvement Rate in %
Energy Consumption	28.37
SLA violation	79.14
Number of Migration	89.47
Energy and SLA violation	84.93

8. Workload Based Analysis

In this section, the evaluation of the proposed *MdSSUR* VM selection policy has been done with different workloads like 20110306, 20110403, and 20110420. These workloads result compared with the metrics like Energy consumption, Number of VM migration, and SLA violation. Fig. 8 depicts energy consumption with different workloads. The average reduction in EC of the proposed MdSSUR VM selection policy is 30.28%, 30.7%, and 33.22% compared to the other mentioned policies using 20110306, 20110403, and 20110420 workloads, respectively. The number of VM migrations mainly controls the overall cost of MC users. The number of VM migrations with various workloads shown in Fig. 9. The number of migrations significantly reduced by the proposed MdSSUR VM selection policy. It decreases on average by 87.42%, 90.98%, and 88.08% compared to the other mentioned VM selection policies using 20110306, 20110403, and 20110420 workloads.

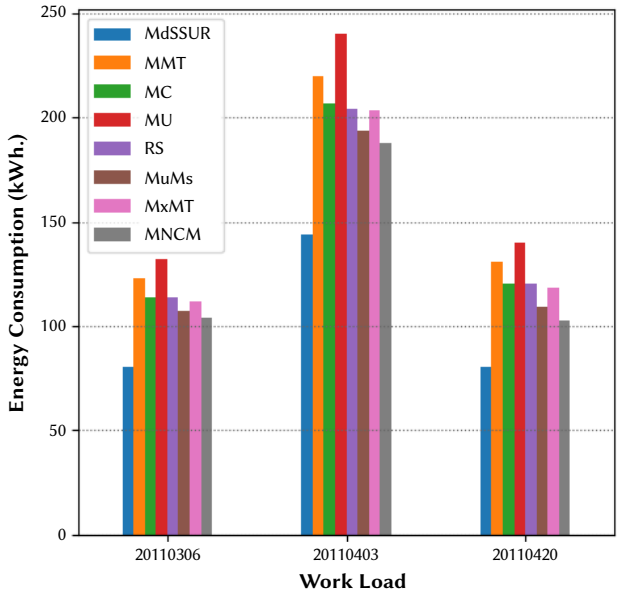


Fig. 8. Total Execution Time (ET) with different workloads.

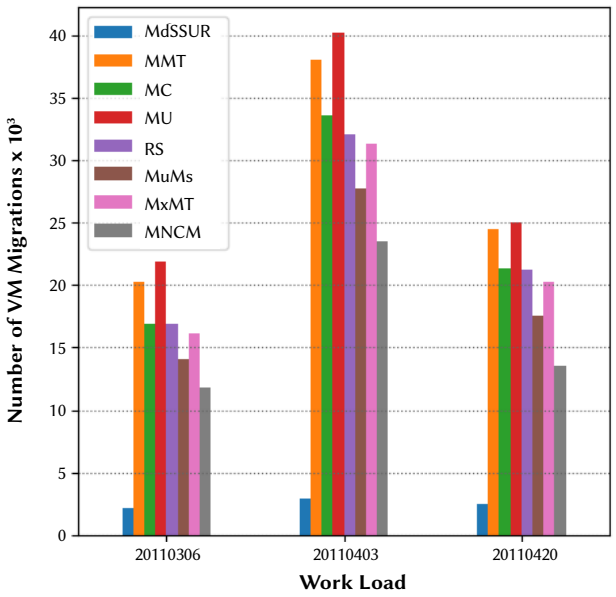


Fig. 9. Number of Migrations with different workloads.

Minimal service level agreement violation increases the QoS provided by the MC service providers. Fig. 10 is to show the SLAV with different workloads. The average reduction of SLAV by the proposed *MdSSUR* VM selection policy is 70.45%, 78.3%, and 79.18% compared to

the other mentioned policies using 20110306, 20110403, and 20110420 workloads, respectively.

Thus, the proposed *MdSSUR* VM selection policy defeats other benchmark mentioned VM selection policies.

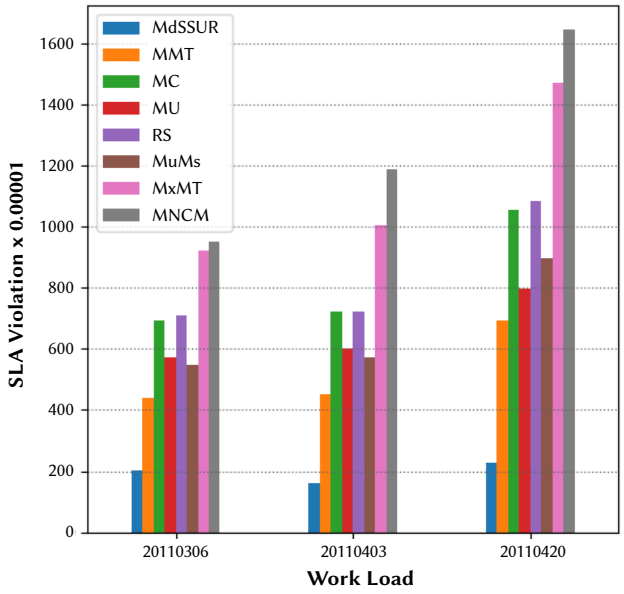


Fig. 10. SLA Violation with different workloads.

VI. Conclusion

Optimal VM selection can decrease VM migrations and increase the throughput of the MC user's request. The emission of greenhouse gases by the MC data centers all over the world needs to be decreased. This paper proposed a novel *MdSSUR* VM selection policy to reduce EC with minimal SLAV. The simulation results have shown that the proposed *MdSSUR* VM selection policy will scale back SLAV and enhance the system performance considerably whereas saving energy. This research plans to incorporate with the Internet of Things (IoT) to enhance the Cloud of Things (CoT) environment.

Acknowledgement

We want to offer our sincere gratitude to Dr. Rajkumar Buyya, one of the renowned researchers behind the inventions and progressive research directions towards cloud computing. One of his excellent contributions to cloud simulation is CloudSim, which has helped thousands of researchers to test different cloud computing algorithms very rapidly. His research work has motivated us to develop the *MdSSUR* VM selection policy.

References

[1] N. Williams, G. S. Blair, "Distributed multimedia applications: A review," *Computer Communications*, vol. 17, no. 2, pp. 119–132, 1994.

[2] M. N. Birje, P. S. Challagidad, R. H. Goudar, M. T. Tapale, "Cloud computing review: concepts, technology, challenges and security," *International Journal of Cloud Computing*, vol. 6, no. 1, pp. 32–57, 2017.

[3] W. Zhu, C. Luo, J. Wang, S. Li, "Multimedia cloud computing," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 59–69, 2011.

[4] A. Vogel, B. Kerherve, G. von Bochmann, J. Gecsei, "Dis-tributed multimedia and QoS: A survey," *IEEE multimedia*, vol. 2, no. 2, pp. 10–19, 1995.

[5] Masanet, Shehabi, Smith, Lei, "Global Data Center Energy Use: Distribution, Composition, and Near-Term Outlook," *Northwestern University: Evanston*, IL, USA, 2018.

[6] E. Innovation, "How Much Energy Do Data Centers Really Use?." https://energyinnovation.org/2020/03/17/how-much-energy-do-data-centers-really-use/, 2020(ac-cessed on 11th September, 2020).

[7] F. Pearce, "Energy Hogs: Can World's Huge Data Centers Be Made More Efficient?." https://e360.yale.edu/features/energy-hogs-can-huge-data-centers-be-made-more-efficient, 2018(accessed on 18th September, 2020).

[8] ATAG, "Air Transport Action Group: Facts and Figure." https://atag.org/facts-figures.html, 2019(accessed on 17th September, 2020).

[9] F. Lombardi, R. Di Pietro, "Secure virtualization for cloud computing," Journal of network and computer applications, vol. 34, no. 4, pp. 1113–1122, 2011.

[10] Y. Xing, Y. Zhan, "Virtualization and cloud computing," Future Wireless Networks and Information Systems, pp. 305–312, 2012.

[11] N. Jain, S. Choudhary, "Overview of virtualization in cloud computing," *Symposium on Colossal Data Analysis and Networking*, pp. 1–4, 2016.

[12] A. Abdelsamea, E. E. Hemayed, H. Eldeeb, H. Elazhary, "Virtual machine consolidation challenges: A review," International Journal of Innovation and Applied Studies, vol. 8, no. 4, p. 1504, 2014.

[13] M. A. Khan, A. Paplinski, A. M. Khan, M. Murshed, R. Buyya, "Dynamic virtual machine consolidation algorithms for energy-efficient cloud resource management: a review," *Sustainable cloud and energy services*, pp. 135–165, 2018.

[14] H. Wang, H. Tianfield, "Energy-aware dynamic virtual machine consolidation for cloud datacenters," *IEEE Access*, vol. 6, pp. 15259–15273, 2018.

[15] A. Beloglazov, R. Buyya, "System, method and computer program product for energy-efficient and service level agreement (SLA)-based management of data centers for cloud computing," *US Patent 9,363,190, Google Patents*, June 7, 2016.

[16] P. G. J. Leelipushpam, J. Sharmila, "Live VM migration techniques in cloud environment–a survey," *IEEE Conference on Information & Communication Technologies*, pp. 408–413, 2013.

[17] A. Choudhary, M. C. Govil, G. Singh, L. K. Awasthi, E. S. Pilli, D. Kapil, "A critical survey of live virtual machine migration techniques," *Journal of Cloud Computing*, vol. 6, no. 1, p. 23, 2017.

[18] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Software: Practice and experience*, vol. 41, no. 1, pp. 23–50, 2011.

[19] A. Beloglazov, R. Buyya, "Optimal online deterministic algo-rithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in cloud data centers," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397–1420, 2012.

[20] R. Yadav, W. Zhang, H. Chen, T. Guo, "Mums: Energy-aware vm selection scheme for cloud data center," *28th International Workshop on Database and Expert Systems Applications, IEEE*, pp. 132–136, 2017.

[21] N. Akhter, M. Othman, R. K. Naha, "Energy-aware virtual machine selection method for cloud data center resource allocation," *arXiv preprint arXiv:1812.08375*, 2018.

[22] A. Beloglazov, R. Buyya, "Energy efficient resource management in virtualized cloud data centers," *10th IEEE/ACM In-ternational Conference on Cluster, Cloud and Grid Comput-ing, IEEE*, pp. 826–831, 2010.

[23] Q. Deng, D. Meisner, L. Ramos, T. F. Wenisch, R. Bianchini, "Memscale: active low-power modes for main memory," *ACM SIGPLAN Notices*, vol. 46, no. 3, pp. 225–238, 2011.

[24] Q. Deng, D. Meisner, A. Bhattacharjee, T. F. Wenisch, R. Bianchini, "MultiScale: memory system DVFS with multiple memory controllers," *Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design*, pp. 297–302, 2012.

[25] J. C. W. Lin, Y. Shao, Y. Djenouri, U. Yun, "ASRNN: A recurrent neural network with an attention model for sequence labeling," *Knowledge-Based Systems*, p. 106548, 2020.

[26] R. Mandal, M. K. Mondal, S. Banerjee, U. Biswas, "An approach toward design and development of an energy-aware VM selection policy with improved SLA violation in the domain of green cloud computing," *The Journal of Supercomputing*, pp. 1–20, 2020.

[27] R. Yadav, W. Zhang, O. Kaiwartya, P. R. Singh, I. A. Elgendy, Y.C. Tian, "Adaptive energy-aware algorithms for minimizing energy consumption

and SLA violation in cloud computing," *IEEE Access*, vol. 6, pp. 55923–55936, 2018.

[28] J. C. W. Lin, L. Yang, P. Fournier-Viger, J. M.-T. Wu, T. P. Hong, L. S. L. Wang, J. Zhan, "Mining high-utility itemsets based on particle swarm optimization," *Engineering Applications of Artificial Intelligence*, vol. 55, pp. 320–330, 2016.

[29] C. Zhang, Y. Wang, Y. Lv, H. Wu, H. Guo, "An Energy and SLA-Aware Resource Management Strategy in Cloud Data Centers," *Scientific Programming*, vol. 2019, 2019.

[30] H. Toumi, B. Marzak, A. Talea, A. Eddaoui, M. Talea, "Use Trust Management Framework to Achieve Effective Security Mechanisms in Cloud Environment," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 4, no. 3, 2017.

[31] Y. Wen, Z. Li, S. Jin, C. Lin, Z. Liu, "Energy-efficient virtual resource dynamic integration method in cloud computing," *IEEE Access*, vol. 5, pp. 12214–12223, 2017.

[32] X. Wu, Y. Zeng, G. Lin, "An energy efficient VM migration algorithm in data centers," *16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science, IEEE*, pp. 27–30, 2017.

[33] J. C. W. Lin, G. Srivastava, Y. Zhang, Y. Djenouri, M. Alo-qaily, "Privacy Preserving Multi-Objective Sanitization Model in 6G IoT Environments," *IEEE Internet of Things Journal*, 2020.

[34] V. K. Solanki, M. Venkatesan, S. Katiyar, "Conceptual Model for Smart Cities: Irrigation and Highway Lamps using IoT," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 4, no. 3, pp. 28–33, 2017.

[35] S. B. Melhem, A. Agarwal, N. Goel, M. Zaman, "Minimizing biased VM selection in live VM migration," *3rd International Conference of Cloud Computing Technologies and Applications, IEEE*, pp. 1–7, 2017.

[36] S. M. Moghaddam, M. O'Sullivan, C. Walker, S. F. Piraghaj, C. P. Unsworth, "Embedding individualized machine learning prediction models for energy efficient VM consolidation within Cloud data centers," *Future Generation Computer Systems*, vol. 106, pp. 221–233, 2020.

[37] H. Peng, L. Liu, L. Ma, W. Zhao, H. Ma, L. Yuntao, "Approximate Error Estimation based Incremental Word Representation Learning," *Data Science and Pattern Recognition*, vol. 4, 2020.

[38] S. A. Makhlouf, B. Yagoubi, "Data-Aware Scheduling Strategy for Scientific Workflow Applications in IaaS Cloud Computing," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 5, no. 4, 2019.

[39] J. S. Pan, X. Wang, S.-C. Chu, T. Nguyen, "A multi-group grasshopper optimisation algorithm for application in capacitated vehicle routing problem," *Data Science and Pattern Recognition*, vol. 4, no. 1, pp. 41–56, 2020.

[40] K. W. Huang, C. C. Lin, Y.-M. Lee, Z.-X. Wu, "A deep learn-ing and image recognition system for image recognition," *Data Science and Pattern Recognition*, vol. 3, no. 2, pp. 1–11, 2019.

[41] M. El Ghazouani, E. Kiram, M. Ahmed, L. Er-Rajy, Y. El Khanboubi, "Efficient Method Based on Blockchain Ensuring Data Integrity Auditing with Deduplication in Cloud," *International Journal of Interactive Multimedia & Artificial Intelligence*, vol. 6, no. 3, 2020.

[42] P. Xu, G. He, Z. Li, Z. Zhang, "An efficient load balancing algorithm for virtual machine allocation based on ant colony optimization," *International Journal of Distributed Sensor Networks*, vol. 14, no. 12, p. 1550147718793799, 2018.

[43] M. Dorigo, M. Birattari, T. Stutzle, "Ant colony optimiza-tion," *IEEE computational intelligence magazine*, vol. 1, no. 4, pp. 28–39, 2006.

[44] S. B. S. Yadav, M. Kalra, "Energy-Aware VM Migration in Cloud Computing," *Proceedings of International Conference on IoT Inclusive Life, NITTTR Chandigarh, India*, Springer, pp. 353–364, 2020.

[45] J. Thaman, M. Singh, "SLA conscious VM migration for host consolidation in cloud framework," *International Journal of Communication Networks and Distributed Systems*, vol. 19, no. 1, pp. 46–64, 2017.

[46] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association*, vol. 74, no. 368, pp. 829–836, 1979.

[47] W. S. Cleveland, *Visualizing data*. Hobart Press, 1993.

[48] R. Buyya, R. Ranjan, R. N. Calheiros, "Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities," *International conference on high performance computing & simulation, IEEE*, pp. 1–11, 2009.

[49] T. Goyal, A. Singh, A. Agrawal, "Cloudsim: simulator for cloud computing infrastructure and modeling," *Procedia Engineering*, vol. 38, pp. 3566–3572, 2012.

[50] A. Beloglazov, *Energy-efficient management of virtual machines in data centers for cloud computing*. PhD dissertation, 2013.

[51] A. EC, "Amazon EC2 instance types." https: //aws.amazon.com/ec2/instance-types/, 2019(accessed on 21st September 2020).

[52] L. Peterson, A. Bavier, M. E. Fiuczynski, S. Muir, "Experiences building planetlab," *Proceedings of the 7th symposium on Operating systems design and implementation*, pp. 351–366, 2006.

[53] K. Park, V. S. Pai, "CoMon: a mostly-scalable monitoring system for PlanetLab," *ACM Special Interest Group in Operating Systems Review*, vol. 40, no. 1, pp. 65–74, 2006.

[54] R. Sahal, M. H. Khafagy, F. A. Omara, "A survey on SLA management for cloud computing and cloud-hosted big data analytic applications," *International Journal of Database Theory and Application*, vol. 9, no. 4, pp. 107–118, 2016.

Nirmal Kr. Biswas

Nirmal Kr. Biswas is currently working as an Assistant Professor at Department of Computer Science & Engineering of Global Institute of Management and Technology, Krishnagar, West Bengal. He has completed Master of Technology (M.Tech) in Computer Science & Engineering in 2010 from Kalyani Government Engineering College. He has completed his Bachelor of Technology (B.Tech) in Information Technology in 2008 from Murshidabad College of Engineering and Technology. His research interests are in Green Cloud Computing, Cloud Computing, Multimedia Computing.

Dr. Sourav Banerjee

Sourav Banerjee achieved Ph.D degree in Computer Science and Engineering from the University of Kalyani in 2018. He completed his B.E in Computer Science and Engineering in the year 2004 and M.Tech in Computer Science and Engineering in 2006. He is currently an Assistant Professor at Department of Computer Science and Engineering of Kalyani Government Engineering College at Kalyani, West Bengal, India. He has authored numerous reputed non-paid SCI journal articles, book chapters and International conferences. He has edited book on Green Cloud Computing. His research interests include Big Data, Blockchain, Cloud Computing, Green Cloud Computing, Cloud Robotics, Distributed Computing and Mobile Communications, IoT. He is a member of IEEE, ACM, IAE and MIR Labs as well. He is a SIG member of MIR Lab, USA. He has published papers in various reputed journals such as, Wireless Personal Communications, Automatic Control and Computer Sciences, Arabian Journal for Science and Engineering, Journal of King Saud University - Computer and Information Sciences, Service Oriented Computing and Applications, Journal of Super computing, Peer-to-Peer Networking and Applications. He is an Editorial board member of Sustainability Journal (SCI, IF:2.576), Wireless Communication Technology. He is the reviewer of IEEE Transactions on Cloud Computing, Wireless Personal Communications,Journal of Ambient Intelligence and Humanized Computing, Journal of Computer Science,Journal of Super computing, Future Internet,International Journal of Computers and Applications, Personal and Ubiquitous Computing, Innovations in Systems and Software Engineering, Journal of Information Security and Applications, etc. He is connected with various international events, like,Workshop on Security and Privacy in Distributed Ledger Technology (IEEE SP-DLT).He is a lead guest editor of Complex and Intelligent Systems Journal (Springer, SCI, IF: 3.79) for the Special Issue on "Advancement and Trends in Green Cloud Computing, Block chain and IoT for Modern Applications and Systems". He is a guest editor of Mathematical Biosciences and Engineering journal (SCI, IF: 1.285) for the Special Issue on "Recent Advancements in Cognitive Green Computing for Smart Cities". A number of worldwide research scholars are attached with him.

Dr. Utpal Biswas

Utpal Biswas received his B.E, M.E, and Ph.D. degrees in Computer Science and Engineering from Jadavpur University, India in 1993, 2001, and 2008 respectively. From 1994 to 2001, he served as a faculty member in the Department of Computer Science and Engineering at NIT, Durgapur, India. He is currently working as a Professor at the University of Kalyani in the Department of Computer Science and Engineering. At the University of Kalyani, he served as Head of the Department of Computer Science and Engineering. He has also served as Dean, Faculty of Engineering, Technology and Management, University of Kalyani. He has over 200 research articles in different journals, book chapters, and conferences. His research interests include optical communication, ad-hoc and mobile communication, semantic web services, E-governance, Cloud Computing, etc. He is engaged in various state government regulatory committees.

A Generalized Wine Quality Prediction Framework by Evolutionary Algorithms

Terry Hui-Ye Chiu, Chienwen Wu, Chun-Hao Chen*

Department of Information and Finance Management, National Taipei University Technology, Taipei, 106 (Taiwan)

Received 30 October 2020 | Accepted 13 March 2021 | Published 21 April 2021

Abstract

Wine is an exciting and complex product with distinctive qualities that makes it different from other manufactured products. Therefore, the testing approach to determine the quality of wine is complex and diverse. Several elements influence wine quality, but the views of experts can cause the most considerable influence on how people view the quality of wine. The views of experts on quality is very subjective, and may not match the taste of consumer. In addition, the experts may not always be available for the wine testing. To overcome this issue, many approaches based on machine learning techniques that get the attention of the wine industry have been proposed to solve it. However, they focused only on using a particular classifier with a specific set of wine dataset. In this paper, we thus firstly propose the generalized wine quality prediction framework to provide a mechanism for finding a useful hybrid model for wine quality prediction. Secondly, based on the framework, the generalized wine quality prediction algorithm using the genetic algorithms is proposed. It first encodes the classifiers as well as their hyperparameters into a chromosome. The fitness of a chromosome is then evaluated by the average accuracy of the employed classifiers. The genetic operations are performed to generate new offspring. The evolution process is continuing until reaching the stop criteria. As a result, the proposed approach can automatically find an appropriate hybrid set of classifiers and their hyperparameters for optimizing the prediction result and independent on the dataset. At last, experiments on the wine datasets were made to show the merits and effectiveness of the proposed approach.

Keywords

Decision Tree, Genetic Algorithm, Machine Learning, Random Forest, Support Vector Machine, Wine Quality Prediction.

DOI: 10.9781/ijimai.2021.04.006

I. Introduction

WINE has always been an essential part of the dinning culture in western countries. With the booming economy in Asia countries in recent decades, wine consumption has increased even more. From the manufacturer point of view, understanding the wine's quality and creating a steady production is an important goal for the industry. However, testing the quality of the wine is complex and diverse. The wine quality is evaluated in terms of subtlety and complexity [1], ageing potential, stylistic purity, varietal expression, ranking by experts, or consumer acceptance, etc. By excluding the controllable object measures, experts' views are very subjective because it can cause the most considerable influence on both winemakers and how consumers think of the wine's quality [2]. Instead of focusing on how experts qualify the wine, focusing on consumer satisfaction based on collectable scientific data is more useful for the majority of wine producers because understanding the desires of the majority of consumers is essential in the production and sales of wine.

Recording the steps of the wine production procedure is to preserve the quality and knowledge of the whole winemaking process. The collected information is the best tool to guarantee the wine quality. The wine industry has currently established the protected designation of origin (PDO) system [3] with the support of analytical chemistry and chemometric tools to obtain information related to a specific wine. With the improvement of technology both in software and hardware, winemakers started to use the collected data to improve the winemaking technique. Due to the high cost and lack of technological resources, it was difficult for most wine industries to classify the wines based on the chemical components. Many algorithms based on machine learning to assess the quality of wine have gained much attention for the wine industry using another approach to determine what attributes make a "good" wine that the consumers can satisfy with them. For instance, Yeo et al. focused on predicting the wine price using a machine learning technique by using past historical wine price data [4]. For wine production, Ribeiro et al. utilized the linear regression, neuron network and decision tree for predicting the wine vilification [5]. Study in [6] collected the wine dataset on the Cabernet Sauvignon characteristics for the cost-efficient prediction.

In 2009, Cortez et al. collected a wine quality dataset which consists of significant larger instances [7]. Then, three machine learning models, including multiple regression, support vector machine (SVM) and neuron network (NN), are trained using the collected wine dataset. It shows that SVM outperforms the other two methods, and indicates the importance of the correct setting of hyperparameters. Over the years, the wine dataset has been adopted in several studies with various methods such as SVM [8], [9], [10], [11], random forest (RF) [11], [12],

* Corresponding author.
E-mail address: chchen@ntut.edu.tw

[13], [14], [15], decision-tree-based algorithms [13], [15], and NN [5], [8], [9] to predict the quality of the wine based on physiochemical characteristics in the wine. In addition, several pieces of research used feature selection to improve the accuracy of wine quality prediction such as recursive feature elimination, principal component analysis (PCA) [11], [15], the statistic-based approaches [6], [10], and the synthetic minority oversampling technique (SMOTE) [14].

Based on the literature, two phenomenons can be found: (1) The SVM-based and RF-based algorithms have been proven to provide good results[6], [7], [8], [9], [10], [11], [12], [13], [14], [16]; (2) Tree-based approaches are also popularly used for wine prediction [5], [17]. However, the past literature mostly focused on using or comparing different machine learning models to find the one that can provide the best prediction result for the specific dataset. In other words, when the wine datasets are changed, the obtained model may not provide the same quality of performance. To solve this problem, in this paper, we firstly propose a generalized wine quality prediction framework which consists of the hybrid model acquisition and online prediction phases. Secondly, based on the framework, the generalized wine quality prediction algorithm based on the genetic algorithms is proposed. It first encodes the classifiers as well as their hyperparameters into a chromosome. The fitness of a chromosome is evaluated by the average accuracy of the employed classifiers. The genetic operations are then performed to generate new offspring. The evolution process is continued until reaching the stop criteria. As a result, the proposed approach can automatically find an appropriate hybrid set of classifiers and their hyperparameters for optimizing the prediction result and is independent on the dataset. Experiments were conducted on the wine datasets to show the merits and effectiveness of the proposed approach. The main contributions of the proposed framework and approach are listed as follows:

1. The proposed framework can use all types of classifiers with their hyperparameters as input in the hybrid model acquisition phase to find the suitable hybrid model and its hyperparameters for wine quality prediction.
2. The proposed framework overcomes the problem of data dependency, which means it provides a mechanism that can automatically obtain not only the appropriate hybrid model but also the hyperparameters for the given dataset no matter where the data is collected, from which areas and countries.
3. Based on the proposed framework, the GA-based generalized wine quality prediction algorithm has been proposed, and the obtained hybrid model and hyperparameters are better than existing approaches in terms of accuracy.
4. Experiments also indicate that when using macro F1-score as a fitness function for the proposed approach, the hybrid model can not only reach a better macro-F1 score but also has similar accuracy when comparing to the existing approaches.

The paper is organized as follows. Section II reviews the past works of predicting the wine quality as well as the basic knowledge used in the proposed approach. In Section III, the detailed components used in the proposed approach are described. In Section IV, the generalized wine quality prediction framework and the proposed approach are stated. The obtained results are analyzed and explained in Section V. Finally, conclusions and future work are drawn in Section VI.

II. Literature Review and Background Knowledge

A. Literature Review

Over the years, several studies used the machine learning techniques to predict wine quality, including utilizing SVM, k-NN, decision tree (DT), random forest, neuron network, regression and others. Before describing them, the recent related studies and methods are we summarized and shown in Table I.

TABLE I. Summary of Recent Studies

	SVM	K-NN	Naïve Bayes	Decision Tree	Random Forest	Neural Network	Regression	Others
Cortez et al. [7] (2009)	x					x		x
Ribeiro et al.[5] (2009)				x		x	x	
Appalasamy et al. [17] (2012)			x	x				
Bhattacharjee et al.[10] (2016)	x		x					
Andonie et al. [6] (2016)		x				x		x
Er et al.[11] (2016)	x	x			x			
Hu et al. [14] (2016)				x	x			x
Petropoulos et al.[18] (2017)								x
Zhang et al. [9] (2017)	x					x	x	
Agrawal et al. [19] (2018)						x		
Gupta [8] (2018)	x					x	x	
Trivedi et al. [12] (2018)					x		x	
Sowmya et al. [20] (2019)				x	x			x
Kumar et al. [21] (2020)	x		x		x			
Mahima et al.[16] (2020)		x			x			
Ozalp et al. [22] (2020)					x			x
Shaw et al. [13] (2020)	x				x	x		

From Table I, according to the used techniques, four types of approaches, including studies using SVM, RF, DT and others, are reviewed as follows.

(1) For studies using SVM, for instance, Cortez et al. [7] produced a large dataset for red and white vinho verde wines, a unique product from the Minho region of Portugal with the most common physicochemical tests selected as features. They selected optimal parameters associated with models by sensitivity analysis. The model selection was guided by parsimony search to find the best model. The results indicated that the SVM outperforms multiple regression and NN. In work presented by Gupta [8], it preprocessed the dataset using linear transformation to remove the inconsistent instances. Then, three models were trained using full features and the selected features by regression. Gupta summarized that SVM was the best model for wine quality prediction based on validating error rate. Also, precisions of SVM and NN using selected features were higher than that using all features. Zhang et al. [9] analyzed the Helan mountain wine dataset by using the SVM, logistic regression and NN as prediction models. The result indicated that classification algorithms were feasible for assessing wine quality, and it also showed the SVM performed better compared to other algorithms. Kumar et al. divided the red wine dataset into 70% and 30% for training and testing for evaluating the performances of SVM, RF and Naïve Bayes (NB) [21]. They used accuracy, recall, precision, F1 score and error rate as performance measurements. Based on the results, they suggested that combining and tuning models can provide better performance.

(2) For studies using RF, for instance, Shaw et al. focused on quality prediction performance analysis for the red wine out of three models, including the SVM, RF and NN [13]. They also indicated that the RF outperformed other models. Trivedi et al. firstly normalized the data and removed the outliner from the dataset, and then reduced the classifying labels of a used dataset from 10 classes (1-10) to 2 classes (bad and good). They discovered the accuracies of RF and logistic regression (LR) could achieve 84% and 76% [12]. Hu et al. focused

on handling data imbalance in white wine, by classifying labels to 3 classes that are low quality (3-4), normal (5-7) and high quality (8-9) [14]. They used synthetic minority over-sampling technique (SMOTE) to preprocess imbalance data and apply the processed data into RF, decision tree (DT) and AdaBoost. The experiments showed RF produces the best results in terms of error rates and receiver operating characteristic (ROC) values. Besides, Mahima et al. transformed the labels of the used wine dataset from 10 classes (1-10) to 3 classes, including bad (1-4), average (5-6) and good (7-10), for evaluating the k-NN and RF by the root-mean-square error (RMSE) [16]. They found that employing the most relevant features on both models provided better performance, and observed that the extreme instances could not be classified appropriately. Ozalp et al. applied a fuzzy logic and the random forest to predict the red wine quality using the instances with three labels that are low, medium and high [22]. Sowmya et al. classified the labels into three groups and used both RF and DT for wine quality prediction [20]. The study also used descriptive statistics to explain the assoication between each wine characteristics and wine quality.

(3) For studies using DT, for instance, Ribeiro et al. used the dataset with 326 samples with the chemical characteristic attributes of wine and subjective attributes from wine taster during the production phase for wine quality prediction by the DT, NN and linear regression [5]. The labels were divided into two classes: medium and good. The results showed that the DT and NN could reach exceedingly high accuracies from 86% to 99%. Appalasamy et al. indicated that the DT performances better than NB [17]. Furthermore, it drilled down the results and found that the accuracy of white wine was affected by a higher number of physicochemistry attributes when comparing to the red wine.

(4) For other studies, for instance, Petropoulos et al. used geographical information to predict the quality of wine grow on different sections in the wine region of Nemea, Greece, using fuzzy logic multi-criteria decision-making system [18]. Andonie et al. used data collected from Cabernet Sauvignon in Washington state with 180 samples for wine quality prediction via classifiers in Weka, including the RF, IBk, multilayer perceptron, KStar, etc [6]. The dataset consists of 32 features, and the six labels. Comparing to other works, it not only focused on finding the best model but also aimed to find a trade-off between the number of used features and accuracy. Bhagyalaxmi et al. proposed a framework by gathering the characteristic of red wine and judging the quality of red wine based on client inclinations [10]. Agrawal et al. used multilayer perceptron model with rectified linear unit for building prediction model, and the best accuracy is 53% for both red and white wines [19].

To summarise, most recent wine quality prediction works used the dataset acquired by Cortez et al. [7], but not all works used both red and white wine dataset for the experimental evaluations. Works in [12] and [21] only used red wine dataset for the experiments. In [12], they discovered RF has a better performance. In [21], they revealed that SVM performs better than RF. Works in [13] and [14] only used white wine dataset for experiments. Both works indicated RF provides better performance on white wine dataset. For works [8], [10], [11], [16], [17], they used both red and white wine datasets for experimental analysis. In [8], [10], they obtained the SVM performs better than other models. [37] [18] indicated that RF is the best among models. In [17], the DT was reported as the most suitable model for wine quality prediction.

B. Classifier

This section briefly describes the classifiers used in this work, including the SVM, random forest, and decision tree.

1. The SVM Classifier

The support vector machine (SVM) is a supervised machine learning model for solving a classification problem [23]. The main concept of SVM is utilized the kernel function to find the hyperplane that can separate instances into categories. As mentioned earlier, SVM [8], [9], [10], [11] have proven to be an effective classifier for wine quality prediction.

There are three hyperparameters in SVM that are the penalty factor C, parameter gamma γ and kernel function *kernel*. The C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. A small value of C tends to emphasize the margin while ignoring the outliers in the training data. A large value of C tends to obtain the best fit for the training data that may cause the overfitting problem. The γ defines the influence degree of a single training. With a small value of γ, the model may not be easy to capture the character of the data. With a large value of γ, the influence area of the support vectors is limited to itself. The final one is *kernel*. There are three types of kernels, including the linear, poly and rbf, can be employed to find the best fit model for the given dataset. Hyperparameter tuning relies more on experimental results than theory, and therefore the best method to determine the optimal settings is by trial and error. By auto fine-tuning the hyperparameters, the SVM can achieve better performance.

2. Random Forest

Random forest (RF) is a supervised learning algorithm, and several studies have shown that using RF can provide a good prediction accuracy [15], [24]. In general, the RF algorithm creates different decision trees using randomly sampled instances. Then, in the prediction phase, based on the prediction results of the trees, a voting technique is used to determine the best solution. Due to using multiple decision trees for prediction, the advantage of RF over other methods is that it can reduce the overfitting. The RF has six hyperparameters: (1) "number of estimators" means the number of trees in the forest, (2) "maximum features" refers to the max number of features considered for splitting a node, (3) "maximum depth" is the maximum number of levels in each decision tree, (4) "minimum samples split" indicates the minimum number of instances placed in a node before splitting the node. (5) "minimum samples leaf" is the minimum number of instances allowed in a leaf node, and (6) "bootstrap" represents a method for sampling instances with or without replacement.

3. Decision Tree

The decision tree (DT) belongs to the supervised learning algorithm. The DT is a tree structure in which each internal node represents a feature, and each leaf node represents a label. The branches represent conjunctions of features that lead to those labels, also known as the decision rules. The main concept behind the DT is to find features which contain the most information. Once the feature is found using the selected criteria, the instances will be split by the feature. The process of finding the feature and split instances is continued until reaching the stopping criterion.

There are many hyperparameters that can be tuned for the DT. In this paper, we focused on six of them as following: (1) "Criterion" represents a function used to measure the quality of a split and could be "gini" for the gini impurity and "entropy" for the information gain, (2) "Splitter" is the strategy used to choose the split at each node. Two options are available. The first one is to choose the best split, and another is to random choose the best split, (3) "Minimum samples split" means the minimum number of samples required to split an internal node, (4) "Minimum samples leaf" is the minimum number of samples required for a leaf node, (5) "Maximum features" indicates the number of features is considered when looking for the best split, and (6) "Maximum depth" is the maximum depth of the tree.

C. The Genetic Algorithms

The basic concept of the genetic algorithms (GA) derived from Charles Darwin's theory of natural evolution and can be used in many fields [23]. For instance, Holland applied GA on adaptive and artificial systems [25]. In GA, each solution is encoded in a string called a chromosome, and could be represented in a binary or decimal form. Two main genetic operators that are crossover and mutation are utilized to generate offspring. The crossover and mutation produce offspring as new possible optimal solutions by swapping or mutating genes of the chromosomes. The fitness function in GA is used to evaluate the fitness of chromosomes in the population. The selection process is employed to generate the next population based on the fitness values of chromosomes. The evolutionary process is continued until reaching the stopping criterion, e.g., reaching a predefined number of generations, obtaining a chromosome with the qualified fitness value. In this study, the GA is utilized to search the suitable set of classifiers and the hyperparameters to form the hybrid model for the different dataset automatically. More detailed explanation of the proposed approach will be stated in the next section.

III. Components of GA-based Hybrid Model

This section describes the main components associated with the GA-based hybrid quality prediction algorithm. Those components include chromosome encoding, initialization of population, fitness function, and lastly crossover and mutation operations.

1. Chromosome Encoding

This paper aims to find an appropriate set of classifiers and their hyperparameters as the hybrid model for wine quality prediction to fit different wine dataset. Hence, the chromosome consists of two major parts, including the hyperparameter and model parts. The encoding schema for a chromosome C_i is shown in Fig. 1.

From Fig. 1, in the first part, it represents the hyperparameters for the k classifiers, which means k sections should be used. Thus, the length of the first part is the sum of the number of hyperparameters used for every classifier. The second part decides what algorithms are selected for the hybrid model, and every classifier is represented by a bit. If the value of w^{model_i} is 1, it means w^{model_i} is a part of the hybrid model. In the following, take three models, SVM, RF and DT, as an example. Assume the numbers of hyperparameters of the three models are 3, 6, and 6. Therefore, in this case, the chromosome consists of 18 genes. The first 15 genes in the chromosome are used to represent the hyperparameters of three models. The first 3 genes belong to SVM, the 4th to 9th genes belong to RF, and the last 6 genes belong to DT. The last 3 genes in the second part decide which model(s) should be activated. It can be represented as follows:

Chromosome C_i: $[i^{svm}, i^{rf}, i^{dt}, i^{w}]$, for $i \in P$ is the population.

For i^w, it consists of three genes $[W_s, W_F, W_T]$, where W_s, W_F, W_T are used to represent the voting weight of SVM, RF, and DT. Each model has it is own hyperparameters. i^{svm} represents the set of hyperparameters for SVM that are [*C, γ, kernel*]. i^{rf} indicates the set of hyperparameters for the RF that are $[R_0, R_1, R_2, R_3, R_4, R_5]$. i^{dt} represents the set of hyperparameters for DT that are $[D_0, D_1, D_2, D_3, D_4, D_5]$ as described in previous section. Hence, a possible chromosome is shown in Fig. 2.

In Fig. 2, the values of the model part are 1, 1 and 0 that means the SVM and the RF are used as the hybrid model. In accordance with the hyperparameter part, the first three genes, 500, 0.001 and *rbf*, represent values of the penalty factor, gamma and kernel function used for SVM. The the 4th to 9th genes, 800, 3, 45, 2, 1 and False, represent values of the number of estimators, maximum features, maximum depth, minimum samples split, minimum samples leaf and bootstrap used for RF. The last six genes, *gini*, *best*, 2, 2, *auto* and 8, in the hyperparameter part represent criterion, splitter, minimum samples split, minimum samples leaf, maximum features and the maximum depth, used for DT.

2. Initialization of Population

Population initialization is the first step in the process of the GA. The population is a set of chromosomes, and the initial population $P(0)$ in this case, is randomly generated. In the previous section, we mentioned that each prediction algorithm has its own set of hyperparameters. For example, SVM consists of three hyperparameters [*C, γ, kernel*].

Although the suitable setting for the three parameters are *kernel* = [rbf], C = [9], and γ = [2, 0.5, 0.125] based on three different datasets (astroparticle, bioinformatic and vehicles) [22], it still cannot guarantee they are suitable for all datasets. Therefore, in order to tune the hyperparameters for every algorithm, based on the chromosome encoding scheme and the population size, the initial population can be generated randomly.

3. Fitness Function

To evaluate the population of genes in the chromosome, the GA requires a fitness function to rank the fitness values of chromosomes based on considering the factors. When designing the fitness function, it should be used to measure how close a chromosome is to the target solution. Designing a useful fitness function is essential to reduce the size of the population and to make the GA more likely to find the optimal solution in less time. In the proposed approach, the average value of the accuracies of active models is employed to calculate the fitness values for a chromosome. Thus, the formula of the fitness function to evaluate a chromosome C_i as $f(C_i)$ is defined as follows:

$$f(C_i) = \frac{\sum_{j=1}^{k} w^k * acc\left(M_j\left(i^{m^j}\right)\right)}{\sum_{j=1}^{k} w^k} \quad (1)$$

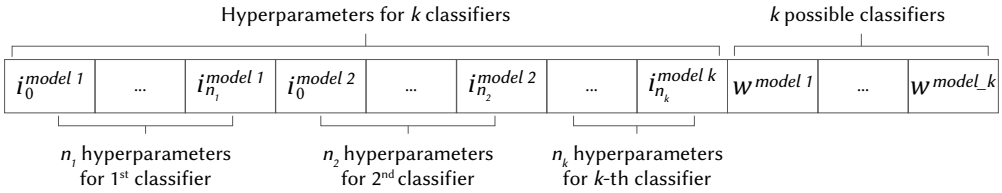


Fig. 1. Encoding schema.

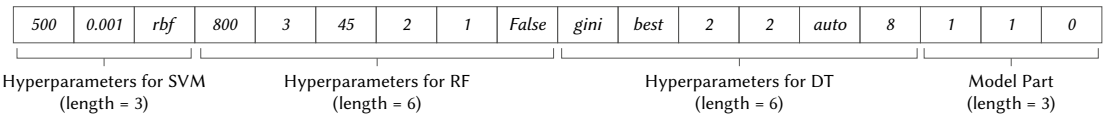


Fig. 2. A possible chromosome.

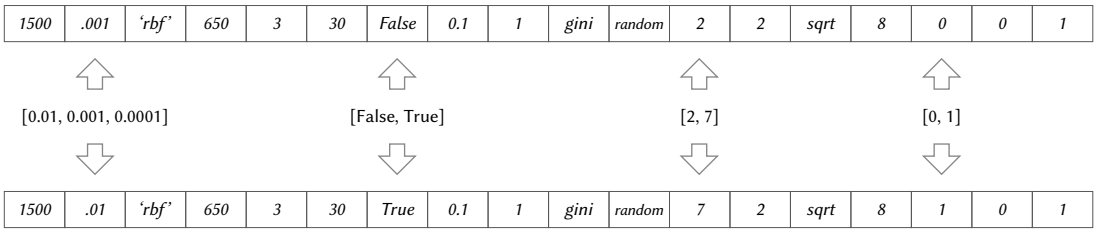


Fig. 3. Illustration of mutation operator.

were M_j is the j-th prediction model adopted in the chromosome C_i. Continue the previous example, when the three prediction models, SVM, RF and DT, are used, the fitness value of a chromosome C_i is calculated as:

$$f(C_i)=\frac{acc(SVM(i^{svm}))w_S+acc(RF(i^{rf}))w_F+acc(DT(i^{dt}))w_T}{w_s+w_F+w_T} \quad (2)$$

where the W_s, W_F, W_T are weights of the three models, and the function acc() is the accuracy function which is utilized to measure the accuracy of a model with an assigned set of parameters. The accuracy is calculated using the formula:

$$acc(M_i)=\frac{Total\ number\ of\ correct\ prediction\ by\ M_i}{Total\ number\ of\ prediction\ by\ M_i} \quad (3)$$

4. Crossover

In this section, the crossover operation performed in this study is described. Based on the crossover operator used in the steady-state GA (SSGA) [26], we made a slight modification and presented the modified crossover operator, steady-state crossover operator (SSCO), for the proposed approach. The difference between the SSCO and that used in SSGA is the way for the selection of parents for crossover. The pseudocode for SSCO is illustrated in Table II.

TABLE II. Pseudocode for the SSCO

Crossover Procedure: SSCO(*P*, *c_rate*, *pSize*)
Input: population *P* Parameters: crossover rate *c_rate* and population size *pSize* Output: newly generated population *P'*
1 Procedure modified_SSGA(){ 2 Create new population *P'* 3 *P'* ← add the "elite" chromosome from *P* 4 Crossover: 5 Select parent C_1 from *P* randomly 6 Select parent C_2 from *P'* randomly 7 Offspring *O* ß uniformCrossover(C_1, C_2, *c_rate*) 8 If *O* is better than the worst chromosome from *P* then 9 Add *O* to *P'* 10 If *P'* < *pSize* 11 GoTo Crossover 12 Else 13 *P* ← *P'* 14 }

From Table II, the SSCO first creates the new population P' (line 2). Then, the elite chromosome is picked from the original population P and copied to the new population P' (line 3). After that, two chromosomes C_1 and C_2 are selected from P and P' (lines 5~6). To make the crossover more effective, the uniform crossover operator is employed for gene exchanging (line 7) [27]. It first generates a number of genes to be exchanged according to the given crossover rate, and the exchanging genes follow the randomly generated positions. At last, the new chromosome O is formed based on the exchanging positions and added to P' (line 8). Take C_1 as base chromosome and C_1as an inserted chromosome as an example. The genes arrangement for C_2 is $[g_{1^*}^2,\ g_2^2,..,g_5^2,\ldots,\ g_8^2,\ldots,g_{k-1}^2,\ g_{k^*}^2]$, where * indicates the genes that will be passed to C_1 to form the new offspring. Hence, the new offspring O is generated as: $[p_{1^*}^2,\ p_2^1,..,p_5^1,\ldots,\ p_8^1,\ldots,p_{k-1}^1,\ p_{k^*}^2]$. The process is continued until *pSize* chromosomes are generated (lines 9~12). In other words, the benefit is that the best chromosome can be utilized as parents to produce the next generation. In addition, the reason to select only the best chromosome is to keep sufficient diversity and avoid premature convergence.

5. Mutation

In biological evolution, due to genes in chromosome may mutate, it provides offspring has the ability to survive when suffering environment changing. Hence, the aim for mutation is to keep the diversify of the population and to prevent the GA trapped in a local optimal [28]. There are several types of mutation. In this study, based on the uniform mutation [29], the modified uniform mutation is employed to mutate randomly selected gene(s) with a mutation probability p_m. In original uniform mutation, the operator mutates the value of the randomly selected gene with the uniform random value between a specified upper and lower bound. Instead of selecting a value between a specific range, the proposed approach only allows the mutation operator to select a value from the given specified list. Continue the previous example, let the second gene of the SVM, the fourth gene of the RF and third gene of the DT are mutated in the hyperparameter part, and let the first gene of the model part are mutated. Assume the specified lists of those genes are [0.01, 0.001, 0.0001], [True, False], [2, 7], and [0, 1], the mutation operator is illustrated in Fig. 3.

From Fig. 3, the values, including 0.01, 'True', 7, and 1, are selected to replace original genes. Then, the chromosome C_i' is generated. After mutation operation, the offspring C_i' will replace the C_i in the population.

IV. Generalized Wine Quality Prediction Framework and Proposed Approach

In this section, the generalized wine quality prediction framework is presented in Section VI.A. Then, based on the framework, the proposed algorithm is described for obtaining the appropriate hybrid model using the GA for wine quality prediction in Section VI.B.

A. Generalized Wine Quality Prediction Framework

As mentioned in the previous section, the existing approaches focus on how to obtain a classifier that can have the best prediction ability on a specific dataset. As to the hyperparameters of the classifier, they can be discovered by different strategies, e.g., the grid search [30], or random search [31]. However, the wine datasets may be collected from different areas and countries, and hyperparameter discovery process may time consuming based on the given searching space. In this paper, we thus propose the generalized wine quality prediction framework for providing a mechanism that can automatically find not only the appropriate hybrid model but also the hyperparameters by the evolution-based algorithms. The framework is shown in Fig. 4.

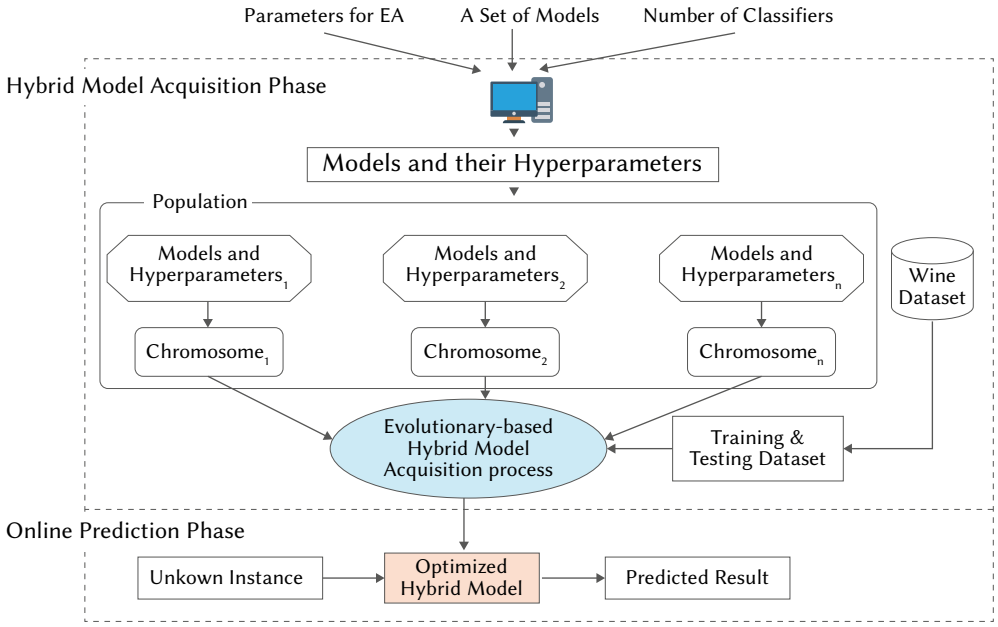


Fig. 4. The generalized wine quality prediction framework.

From Fig. 4, the proposed framework contains two phases, including the hybrid model acquisition and online prediction phases. In the first phase, according to the types of classifiers and their hyperparameters, the population is initialized. The initialized population is then sent into the evolutionary-based hybrid model acquisition module. Operations will generate possible offspring. After generations, the hybrid model with the best fitness value is outputted. Note that any evolutionary-based algorithms can be employed for searching the hybrid model as well as their hyperparameters. In the online prediction phase, the unknown instance can be identified by the optimized hybrid model.

B. GA-based Generalized Wine Quality Prediction Algorithm

Based on the proposed framework, the GA-based generalized wine quality prediction algorithm is stated in this section. The pseudocode of the proposed algorithm is shown in Table III.

In Table III, the proposed approach first divides the dataset into training and testing datasets (line 2). Then, the initial population is generated based on the given set of models M, the number of classifiers num_c, and the hyperparameters HP_M (lines 4~7). The fitness function defined in formula (1) is utilized to evaluate the quality of every chromosome (lines 9~12). Each chromosome represents a model M. Using the given training and testing datasets D_{train} and D_{test}, the fitness value $fValue$ of a chromosome is calculated (line 10). During this step, the encoded model is trained and tested, and a performance score for each model is returned as the fitness value in the end. Then the fitness value of a chromosome is updated the population (line 10). Note that other criteria can also be used as the fitness function, e.g., macro-F1 score. The two genetic operators are performed on the population to generate new offspring (lines 13~14). The newly generated population will replace the previous population (line 15). After reaching the predefined number of generations num_gene, the best chromosome is outputted as the final hybrid model (line 17). The best chromosome consists of hyperparameters for the hybrid model. Because many works reported that the classification techniques, including the SVM, RF and DT are commonly used for wine quality prediction. Therefore, in this paper, three models but not limited to are used as the set of models to construct the hybrid model. Although there are many existing wine quality prediction approaches, they focus only on certain dataset or classifiers. Thus, the proposed algorithm's advantage is that it is a general algorithm and can be employed to find the appropriate hybrid model and its hyperparameters no matter what kinds of wine datasets are given.

TABLE III. Pseudocode of the GA-based Generalized Wine Quality Prediction Algorithm

The proposed algorithm: GA_hybrid()
Input: dataset D, a set of models M, hyperparameter for models HP_M, number of classifiers num_c Parameters: population size $pSize$, number of generations num_gene, crossover rate c_rate, mutation rate m_rate Output: The hybrid model and hyperparameters
1 Procedure GA_hybrid(){ 2 D_{train}, D_{test} ← trainingandTestingGeneration(D) 3 $population \leftarrow \phi$ 4 For i = 0 to $pSize$ 5 C_i ← generateModelandHyperpara(M, num_c, HP_M) 6 $population \leftarrow population \cup C_i$ 7 End For i 8 For j = 0 to num_gene 9 For i = 0 to $pSize$ 10 $fValue_i$ ← calculateFitness(C_i, M, D_{train}, D_{test}) 11 $Population$ ← updateFitness($population$, $fValuei$) 12 End For i 13 $newPopulation$ ← SSCO($population$, c_rate, $pSize$) 14 $newPopulation$ ← executeMutation($newPopulation$, m_rate) 15 $population \leftarrow newPopulation$ 16 End For j 17 $bestChromosome$ = findBestSolution($population$) 18 Output $bestChromosome$ 19 } 20

V. Experimental Results

A. Dataset Descriptions and Baseline Models

The wine dataset from the UCI database consists of two sets of wine datasets that are red and white wine datasets [7]. The red wine and white datasets contain 1599 and 4898 instances, respectively. Both datasets contain 11 physiochemical variables, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, Sulphates, and alcohol. The attribute

"sensory" is a quality rating (class label) which is from 0 (very bad) to 10 (excellent).

The datasets were collected from May 2004 up to February 2007 using the only protected designation of origin samples that were tested at the official certification entity (CVRVV). The CVRVV is an inter-professional organization to improve the quality and marketing of Vinho Verde. The datasets were recorded by a computerized system (iLab), which automatically manages the wine sample testing process from producer requests to laboratory and sensory analysis [1]. The statistical details of the datasets for the physiochemical variables are shown in Table IV.

TABLE IV. The Statistics Summary for the Physiochemical Features of the Datasets

	Red Wine			White Wine		
Attribute	**Min.**	**Max.**	**Mean**	**Min.**	**Max.**	**Mean**
Fixed acidity	4.6	15.9	8.3	3.8	14.2	6.9
Volatile acidity	0.1	1.6	0.5	0.1	1.1	0.3
Citric acid	0	1	0.3	0	1.7	0.3
Residual sugar	0.9	15.5	2.5	0.6	65.8	6.4
Chlorides	0.01	0.61	0.08	0.01	0.35	0.05
Free sulfur dioxide	1	72	14	2	289	35
Total sulfur dioxide	6	289	46	9	440	138
Density	0.99	1.01	0.996	0.99	1.04	0.994
Sulphates	0.3	2	0.7	0.2	1.1	0.5
pH	2.7	4	3.3	2.7	3.8	3.1
Alcohol	8.4	14.9	10.4	8	14.2	10.4

To show the merits of the proposed approach, we compare the hybrid model against the SVM, RF and DT with the hyperparameters that were discovered using the grid search and random search. The *n*-fold cross-validation is utilized to construct models. The setting for each model is displayed in Table V.

TABLE V. Hyperparameter Setting for the Baseline Models

Model	**Hyperparameters Setting**
SVM	$C = 1500, \gamma = 0.0001,$ $kernel = rbf$
RF	$R_o = 1400, R_1 = 3, R_2 = 20,$ $R_3 = 4, R_4 = 3, R_5 = True$
DT	$D_o = gini, D_1 = 6, D_2 = 5,$ $D_3 = auto, D_4 = 3, D_5 = random$

B. Experimental Setting

In this section, we explain the experimental setting of the proposed algorithm. There is no specific rule to set proper hyperparameters for each model. It is a tedious but crucial task, as the performance of a classifier is highly dependent on the choice of hyperparameters. In order to find the appropriate initial parameter setting for each model, we executed a grid search and random search to find the possible parameters for each classifier. The ranges of hyperparameters for SVM i^{svm}, RF i^{rf}, DT i^{dt} and weight option i^{w} are shown in Table VI.

Based on the parameters listed in Table VI, the number of chromosomes that can be created is 132,7104. This number is too large and unable to complete the evolution process in a reasonable time. Therefore, the population size of the proposed algorithm was set at 500. Hence, it randomly selected 500 chromosomes to form the initial population. The number of generations was set at 100. The crossover and mutation rates were set at 50% and 1%.

TABLE VI. The Ranges of Hyperparameters for the Used Models

Model	**Parameters**	**Ranges**	**Data types**
i^{svm}	C	[500, 1000, 1500]	Integer
	γ	[0.01,0.001, 0.0001]	Float
	kernel	['rbf', 'poly']	String
i^{rf}	R_o	[650, 733, 800]	Integer
	R_1	[3, 'sqrt']	String
	R_2	[30, 45, None]	Integer
	R_3	[0.1, 2]	Float
	R_4	[1, 3]	Integer
	R_5	[False, True]	Boolean
i^{dt}	D_o	['entropy', 'gini']	String
	D_1	['best', 'random']	String
	D_2	[2, 7]	Integer
	D_3	[1, 2]	Integer
	D_4	['sqrt', 'auto']	String
	D_5	[9, 8]	Integer
i^{w}	w_s	[0, 1]	Integer
	w_F	[0, 1]	Integer
	w_T	[0, 1]	Integer

In the following, the performance measurements of a classifier are described. The accuracy of a classifier is one way to measure how often the algorithm correctly classifies an instance. The formula is shown as follows:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

where *TP* is the true positive, which means the number of positive instances that are classified to the positive class. *TN* is the true negative, which means the number of negative instances that are classified to the negative class correctly. *FP* is false positive, which means the number of negative instances that are classified to the positive class. *FN* is false negative, which means the number of positive instances that are classified to the negative class.

In the multi-class classification problem, micro and average accuracy, precision, and recall are always the same [32]. Therefore, we use the macro-averaging measurements that are macro-precision and macro-recall for additional measurement reference. Also, based on past work [11], it indicated the wine dataset is imbalanced, only using accuracy may not provide a clear picture. Thus, the macro-F1 score is also utilized for a more detailed comparison. The definition of precision to evaluate a multi-class classifier is shown as follows:

$$P_{Classifier} = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FP_c} \tag{5}$$

where TP_c and FP_c represent the true positive and the false positive for class *c*. When precision is one, it means the prediction ability of the classifier is perfect. The macro-precision will be lower than average precision. That is because although the model performs exceptionally well on some specific classes, it may perform poorly on some classes, hence downgrading the value of the macro-precision score. The macro-precision is given as follows:

$$Macro\text{-}Precision\ MP_{Classifier} = \frac{\sum_c P_c}{number\ of\ classes} \tag{6}$$

The macro- precision is performed by first computing the precision of every class, and then taking the average of all precisions.

Another metric often used to evaluate performance other than

accuracy is the recall. There is a trade-off between precision and recall. It means higher the recall lower the precision and vice versa. The recall measures the percentage of total relevant results correctly classified by the algorithm. This value is an important indication of how many predictions are correctly predicted. The definition of recall to evaluate a multi-class classifier is shown as follows:

$$recall\ R_{classifier} = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FN_c} \tag{7}$$

where TP_c and FN_c represent the true positive and false negative for class c. When the recall is one, it means that all truly positive samples were predicted as the positive class. Similar to micro-precision, the value will be lower if one class performs poorly. The macro-recall is given as follows:

$$Macro\text{-}Recall\ MR_{classifer} = \frac{\sum_c R_c}{number\ of\ classes} \tag{8}$$

Accuracy is useful when the class distribution in the dataset is even, but F1 score is a better metric when the dataset has imbalanced classes. F1 score is simply a harmonic mean between precision and recall. The definition of F1 score to evaluate a multi-higher class classifier is shown as follows:

$$F1_{classifier} = \frac{2 \times (P_c \times R_c)}{(P_c + R_c)} \tag{9}$$

where P_c and R_c represent the precision and recall for class c. Maximizing the F1 score is like finding the best balancing value between precision and recall. Since we are processing multi-class dataset, we would prefer to use macro-F1 score for comparison. The macro-F1 score calculation is given as follows:

$$Macroi\text{-}F1 \quad F1_{classifer} = \frac{\sum_c F1_c}{number\ of\ classes} \tag{10}$$

There is no defined range of F1 score to determine the performance of the model. We can maximize the macro-F1 score to find the best-balanced value between precision and recall.

C. Experimental Results

Since most of the past works mainly focus on accuracy, we thus compare the accuracy of the proposed approach against others. Also, most works set the training and testing datasets ratio to 80% and 20%. Therefore, we also set the same ratio for the training process. For comparisons, we include three mentioned classification models, the SVM, RF, and DT, as the baseline models. We also compare the proposed approach to the works of Cortez et al. [7] and Appalasamy et al. [17] for performance evaluations. However, both works did not provide enough information to calculate precision and recall. Therefore, the comparison results of baseline and proposed approach in terms of accuracy, precision and recall on the testing datasets are shown in Table VII.

TABLE VII. Comparison of Proposed Approach with Different Models

	Red Wine			White Wine		
Models	Accuracy	Precision	Recall	Accuracy	Precision	Recall
SVM	0.57	0.39	0.37	0.51	0.41	0.35
RF	0.58	0.51	0.36	0.66	0.43	0.41
DT	0.53	0.37	0.34	0.58	0.32	0.32
Cortez et. al [7]	0.45	-	-	0.51	-	-
Apalasamy et. al [17]	0.62	-	-	0.65	-	-
Proposed Approach	0.72	0.34	0.36	0.68	0.57	0.37

From Table VII, we can observe that the accuracies of the proposed approach on red wine and white wine are 72% and 68% that is better than existing approaches. It is also interesting to see that the proposed model has lower precision and recall for red wine than most of the baseline models. For white wine, accuracy and precision are higher than all models, but the recall is slightly lower than the RF. These results indicated that using accuracy as fitness function for finding the hybrid models are good on white wine but a little worse on red wine dataset. To further examine the performances of the proposed approach, different training and testing datasets are used to obtain the hybrid models for red wine and white wine datasets. The results of them are shown in VIII and Table IX.

TABLE VIII. Result for Different Testing Data Ratio (Red Wine)

Red Wine	Testing Dataset Size Percentage			
	10%	*20%*	*30%*	*40%*
Accuracy	0.73	0.72	0.69	0.67
Macro-Precision	0.42	0.34	0.34	0.34
Macro-Recall	0.47	0.36	0.36	0.33
Macro-F1 score	0.45	0.35	0.35	0.33

TABLE IX. Result for Different Testing Data Ratio (White Wine)

White Wine	Testing Dataset Size Percentage			
	10%	*20%*	*30%*	*40%*
Accuracy	0.70	0.68	0.67	0.66
Macro-Precision	0.64	0.57	0.49	0.52
Macro-Recall	0.62	0.37	0.37	0.36
Macro-F1 score	0.49	0.40	0.40	0.40

From Table VIII, when the testing ratios were set at 10% or 20%, the hybrid models provide the highest accuracies than others, and the accuracies were gradually decreased along with the increasing of ratios. The macro-precision and macro-recall are low and almost similar for red wine dataset. That means the amount of false-positive is very close or equal to the false negative. Besides, the macro-F1 score also dropped when the ratio larger and equal to 20%.

From Table IX, the hybrid model on white wine dataset shows a different result, where the macro-precision is always higher than macro-recall. When the ratio was set at 10%, the accuracy and macro-F1 score are at the highest, and the macro-precision and macro-recall are at the closest. The macro-F1 score is comparative constant with the change of ratio from 20% to 40%. When the macro-F1 score is low, the macro-precision and macro-recall score for red wine indicates the data is highly skewed on certain classes. The white wine dataset is also skewed, but the distribution is more even when comparing to red wine. It is interesting to note that when the testing and training ratio is 10% to 20%, the proposed approach can reach the best performance.

Since the work has proven the datasets for both red and white wine are imbalance [14], it is more reasonable to focus on macro-F1 score instead of accuracy. Macro-F1 score is instrumental in most scenarios when working with imbalanced datasets. Under this condition, we change the fitness function to focus on finding the hyper model that can provide the highest F1 score. The results using the proposed approach with the F1 score as a fitness function for red and white wine datasets are shown in Table X and Table XI.

The result shows that if we focused on improving the macor-F1 score, the accuracies would drop under all conditions. It is also interesting to see both red wine and white wine datasets behave similarly. That is when the increase in the ratios, the macro-F1 score and accuracy are decreasing. The results also show that when the ratio is 10%, the obtained hybrid model has the best performances of 0.59 and 0.58 on red and white wine datasets. Overall speaking, the

proposed approach using macro-F1 is better than that using accuracy as the fitness function.

TABLE X. Results for Using F1 Score As Fitness Score (Red Wine)

Red wine	Testing dataset size percentage			
	10%	*20%*	*30%*	*40%*
Macro-F1	0.59	0.46	0.42	0.36
Accuracy	0.65	0.61	0.62	0.57
Macro-Precision	0.57	0.44	0.41	0.36
Macro-Recall	0.63	0.48	0.45	0.35

TABLE XI . Results for F1 Score As Fitness Score (White Wine)

White wine	Testing dataset size percentage			
	10%	*20%*	*30%*	*40%*
Macro-F1	0.58	0.41	0.40	0.38
Accuracy	0.65	0.68	0.67	0.65
Macro-Precision	0.66	0.51	0.50	0.52
Macro-Recall	0.55	0.37	0.37	0.35

D. Results of Wilcoxon Signed-Rank and Friedman Tests

We used the Wilcoxon signed-rank test to verify whether the proposed approach is statistically significance at a confidence level at 95%. Since we were unable to retain further experimental data from Cortez et al. [7] and APalasamy et al. [17], we compared the proposed model (***P***) with the baseline models (SVM, DT and RF). With the accuracies of each model A_{SVM}, A_{DT}, A_{RF} and A_P, the Wilcoxon signed-rank test results on red and white wines against the proposed model are summarized in Table XII and Table XIII.

TABLE XII. The Results of the Wilcoxon Signed-rank Test (Red Wine)

	SVM-*P*	DT-*P*	RF- *P*
H0: Null hypothesis	H_0: $A_{SVM} = A_P$	H_0: $A_{DT} = A_P$	H_0: $A_{RF} = A_P$
Ha: Alternative hypothesis	Ha: $A_{SVM} < A_P$	Ha: $A_{DT} < A_P$	Ha: $A_{RF} < AP$
z-value (two tail)	-2.0896	-1.9604	-1.6803
p-value (two tail)	0.0362	N/A	N/A
T_{wilcox}	$T_{wilcox}(10) = 6.5$	$T_{wilcox}(8) = 4$	$T_{wilcox}(8) = 6$

According to the Wilcoxon test for red wine, the p-value for the SVM-*P* pair is smaller than the threshold value of 0.05. Therefore, the null hypothesis is rejected. In addition, for SVM-*P*, the $T_{wilcox}(10)$ is 6.5 which is smaller than the critical value for Wilcoxon at $N = 10$ ($p < .05$) is 8. Since both p-value and T_{wilcox} all below the threshold, the null hypothesis can be rejected. That indicates the proposed approach is significantly better than the SVM. However, DT-*P* and RF-*P* pairs show different results, because the the Wilconx test data size *N* is 8, which is not large enough for the distribution of the Wilcoxon statistic to form a normal distribution. Therefore, it is not possible to calculate accurate p-value.

TABLE XIII. The Results of the Wilcoxon Singed-rank Test (White Wine)

	SVM-*P*	DT-*P*	RF- *P*
H_0: Null hypothesis	H_0: $A_{SVM} = A_P$	H_0: $A_{DT} = A_P$	H_0: $A_{RF} = A_P$
H_a: Alternative hypothesis	H_a: $A_{SVM} < A_P$	H_a: $A_{DT} < A_P$	H_a: $A_{RF} < A_P$
z-value (two tail)	-2.8031	-2.8031	-2.8031
p-value (two tail)	0.00512	0.00512	0.00512
T_{wilcox}	$T_{wilcox}(10) = 0$	$T_{wilcox}(10) = 0$	$T_{wilcox}(10) = 0$

For white wine, it is interesting to see all null hypothesis can be rejected since $T_{wilcox}(10) = 0$ for all three sets of experiments. In short, given the accuracy for each method on the same dataset, the proposed model performed better than all other models for white wine dataset. However, for red wine, the proposed model performed better than SVM, but cannot make a conclusion for DT and RF for red wine dataset. We can only conclude that the data size for red wine used for each test is small and does not provide enough information to make an effective conclusion.

A Friedman test was then conducted on ten runs for red and white wines to examine performances (accuracies) of the four different models on 10 datasets. The Friedman test is a non-parametric equivalent of the repeated measures ANOVA [33]. Results showed that different red wine models produce statistically significant differences in terms of accuracy with $Q = 149.64$ and $p < 0.000001$. For white wine, it also showed that different models also perform statistically significant difference in terms of accuracy with $Q = 168.48$ and $p < 0.000001$.

E. Discussion

There are some works in the recent two years that also conduct experiments on the same wine dataset. However, those studies divided the instances into two labels [12] or three labels, such as [34], [16], [22] and [20] for building models. Therefore, it makes the comparison slightly unfair due to the different standard. Other studies like [19], [15], [13] and [21] either have lower results than Appalasamy's work [17]. In this paper, we thus compared the proposed method with the models proposed by Cortez et al. [7] and Appalasamy et al. [17] because they provided detailed description of each evaluation measurement matrices.

Evolutionary Algorithms (EA) refers to a set of biologically-inspired algorithms, for example, the Genetic Algorithms (GA), the Particle Swarm Optimization (PSO) [35], etc. GA is a stochastic search method that mimics the metaphor of natural biological evolution. For the PSO, it is inspired by the social behaviours of animals, and by updating the position and velocity of each individual to find solutions. Recently, PSO gained some attention in the field of the next-generation wireless network [36].

The differences between the GA and PSO are stated as follows. Based on [37], the PSO performs better in terms of the computational efficiency than the GA for solving the unconstrained non-linear problems with continuous design variables. However, the GA performs better when applied to the constrained non-linear problems with continuous or discrete design variables. For the problem to be solved in this paper, variables are constrained, non-linear and discrete. Therefore, the GA is adopted to deal with the hybrid model optimization algorithm. However, if the problem can be mapped to the unconstrained non-linear problems, the PSO will be a good methodology to be employed for searching the solution. In the future, we will continue to enhance the framework and to try and design different approaches to tune the performances.

VI. Conclusion and Future Work

In this paper, unlike most past works focusing on which classification model provides the best performance in predicting wine quality. Instead, we have proposed a generalized wine quality framework which consists of the hybrid model acquisition and online prediction phases. Based on the framework, the GA-based generalized wine quality prediction algorithm has been proposed. The proposed approach first encodes a set of classifiers and hyperparameters into a chromosome. The fitness functions including the accuracy and macro-F1 score are employed to evaluate the goodness of every chromosome. The steady-state crossover operator and uniform operator are applied on the population to generate new offspring.

After the evolution process, the appropriate hybrid model and the hyperparameters are used for wine quality prediction. Experiments on the red and white wine datasets indicate that the proposed approach is better than other existing approaches in terms of accuracy. In addition, when using macro-F1 score as the fitness function, although the accuracy of the hybrid model is decreasing, the macro-F1 score, macro-precision and macro-recall are increasing. In the future, under the proposed framework, other types of evolutionary algorithms can be employed to get a more solid classifier. In addition, more classifiers or other ML approach like different neural networks [38] can also be considered to construct the hybrid model.

Acknowledgments

This research was supported by the Ministry of Science and Technology of the Republic of China under grants MOST 108-2221-E-032-037 and 109-2622-E-027-032.

References

[1] B. C. Smith, "Getting more out of wine: Wine experts, wine apps and sensory science," *Current Opinion in Food Science,* vol. 27, pp. 123-129, 2019.

[2] J. M. Cardebat and F. Livat, "Wine expert's rating: A matter of taste?," *International Journal of Wine Business Research,* Vol. 28, pp. 43-58, 2016.

[3] B. V. Canizo, L. B. Escudero, R. G. Pellerano, and R. G. Wuilloud, "10 - quality monitoring and authenticity assessment of wines: Analytical and chemometric methods," in *Quality control in the beverage industry*, A. M. Grumezescu and A. M. Holban, Eds.: Academic Press, pp. 335-384, 2019.

[4] M. Yeo, T. Fletcher, and J. Shawe-Taylor, "Machine learning in fine wine price prediction," *Journal of Wine Economics,* Vol. 10, No. 2, pp. 151-172, 2015.

[5] J. Ribeiro, J. Neves, J. Sanchez, M. Delgado, J. Machado and P. Novais, "Wine vinification prediction using data mining tools," *International conference on European computing conference*, pp. 78-85, 2009.

[6] R. Andonie, A. M. Johansen, A. L. Mumma, H. C. Pinkart, and S. Vajda, "Cost efficient prediction of cabernet sauvignon wine quality," *IEEE Symposium Series on Computational Intelligence*, pp. 1-8, 2016

[7] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems,* Vol. 47, no. 4, pp. 547-553, 2009.

[8] Y. Gupta, "Selection of important features and predicting wine quality using machine learning techniques," *Procedia Computer Science,* Vol. 125, pp. 305-312, 2018.

[9] Z. Lingfeng, F. Feng, and H. Heng, "Wine quality identification based on data mining research," *International Conference on Computer Science and Education* , pp. 358-3612, 2017.

[10] S. Bhattacharjee and M. R. Chaudhuri, "Understanding quality of wine products using support vector machine in data mining.," *Prestige International Journal of Management & IT- Sanchayan,* Vol. 5, no. 1, pp. 67-80, 2016.

[11] Y. Er and A. Atasoy, "The classification of white wine and red wine according to their physicochemical qualities," *International Journal of Intelligent Systems and Applications in Engineering,* pp. 23-23, 2016.

[12] A. Trivedi and R. Sehrawat, "Wine quality detection through machine learning algorithms," *International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering*, pp. 1756-1760, 2018.

[13] B. Shaw, A. K. Suman, and B. Chakraborty, "Wine quality analysis using machine learning," *Emerging technology in modelling and graphics*: *Springer*, pp. 239-247, 2020.

[14] G. Hu, T. Xi, R. Mohammed, and H. Miao, "Classification of wine quality with imbalanced data," *IEEE International Conference on Industrial Technology*, pp. 1712-1217, 2016.

[15] S. Aich, A. A. Al-Absi, K. L. Hui, J. T. Lee, and M. Sain, "A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques," *International Conference on Advanced Communication Technology*, pp. 139-143, 2018.

[16] G. U. Mahima, Patidar Y., Agarwal A., Singh K.P., "Wine quality analysis using machine learning algorithms," *Micro-Electronics and Telecommunication Engineering: Proceedings of 3rd ICMETE 2019,* pp. 11-18, 2020.

[17] P. Appalasamy, A. Mustapha, N. Rizal, F. Johari, and A. Mansor, "Classification-based data mining approach for quality control in wine production," *Journal of Applied Sciences,* Vol. 12, pp. 598-601, 2012.

[18] S. Petropoulos, C. S. Karavas, A. T. Balafoutis, I. Paraskevopoulos, S. Kallithraka, and Y. Kotseridis, "Fuzzy logic tool for wine quality classification," *Computers and Electronics in Agriculture,* Vol. 142, pp. 552-562, 2017.

[19] G. Agrawal and D.-K. Kang, "Wine quality classification with multilayer perceptron," (in En), *International Journal of Internet, Broadcasting and Communication,* Vol. 10, no. 2, pp. 25-30, 2018.

[20] D. Sowmya, Sayyed Johar, M. Ganavi, and N. S. Nayak, "Analyzing wine types and quality using machine learning techniques," *International Journal of Engineering Applied Sciences and Technology,* Vol. 4, no. 5, pp. 519-529, 2019.

[21] S. Kumar, K. Agrawal, and N. Mandan, "Red wine quality prediction using machine learning techniques," *International Conference on Computer Communication and Informatics*, pp. 1-6, 2020.

[22] A. E. Ozalp and I. Askerzade, "A data science study for determining food quality: An application to wine," *Communications*, Vol. 68, No. 1, pp. 762-770, 2018.

[23] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Transactions On Neural Networks*, Vol. 10, No. 5, pp. 988-999, 1999.

[24] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news,* Vol. 2, pp. 18-22, 2002.

[25] J. H. Holland, "Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence". *MIT press*, 1992.

[26] A. Rogers and A. Prügel-Bennett, "Modelling the dynamics of a steady-state genetic algorithm," *Foundations of genetic algorithms,* Vol. 5, pp. 57-68, 1999.

[27] W. M. Spears and K. D. De Jong, "On the virtues of parameterized uniform crossover," *Naval Research Lab Washington DC*, 1995.

[28] I. Korejo, S. Yang, K. Brohi, and Z. Khuhro, "Multi-population methods with adaptive mutation for multi-modal optimization problems," *International Journal on Soft Computing, Artificial Intelligence and Application,* Vol. 2, pp. 1-19, 2013.

[29] N. Soni and T. Kumar, "Study of various mutation operators in genetic algorithms," *International Journal of Computer Science and Information Technologies,* Vol. 5, no. 3, pp. 4519-4521, 2014.

[30] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," *National Taiwan University*, 2003.

[31] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research,* Vol. 13, pp. 281–305, 2012.

[32] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information processing & management,* Vol. 45, No. 4, pp. 427-437, 2009.

[33] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research,* Vol. 7, pp. 1-30, 2006.

[34] D. Bhagyalaxmi, R. Manjula, and V. Ramanababu, "A new framework approach to predict the wine dataset using cluster algorithm," *International Conference on Computational Intelligence and Informatics*, pp. 39-48, 2020.

[35] H. Al-Sahaf, Y. Bi, Q. Chen and A. Lensen, "A survey on evolutionary machine learning," *Journal of the Royal Society of New Zealand,* Vol. 49, No. 2, pp. 205-228, 2019.

[36] J. C. W. Lin, G. Srivastava, Y. Zhang, Y. Djenouri, and M. Aloqaily, "Privacy preserving multi-objective sanitization model in 6g iot environments," *IEEE Internet of Things Journal,* pp. 1-1, 2020.

[37] R. Hassan, B. Cohanim, O. De Weck, and G. Venter, "A comparison of particle swarm optimization and the genetic algorithm," *AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference*, AIAA 2005- 1897, 2005.

[38] J. C.-W. Lin, Y. Shao, Y. Djenouri, and U. Yun, " ASRNN: A recurrent neural network with an attention model for sequence labeling," *Knowledge-Based Systems,* Vol212, 106548, 2020.

Terry Hui-Ye Chiu

Terry Hui-Ye Chiu received her B.S. in Information Technology and M.S. in Computer Science from the University of Auckland, New Zealand, in 2001. She is pursuing a Ph.D degree in the Department of Information and Finance Management at National Taipei University of Technology, Taipei, Taiwan. Her research interests include machine learning, data mining, financial technology application.

Chienwen Wu received

Chienwen Wu received his PhD degree in Computer Science from the University of Illinois at Urbana-Champaign in Illinois. Currently, Chienwen Wu is an Associate Professor of Information and Finance Management at National Taipei University of Technology. His research areas include warehouse management, combinatorial optimization, meta-heuristics and machine learning.

Chun-Hao Chen

Chun-Hao Chen is an associate professor at Department of Information and Finance Management at National Taipei University of Technology, Taipei, Taiwan. Dr. Chen received his Ph.D. degree with major in computer science and information engineering from National Cheng Kung University, Taiwan, in 2008. He has a wide variety of research interests covering data mining, time series, machine learning, evolutionary algorithms, and fuzzy theory. Research topics cover portfolio selection, trading strategy, business data analysis, time series pattern discovery, etc. He serves as the associate editor of the International Journal of Data Science and Pattern Recognition, and IEEE Access.

Optimal QoE Scheduling in MPEG-DASH Video Streaming

Shin-Hung Chang[1]*, Min-Lun Tsai[1], Meng-Huang Lee[2], Jan-Ming Ho[3]

[1] Department of Computer Science and Information Engineering, Fu Jen Catholic University, Taipei (Taiwan)
[2] Department of Information Technology and Management, Shih Chien University, Taipei (Taiwan)
[3] Institute of Information Science, Academia Sinica, Taipei (Taiwan)

Received 30 October 2020 | Accepted 13 January 2021 | Published 25 June 2021

unir
LA UNIVERSIDAD EN INTERNET

Abstract

DASH is a popular technology for video streaming over the Internet. However, the quality of experience (QoE), a measure of humans' perceived satisfaction of the quality of these streamed videos, is their subjective opinion, which is difficult to evaluate. Previous studies only considered network-based indices and focused on them to provide smooth video playback instead of improving the true QoE experienced by humans. In this study, we designed a series of click density experiments to verify whether different resolutions could affect the QoE for different video scenes. We observed that, in a single video segment, different scenes with the same resolution could affect the viewer's QoE differently. It is true that the user's satisfaction as a result of watching high-resolution video segments is always greater than that when watching low-resolution video segments of the same scenes. However, the most important observation is that low-resolution video segments yield higher viewing QoE gain in slow motion scenes than in fast motion scenes. Thus, the inclusion of more high-resolution segments in the fast motion scenes and more low-resolution segments in the slow motion scenes would be expected to maximize the user's viewing QoE. In this study, to evaluate the user's true experience, we convert the viewing QoE into a satisfaction quality score, termed the Q-score, for scenes with different resolutions in each video segment. Additionally, we developed an optimal segment assignment (OSA) algorithm for Q-score optimization in environments characterized by a constrained network bandwidth. Our experimental results show that application of the OSA algorithm to the playback schedule significantly improved users' viewing satisfaction.

Keywords

Integer Programming, MPEG-DASH, Motion Vector, Quality Of Experience (QoE), Video Streaming.

DOI: 10.9781/ijimai.2021.06.003

I. Introduction

Recently, the rapid increase in network bandwidth has enabled Internet services to offer users more diverse choices with the result that Internet services have become inseparable from daily life. Owing to the changes in the habits and customs of users seeking entertainment, video streaming services have become part of life. For example, people can instantly receive the latest news, watch movies, or listen to music online. Although the increase in network bandwidth has brought additional Internet services, network congestion has become more problematic, especially for video streaming services. HTTP is a popular protocol for sending and receiving web pages. The reasons for adopting HTTP include its ability to easily penetrate client firewalls because network port 80 is always open for browsing webpages. Moreover, HTTP is compatible with content delivery networks (CDNs). Thus, it is helpful for deploying CDNs [1]. The well-known video streaming technology based on HTTP is dynamic adaptive streaming over HTTP (DASH), also known as MPEG-DASH [2], [3], [4]. DASH technology divides entire streaming videos into a series of small video segments and sequentially transmits each video segment to the users. The DASH client selects different video segments according to the network bandwidth conditions, and the length of each video segment is fixed.

With the advent of the 5G era, network bandwidth has increased tremendously, and video streaming providers such as YouTube, Netflix, Hulu, and Tudo have not only provided users with faster and more convenient services, but also improved quality of service (QoS) [5]. Although DASH streaming can provide different video resolutions to users according to the network bandwidth, it cannot ensure user viewing satisfaction. Therefore, video streaming service providers can improve user satisfaction by reducing the influence of network congestion. Although a certain QoS can be maintained, user satisfaction is not ensured. Quality of experience (QoE) is another indicator that defines user satisfaction, which reflects the subjective feelings of the users and originates directly from the user. Thus, QoE can be used to provide services that meet user expectations. The factors influencing QoE can be divided into three types: human, system, and context [6]. Briefly, the service either satisfies or disappoints the user.

Traditional DASH video playback scheduling algorithms often focus on smooth playback of the entire video when the network bandwidth is constrained [7], [8], [9]. These algorithms cannot adjust the video playback scheduling according to the user's true

* Corresponding author.
E-mail address: shchang@csie.fju.edu.tw

viewing QoE indicators. In this paper, we discuss the influence of different video resolutions on user viewing satisfaction based on two evaluation methods. The first method involves a series of click density (CD) experiments that can be performed to collect factors with which users are dissatisfied and define a quality score, termed the Q-score, based on the unsatisfactory click counts. However, many different types of videos need to be quantized into Q-scores, and it is impractical to perform CD experiments for each video. Instead, we proposed a second method in which the Q-score can be automatically matched with the amounts of motion vector variations in different scenes with different resolutions, thereby improving the efficiency of Q-score quantification.

Finally, we propose an algorithm, the optimal segment assignment (OSA) algorithm, which uses integer programming to schedule DASH video streaming for QoE optimization [10]. The main objective of OSA in terms of video playback scheduling is that when a large Q-score difference exists between resolutions, the OSA algorithm assigns a high-resolution video segment, and vice versa. This design concept can enhance the overall satisfaction of users and effectively utilize network bandwidth.

The remainder of this paper is organized as follows. Section II introduces related work. Section III presents the problem formulation and proposed algorithms, including the click density (CD) experimental environment and method, motion vector variation measurement using the block matching algorithm, distribution of unsatisfied click counts with different resolutions, the method for quantifying the Q-score, and the proposed OSA algorithm. Section IV presents the experimental results for user satisfaction obtained using our proposed method and the traditional popular algorithm. Finally, Section V presents our conclusions and topics for future work.

II. Related Work

A. User Feeling Measurement

Many factors affect the feelings of users, and it is not easy to collect and quantify these subjective feelings. Many previous researchers have focused on mapping network-related factors, such as the average playback bit rate, resolution switch count, buffer status, and download video quality, to judge users' quality of experience. Sakamoto et al. [11] proposed a bitrate selection method for adaptive video streaming for MPEG-DASH to improve the QoE by minimizing the bit rate fluctuation. Cetinkaya et al. [12] constructed a software-defined network (SDN) to reroute DASH flows to provide a fairness streaming service among DASH clients. Gao et al. [13] provided a deep learning model to characterize personalized QoE with temporal, spatial, and periodic correlations. With this classification of characteristics, they claimed that they could improve the personalized QoE for each specific user. They claimed that the results were more effective for QoE evaluation. Bentaleb et al. [14] proposed a software-defined network, named SDNDASH, based on dynamic network resource allocation and management. This architecture can manage and allocate network resources dynamically to improve the QoE for each user. Li et al. [15] proposed a QoE-driven mobile edge caching placement mechanism for dynamic adaptive video streaming with different rate-distortion characteristics. Zhao et al. [16] provided a robust adaptive algorithm at the client side for smoothing the streaming experience. They claimed that this mechanism can work stably under different network conditions. Lee et al. [17] proposed a segment-adjusted scheme based on the playback buffer status and network characteristics of the content. Cao et al. [18] proposed a QoE-friendly resolution-adaptation method that switches resolution less frequently and achieves smooth changes in resolution. Muller et al. [7] proposed a buffer level (BL) algorithm, which sets a 30s buffer to compensate for large bandwidth variations. The BL algorithm determines the resolution of the next requested video segment according to the state of the buffer occupancy of the client. When the buffer occupancy is at a lower level, a higher-resolution segment is retrieved during the next segment-retrieving cycle. On the other hand, when the buffer occupancy is at a higher level, a lower-resolution segment is retrieved. Alzahrani et al. [19] applied a machine-learning model to handle rate control to select the best network quality. Huang et al. [20] defined an integrated user QoE model and optimally controlled playback freezing, bitrate switch, and video playback quality by stabilizing the client buffer state. Yin et al. [21] used a theoretical approach for controlling video streaming over HTTP. Xin et al. [22] proposed a trunk-based request strategy to guarantee QoE in a P2P-VOD system.

Generally, if network traffic is sufficient to stream smooth video playback to each user, the user will have the best video viewing quality. However, these network-related factors may not truly reflect users' QoE. Chen et al. [23] proposed the OneClick experiment to describe the influences of various network factors, such as bandwidth, loss rate, and delay, on users' listening satisfaction. The original audio clip was divided into different clips according to various network factors, and each clip was cut into several audio segments. To provide the sound clip material to the user as a test, the first sound segment is the original sound segment, and the subsequent audio segments correspond to different situations and are randomly arranged to form a brand new testing sound clip, which does not overlap with the other segments. Because the experiment is intended to test sound satisfaction, the user must wear headphones to listen to the sounds to eliminate interference in the form of external sound, and the experiments are conducted through a keyboard and computer screen display. If the user is dissatisfied with the sound quality of the current segment, the blank key on the keyboard can be tapped, which is designated as an unsatisfactory click record. After the test, the database collected the satisfaction of different users for different audio segments. Because the user reacts by clicking after listening to the sound, there is a slight delay. Therefore, using a modified feedback delay time, it is possible to determine the actual time point of user dissatisfaction. The OneClick experiment is convenient because it can be performed on any computer at any time, as long as the computer system includes a keyboard, monitor, and headphones. In addition, the participants in the experiment did not need training beforehand; rather, they only needed to click on the appropriate key to indicate that the sound was unsatisfactory. Therefore, the OneClick experiment can reflect user listening satisfaction.

Additionally, many previous researchers have studied users' QoE based on the video content and users' viewing characteristics. Yue et al. [24] proposed a hybrid neural network model that integrates a deep neural network (DNN) and a recurrent neural network model (RNN) to learn an attention mechanism for user behavior analysis. Dimopoulos et al. [25] proposed a mechanism for detecting users' QoE degradations with three key influencing factors: stalling, average video quality, and quality variations. Zhao et al. [26] reviewed the QoE strategy for video transmission, including context and human factors. Engelke et al. [27] reviewed several psychophysiology-based QoE assessment methods. Hu et al. [28] proposed a semantic-aware adaption scheme termed SMA-PANDA to adapt video segments to DASH streaming. They used the k-means algorithm to classify motions into three types of video: slow moving, general walking, and rapid moving in a soccer sport movie. They mentioned that three types of video segments can be scheduled in a video playback to achieve a high QoE for users. However, except for the OneClick experiment, previous studies might not touch the users' true feelings. In this paper, we propose a segment assignment algorithm and test real users' viewing QoE using a series of sensory experiments.

B. Motion Vector in Motion Estimation Algorithm

Barjatya et al. [29] compared seven different block-matching algorithms to predict the movement of macroblocks while compressing a video. Among the motion estimation methods, the exhaustive search approach is the most accurate; however, the processing time is relatively long, and the motion estimation search process is employed. Motion estimation involves calculation of the change in the position of an object between two successive frames and the background in a macroblock to corresponding positions, and the calculated amount and direction of movement are recorded in a matrix to form a motion vector. Considering that the time cost is limited, a macroblock cannot be searched in the range of an entire frame. Therefore, the search range must be fixed; generally, the search parameter extends 7 pixels out of a macroblock with a length and width of 16 pixels. The search range will be formed as a rectangle area with a length and width of 25 pixels.

A larger number of search parameters increases the search cost. A macroblock is similar to another macroblock and is calculated on the basis of the cost function. The cost is minimized when the macroblock most closely corresponds to the current macroblock, that is, the most similar macroblock. There are many measures of cost, the most widely known and inexpensive being the mean absolute difference and mean squared error. Hosur et al. [30] used a motion vector for fast-motion estimation. Tourapis et al. [31] proposed an enhanced block-based search algorithm for motion estimation. Arora et al. [32] identified the initial search center dismisses that will be applied in any fast block-matching algorithm to find the motion vectors, rather than using a fixed search pattern. Kamble et al. [33] proposed a modified diamond search algorithm, which employed a small diamond-shaped search pattern in the initial step and a large diamond shape in further steps for handling fast motion estimation.

III. Problem Formulation and Proposed Algorithm

A. Problem Formulation

In this paper, we propose a measurement method for evaluating users' QoE, namely the Q-score. In this section, we formulate the problem to maximize users' QoE and propose an algorithm for determining a playback schedule that maximizes the Q-score. To define the problem clearly, we present some notations, as listed in Table I.

TABLE I. Notations for Problem Formulation and Model

Symbol	Definition
L	Startup latency.
N	Number of video segments in a DASH video.
ΔT	Playback duration of each video segment.
K	Number of different resolutions in a DASH video.
S_i	i^{th}-resolution video segment, where $1 \le i \le K$.
$\|S_i\|$	Size of the i^{th}-resolution video segment, where $1 \le i \le K$ and $\|s_1\| > \|s_2\| > \cdots > \|s_K\|$.
T	A DASH playback schedule consisting of t_j $T = [t_j \| 1 \le j \le N]$, where $t_j \leftarrow s_i$ and $1 \le i \le K$.
$q_{i,j}$	The specific Q-score gain while the i^{th}-resolution video segment, s_i, is scheduled into t_j of T.
$Q(T)$	The accumulated Q-score gain in a playback schedule T. $Q(T) = \sum_{j=1}^{N} q_{i,j}$, where $1 \le i \le K$.

In this study, the measured network bandwidth R was first assumed to be limited and fixed. The time required to download a video was divided into three periods. The first is the startup latency L, which involves preloading a video to start playback. Second, we assume that the video has N segments for playback and that the duration of each video segment is ΔT. Therefore, the total download time is $L + (N - 1) \times \Delta T$. Note that the last video segment should be downloaded successfully before it plays back.

In DASH video streaming technology, segments with the same video resolution are encoded with the same data size. We assume that there are K resolutions in total, and the video segments of each resolution are denoted as s_i and $|s_i|$ indicates the size of the segment s_i, where $1 \le i \le K$. The K^{th} resolution has the lowest resolution, thus $|s_1| > \cdots > |s_K|$. A playback schedule, T, consists of arranging a video segment, s_i, into each time slot, t_j, and formed as $T = [t_j | 1 \le j \le N]$, where $t_j \leftarrow s_i$, where $1 \le i \le K$. Additionally, while each video segment, s_i, is scheduled into t_j, the specific Q-score gain, $q_{i,j}$, is accumulated in $Q(T)$. Therefore, $Q(T) = \sum_{j=1}^{N} q_{i,j}$, where $1 \le i \le K$. The major goal of this study is to find a DASH playback schedule, T, which maximizes $Q(T)$.

To distinguish different user experience levels, we convert users' QoE into a Q-score and use the proposed algorithm to determine a playback schedule that maximizes the Q-score gain. We observed that the Q-score of a high-resolution video segment is greater than that of a low-resolution video segment for the same scenes, as confirmed by our proposed Click Density (CD) experiment, presented in the following sections.

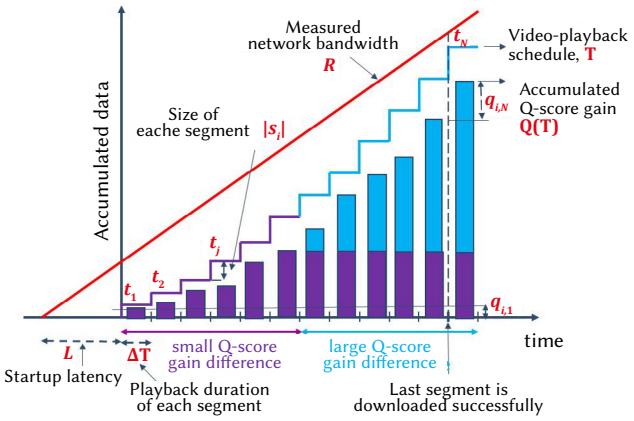


Fig. 1. Video-playback schedule with accumulated Q-score.

Fig. 1 presents an example of the video-playback schedule, T, in which the proposed algorithm allocates a different Q-score gain to video segments with a different resolution in a distinct scene. With the measured network bandwidth, R, low-resolution or high-resolution video segments occur in each playback duration, ΔT, of the playback schedule. In Fig. 1, we assume that the difference in the Q-score gain between high-resolution and low-resolution video segments is small in the first half of the video and large in the second half of the video. Then, the proposed OSA algorithm assigns low-resolution video segments in the area in which the Q-score gain difference between resolutions is smaller. Conversely, high-resolution video segments are chosen when the Q-score gain difference between resolutions is larger. Thus, the proposed OSA algorithm is designed to maximize users' Q-score gain.

Fig. 2 presents our framework for the designed CD experiment and the proposed OSA algorithm for calculating playback scheduling while maximizing the Q-score for video stream V. In our sensory experiments, we merged several different types of video into a test video (as shown in ①) and designed a click density (CD) experiment (as shown in ②) to obtain a Q-score model (as shown in ④) by a subjective test (as shown in ③). Then, we designed a motion vector estimation algorithm to calculate the average motion vector of this test video (as shown in ⑤).

Using regression analysis, we obtained a set of Q-score equations of different resolutions (as shown in ⑥), which transfers the Q-score of different resolutions obtained from the subject test results, QoE(CD), to that obtained by calculating the average motion vector, QoE(MV), (as shown in ⑦). Therefore, we obtain the Q-score of each segment with different resolutions, instead of performing subjective tests for each video stream. Furthermore, given a network condition *R*, we obtain the optimal playback scheduling with the maximum Q-score, Q(T), by the integer programming algorithm (as shown in ⑧). Hereafter, given a new video stream *V* (as shown in ⑨) and a network constraint *R* (as shown in ⑪), the playback schedule with maximum Q-score (as shown in ⑫) is arranged by calculating the average motion vectors (as shown in ⑩) instead of starting the subjective tests all over again.

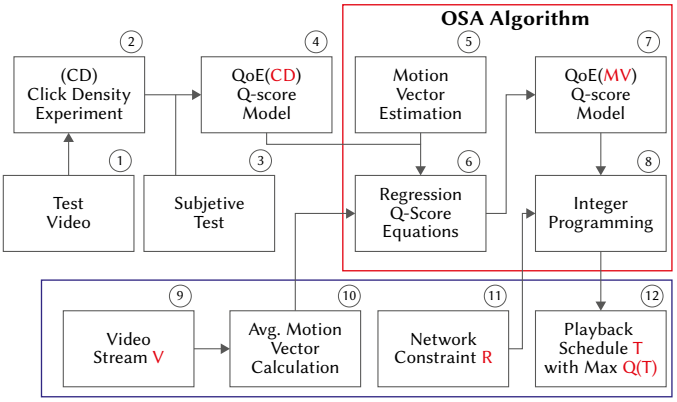


Fig. 2. Framework for the CD experiment and the OSA algorithm.

B. Test Video Sequence for Click Density (CD) Experiment

YouTube is currently one of the largest platforms for video streaming and includes a wide variety of video types [35]. For example, movies can be divided into science fiction films, action adventure films, literary and romantic films, and so on. In addition, YouTube uses a 16:9 aspect ratio player to match computer limitations, and the video resolution is proportional to the recommended aspect ratio. Because different video types correspond to different individual bitrates, different resolutions have different bitrates. To verify the user satisfaction based on the QoE in different video scenes, we chose three different categories—action adventure, sport, and family love films—as test materials. Action adventure films include fast-motion scenes, and the shots move quickly. Family films involve slow-motion scenes and fixed shots. Although sports and action adventure films both include fast-motion scenes, they differ in terms of shot distance. All of the videos were sourced from YouTube: the action adventure films were "Casino Royale" and "The Avengers," the family love film was "Why Him," and the sports film was "Lionel Messi - Skills & Goals."

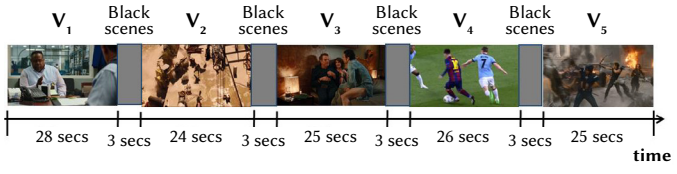


Fig. 3. Test video sequence composed of different types of video.

We merged these different types of video into a test video (as presented in Fig. 3) and inserted fast motion scenes between the slow motion scenes. To prevent adverse effects as a result of watching a video clip from a fast motion scene followed by one from a slow motion scene or conversely, we inserted black scenes with a duration of 3 seconds between consecutive video clips. Because the impact of resolution on the QoE for different types of videos was the focus of our study, we excluded the two most important influencing factors: sound and subtitles. Sound is an important influencing factor in videos; for example, horror movies are often paired with thrilling sound effects, and sound effects can integrate users into the story. In addition, subtitles significantly influence the visuals at different resolutions. For example, the degree to which the edges of subtitles are blurred is more dramatic in low-resolution videos. Thus, the test video sequence was composed without sound or subtitles. We used FFmpeg open-source tools to cut the video clips to produce the corresponding scenes and merge them into a test video sequence. The test video sequence consisted of five video clips that were presented at 30 fps. Details of the video clips are listed in Table II.

TABLE II. Test Video Sequence Information

Video Clip no.	Video Source	Type	Duration (secs)	Movement	
				Camera	Object
v1	*Why Him*	Family	28	Fixed	Slow
v2	*Casino Royale*	Adventure	24	Moving	Fast
v3	*Why Him*	Family	25	Fixed	Slow
v4	*Lionel Messi - Skills & Goals*	Sport	26	Moving	Fast
v5	*The Avengers*	Adventure	25	Moving	Fast

Most videos on YouTube have a maximum resolution of 1080p, so we chose resolutions of 1080p, 720p, 480p, and 360p as the test resolutions, and the corresponding bitrates are provided in Table III. The bitrates were chosen to enable users to perceive the differences between the resolutions.

TABLE III. Resolutions and Corresponding Bitrates

Resolution	Video size	Bitrate (kbps)
1080p	1920×1080	3000
720p	1280×720	1500
480p	854×480	500
360p	640×360	200

C. Proposed Click Density (CD) Experiment

We followed the concept of the OneClick experiment and modified it as the click density experiment for the video quality of the user experience test. Our click density experiment was modified to include two improvements. First, we changed the response device from the keyboard to the mouse. The click sensitivity with the mouse is higher than the keyboard effect, thus the differences in user satisfaction during the video test could be measured more accurately. Second, to retain the ability to conduct the test on any computer and facilitate the process for the test subjects, we set up a video test website. The following steps were performed to instruct the test subjects on how to use the video test website. Initially, a silent test video with a length of 2 min 20 s was played twice. First, the test video was played exactly as obtained from the source, which means that the highest bitrate and resolution were employed. The second time, one of the four test resolutions was randomly selected. Then, if the subject was dissatisfied with the current video segment during the video playback, the subject could click on the "unsatisfied button." If the unsatisfied button was clicked consecutively, it turned red to remind the subjects that they were highly dissatisfied with the current state and to facilitate user distinction of the level of dissatisfaction during the test. Finally, we reminded the subjects to use a monitor with at least 1024×768 resolution when performing the video test experiments and recommended not conducting the test on devices such as mobile phones or tablet PCs.

Considering that individuals have different levels of subjective

judgment, we used at least 30 test subjects to obtain data for each resolution. Then, we removed the black scenes between the video clips and modified the response delay times for the subjects. Finally, we aligned the unsatisfied click counts such that they corresponded with each video segment. The basic unit of the video was a video segment, each video segment lasted for 2 s, and the total length of the video was 64 segments.

Fig. 4 shows the average normalized click-count distributions corresponding to the four resolutions. There were 30 test subjects for 1080p, 30 for 720p, 37 for 480p, and 42 for 360p. Video segments 1–14 and 27–39 were slow-motion scenes, whereas video segments 15–26, 40–52, and 53–64 were fast motion scenes. As shown in Fig. 4, the average number of unsatisfied clicks for the high-resolution and low-resolution cases contrast each other. In other words, the lower the resolution, the higher the number of unsatisfied clicks. Moreover, with a low resolution, such as 360p or 480p, the average number of unsatisfied clicks was significantly higher for the fast motion scenes (video segments 15–26, 40–52, and 53–64) than in the slow motion scenes (video segments 1–14 and 27–39). With a high resolution of 720p ps, the same phenomenon is evident. However, the difference at 1080p was insignificant. To summarize, different resolutions have different QoE ranges for different scenes.

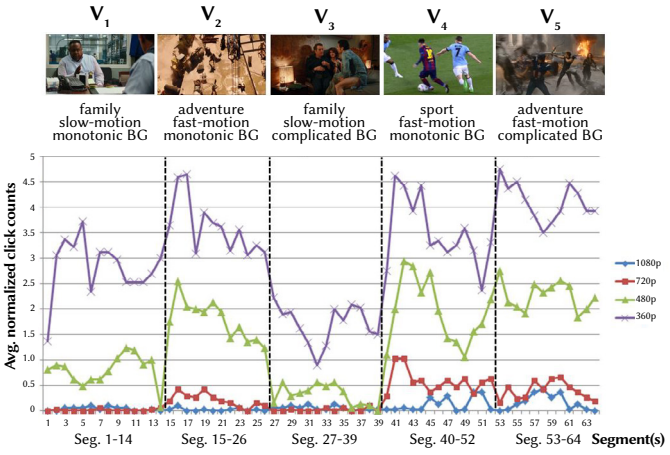


Fig. 4. Normalized dissatisfied click counts distribution.

D. QoE Quantification Q-score

In this section, we present the method for converting the user's QoE into a Q-score, which is defined in this paper. The absolute category rating (ACR) is a method for assessing differences in the QoE level in video or audio tests, and it consists of five levels: excellent, good, fair, poor, and bad. The number of unsatisfied clicks was converted into ACRs and 30 people were randomly selected from the test samples, thus the total number of samples from the four resolutions was 120. Because situations with zero clicks are equivalent to no dissatisfaction, they were evaluated as excellent.

As shown in Fig. 5, in addition to the zero-click case, the other unsatisfied clicks were sorted from smallest to largest according to the satisfaction level, and in the sample used in our CD experiment, the minimum number of clicks was 1 and the maximum was 19.

We then used the quartile method to match the rating quality. Zero unsatisfied clicks are equivalent to the highest Q-score of 5 points (as indicated in Table IV), and the remaining one to four points are allocated by different percentiles defined by the first quartile (Q1 = 2), second quartile (Q2 = 5), and third quartile (Q3 = 9).

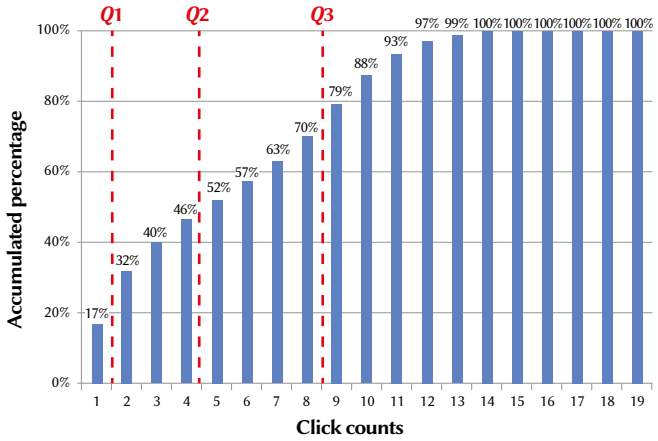


Fig. 5. Distribution of number of clicks excluding the zero click case.

TABLE IV. Correspondence Between Number of Clicks and ACR Rating

Q-score	Number of clicks (x)
5	0
4	$x < 2$
3	$2 \le x < 5$
2	$5 \le x < 9$
1	$x \ge 9$

Fig. 6 presents the average Q-score of each video segment with four resolutions. It is evident that in the slow motion scenes (video segments 1–14 and 27–39), the Q-scores corresponding to 1080p and 720p resolution fall between 5 and 4.9 points, whereas in the fast motion scenes (video segments 15–26, 40–52, and 53–64), the Q-score corresponding to the 720p resolution is obviously lower. However, the Q-score corresponding to a 1080p resolution exhibits no obvious change. Conversely, the Q-scores corresponding to the 480p and 360p resolutions were significantly higher in the slow motion scenes than in the fast motion scenes. Additionally, the difference in the Q-scores in the fast motion scenes was greater than that in the slow motion scenes. In other words, the effect is the same as that in the click distribution from the CD experimental results.

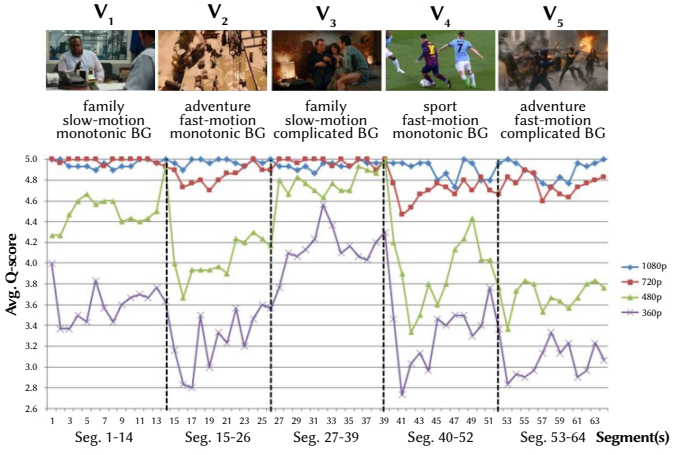


Fig. 6. Average Q-score distribution applied in our study.

E. Q-score Prediction By Variation in the Motion Vector

Conducting a CD experiment to obtain the user's Q-score before scheduling the playback of each video is impractical. Therefore, we used the exhaustive search method in the block-matching algorithm to perform motion vector calculations for each macroblock in a video

frame. We performed calculations for different resolutions of the motion vectors. Each video segment was 2 s long and the video was played at 30 fps, thus a video segment generated 60 frames. The length of the video segments was 64, and there were 3840 frames in total. Because our objective was to calculate the motion vector variation between frames, we compared the current frame with the frame that was two frames away from the current frame. For example, the 1st frame was compared with the 3rd frame, the 2nd frame was compared with the 4th frame, and so on (as presented in Fig. 7).

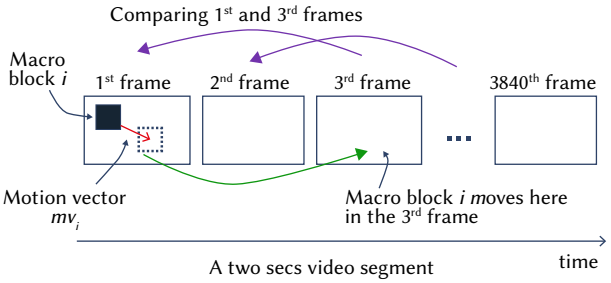


Fig. 7. Calculation of the motion vector, mv_i, between two frames.

After calculating all the motion vectors of the current frame, we averaged the vector length $|mv_i|$, for the current frame. The average motion vector variation is calculated using (1):

$$\frac{\sum_{i=1}^{M}|mv_i|}{M} \quad (1)$$

where M indicates the number of macro blocks in a video frame.

By comparing the corresponding frames in sequence, we obtained an average motion vector length of 3838 frames at four different resolutions. Fig. 8 compares the average motion variations of frames with two different resolutions, and the former resolution (indicated by "r1") and the latter resolution (indicated by "r2"). Taking the comparison of the 1080p and 720p resolutions as an example, the number of larger average motion vectors of frames at 1080p resolution is 3643 more than that at 720p resolution (represented as "r1 > r2"), which accounts for 95% of the total. Conversely, the number of smaller average motion vectors of the frames in the 1080p resolution video is 195 less than that in the 720p resolution video (represented as "r1 < r2"), accounting for 5% of the total. In Fig. 8, we demonstrate that the average motion vector length in high-resolution video is always larger than that in low-resolution video. Thus, the average motion vector length of the video segments can be used to represent the video quality among videos with different resolutions.

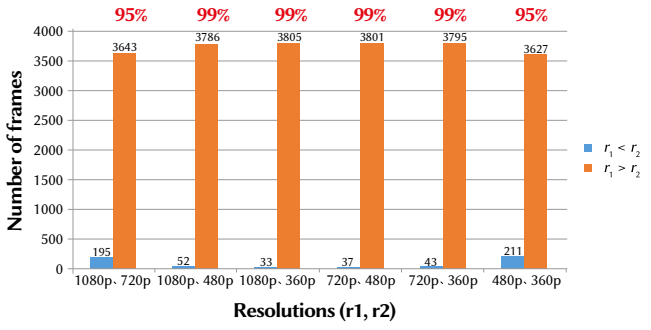


Fig. 8. Comparison of the average motion vector length of corresponding frames between two different resolutions.

Fig. 9 shows the average motion vectors of video segments with the four different resolutions. The frames were assigned to the video segments as follows: frames 1–60 constituted the first video segment, frames 61–120 were the second video segment, and the final frames (frames 3781–3838) formed the 64th video segment, which had only 58 frames. Therefore, Fig. 9 also shows that the higher the resolution, the higher the average motion vector length regardless of whether the scene contained fast or slow motion actions.

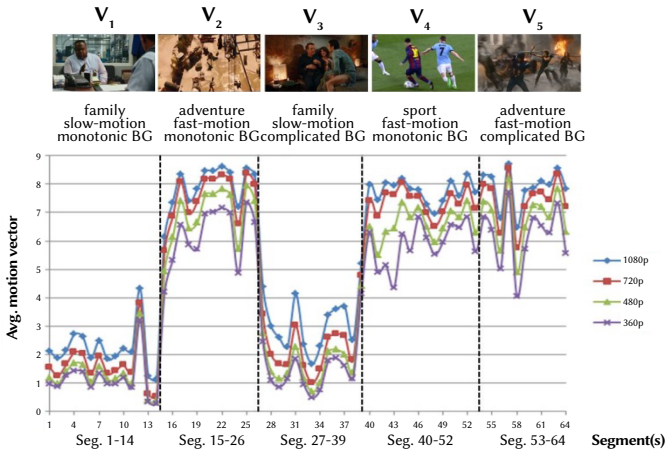


Fig. 9. Average length of motion vectors of video segments with different resolutions.

Furthermore, to evaluate whether the average motion vector length of video segments with four different resolutions was significant for the Q-score prediction model, we performed multiple regression and the Q-score was quantified by conducting our proposed CD experiment. Tables V–VIII contain the results of the ANOVA analysis of the motion vector and Q-score. Because multiple regression equations are established with different influencing factors, they are either highly or poorly correlated. In addition to the Q-scores corresponding to 1080p resolution, the others are significant for the Q-score prediction model. However, the p-values between groups indicate no significant difference, which means that the influencing factors may affect each other and cause the model to inaccurately predict the Q-score. Thus, we excluded the lowest relevant influencing factors and performed regression analysis of the other influencing factors until all of the influencing factors exhibited significant differences.

TABLE V. Results of ANOVA Analysis of Motion Vectors with Four Resolutions and Q-Score for 1080P Resolution

ANOVA analysis	Degree of freedom (DF)	Sum of squares (SS)	Mean of sum (MS)	F	Significance F
Regression	4	0.013	0.003		
Residual	59	0.294	0.005	0.631	0.643
Total	63	0.307	-		

	Coefficient	Standard error	t-statistic	p-value
Intercept	4.964	0.052	94.883	3.3×10^{-66}
1080p avg. mv	-0.028	0.075	-0.369	0.714
720p avg. mv	0.063	0.131	0.48	0.633
480p avg. mv	-0.038	0.111	-0.343	0.733
360p avg. mv	-0.005	0.048	-0.103	0.919

TABLE VI. Results of ANOVA Analysis of Motion Vectors with Four Resolutions and Q-Score for 720P Resolution

ANOVA analysis	Degree of freedom (DF)	Sum of squares (SS)	Mean of sum (MS)	F	Significance F
Regression	4	0.817	0.204		
Residual	59	0.4	0.007	30.15	1.14×10^{-13}
Total	63	1.217	-		

	Coefficient	Standard error	t-statistic	p-value
Intercept	4.894	0.061	80.237	6.05×10^{-62}
1080p avg. mv	0.32	0.088	3.644	5.69×10^{-4}
720p avg. mv	-0.7	0.152	-4.594	2.34×10^{-5}
480p avg. mv	0.473	0.129	3.664	5.33×10^{-4}
360p avg. mv	-0.103	0.056	-1.826	7.29×10^{-2}

TABLE VII. Results of ANOVA Analysis of Motion Vectors with Four Resolutions and Q-Score for 480P Resolution

ANOVA analysis	Degree of freedom (DF)	Sum of squares (SS)	Mean of sum (MS)	F	Significance F
Regression	4	8.983	2.246		
Residual	59	3.851	0.065	34.402	8.21×10^{-5}
Total	63	12.834	-		

	Coefficient	Standard error	t-statistic	p-value
Intercept	4.473	0.189	23.625	9.11×10^{-32}
1080p avg. mv	0.644	0.272	2.366	2.13×10^{-2}
720p avg. mv	-1.109	0.473	-2.346	0.022
480p avg. mv	0.194	0.401	0.484	0.63
360p avg. mv	0.219	0.174	1.254	0.215

TABLE VIII. Results of ANOVA Analysis of Motion Vectors with Four Resolutions and Q-Score for 360P Resolution

ANOVA analysis	Degree of freedom (DF)	Sum of squares (SS)	Mean of sum (MS)	F	Significance F
Regression	4	6.598	1.649		
Residual	59	5.685	0.096	17.118	2.27×10^{-9}
Total	63	12.283	-		

	Coefficient	Standard error	t-statistic	p-value
Intercept	3.562	0.23	15.485	1.9×10^{-22}
1080p avg. mv	0.836	0.331	2.528	0.014
720p avg. mv	-1.497	0.574	-2.608	0.012
480p avg. mv	0.648	0.487	1.331	0.189
360p avg. mv	-0.018	0.212	-0.086	0.932

The Q-score model prediction for the 1080p resolution is of little relevance to the other influencing factors, and the Q-score for 1080p resolution was always greater than 4.7 points; in other words, most users were satisfied with the 1080p resolution. Therefore, we directly set the Q-score prediction model for 1080p resolution to a maximum of five points. Tables IX–XI present the results of the ANOVA analysis for the other three resolutions and their Q-scores, and all of the influencing factors have significant differences at the $\alpha = 0.05$ level. The coefficients of determination R2 for the other three resolutions were 0.653, 0.699, and 0.537 for the 720p, 480p, and 360p resolutions, respectively.

TABLE IX. Results of ANOVA Analysis of Motion Vectors with Three Resolutions and Q-Score for 720P Resolution

ANOVA analysis	Degree of freedom (DF)	Sum of squares (SS)	Mean of sum (MS)	F	Significance F
Regression	3	0.795	0.265		
Residual	60	0.422	0.007	37.63	8.3×10^{-14}
Total	63	1.217	-		

	Coefficient	Standard error	t-statistic	p-value
Intercept	4.917	0.061	80.679	6.8×10^{-63}
1080p avg. mv	0.261	0.083	3.138	2.63×10^{-3}
720p avg. mv	-0.538	0.126	-4.256	7.41×10^{-5}
480p avg. mv	0.268	0.065	4.135	1.12×10^{-4}

TABLE X. Results of ANOVA Analysis of Motion Vectors with Three Resolutions and Q-Score for 480P Resolution

ANOVA analysis	Degree of freedom (DF)	Sum of squares (SS)	Mean of sum (MS)	F	Significance F
Regression	3	8.968	2.989		
Residual	60	3.867	0.064	46.384	1.23×10^{-15}
Total	63	12.834	-		

	Coefficient	Standard error	t-statistic	p-value
Intercept	4.504	0.177	25.431	7.89×10^{-34}
1080p avg. mv	0.565	0.217	2.609	1.14×10^{-2}
720p avg. mv	-0.914	0.244	-3.741	4.13×10^{-4}
360p avg. mv	0.292	0.085	3.424	1.12×10^{-3}

TABLE XI. Results of ANOVA Analysis of Motion Vectors with Three Resolutions and Q-Score for 360P Resolution

ANOVA analysis	Degree of freedom (DF)	Sum of squares (SS)	Mean of sum (MS)	F	Significance F
Regression	3	6.597	2.199		
Residual	60	5.686	0.095	23.206	4.28×10^{-10}
Total	63	12.283	-		

	Coefficient	Standard error	t-statistic	p-value
Intercept	3.566	0.224	15.948	3×10^{-23}
1080p avg. mv	0.826	0.306	2.702	8.95×10^{-3}
720p avg. mv	-1.469	0.464	-3.165	2.44×10^{-3}
480p avg. mv	0.611	0.238	2.573	1.26×10^{-2}

In the regression model, the predicted Q-score is y, the influencing factors corresponding to 1080p, 720p, 480p, and 360p resolutions are x_1, x_2, x_3, and x4, respectively, and α_i is the coefficient of the influencing factor. The regression formula is presented in (2):

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_4 x_4 \quad (2)$$

The results led us to conclude that the variation in the average motion vector is higher, regardless of whether the scenes were contained in fast or slow motion segments. To evaluate whether the average motion vector variation of video segments with four different resolutions was significant for the Q-score prediction model, we performed multiple regression and ANOVA analyses. In the regression model, the predicted Q-score of the video segment was y. The calculated average motion vector variations in the 1080p, 720p, 480p,

and 360p resolutions are x_1, x_2, x_3, and x_4, respectively. We obtained regression equations to predict the Q-score of a video segment with different resolutions (as presented in Table XII).

TABLE XII. Equations for Q-score Prediction with Four Resolutions

Resolution	Q-score regression equation (y)
1080p	y=5
720p	$y=4.917+0.261x_1-0.538x_2+0.268x_3$
480p	$y=4.504+0.565x_1-0.914x_2+0.292x_4$
360p	$y=3.566+0.826x_1-1.469x_2+0.611x_3$

Fig. 10 shows the multiple regression results for the Q-score prediction model with four resolutions. In the slow-motion scenes, the Q-score differences between resolutions were small, whereas they were larger in the fast motion scenes. This phenomenon is the same as that in the original Q-score distribution obtained from the CD experiment. Therefore, we also used the correlation coefficient to analyze the difference between the predicted and original Q-scores. Fig. 10 shows the multiple regression results for the Q-score prediction model with four resolutions, which are similar to the results presented in Fig. 6.

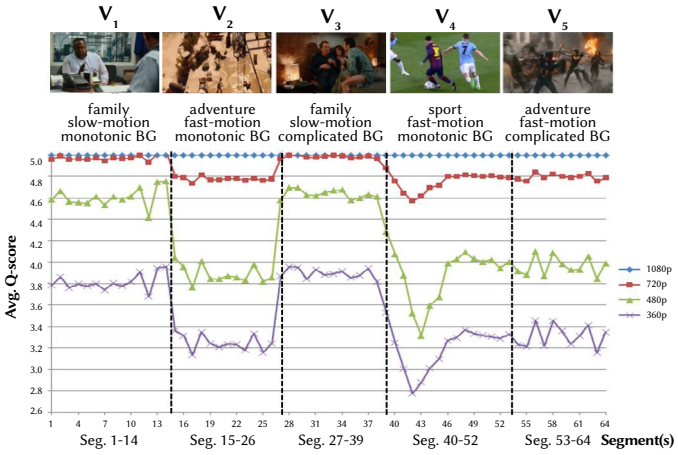
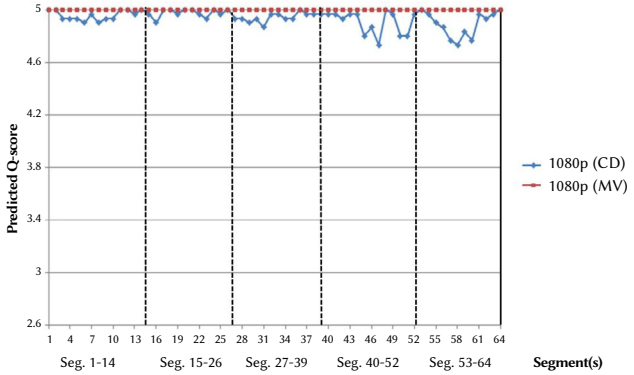


Fig. 10. Average Q-score distribution based on multiple regression analysis.

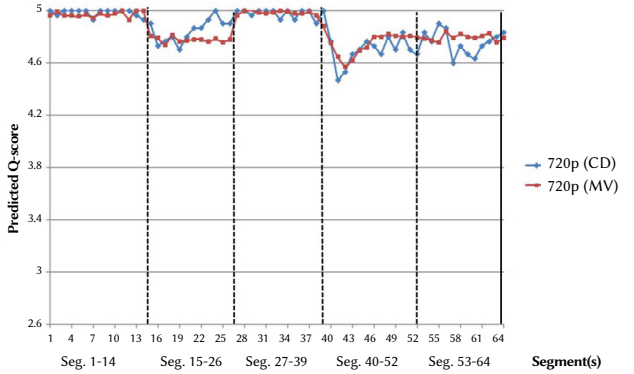
Fig. 11 shows the predicted Q-scores with the click density (CD) and motion vector (MV) at four resolutions. Because the Q-score predictions from motion vectors in the video with 1080p resolution always equal 5 points, the correlation calculation could not be applied. However, Fig. 11(a) shows that the Q-score traces between these two methods are similar. Fig. 11(b), Fig. 11(c), and Fig. 11(d) present the results for the 720p (correlation = 0.812), 480p (correlation = 0.836), and 360p (correlation = 0.733) resolutions, respectively. From these results, we conclude that the predicted Q-score from the motion vector method is close to the true user QoE.

F. Integer Programming Algorithm

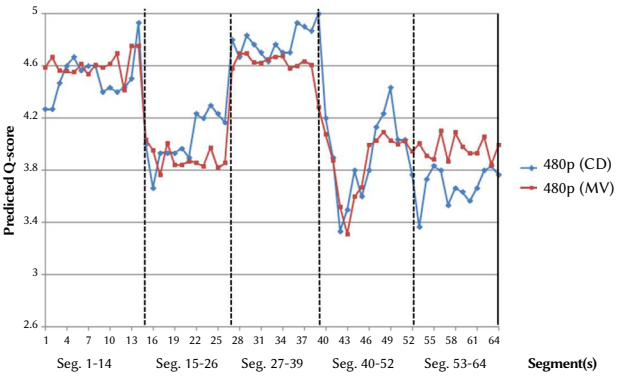
The differences in the Q-scores between the different resolutions in the slow-motion scenes calculated with the proposed Q-score prediction model were small, but were larger for the fast motion scenes. This finding is the same as that for the original Q-score distribution obtained from the CD experiment. To schedule video playback in which high-resolution video segments are assigned to fast motion scenes and low-resolution video segments to slow motion scenes, we applied integer programming to maximize the overall average Q-score gain [10]. The playback schedule T, is a combination of N video segments. We present this playback scheduling, T, in (3). Each video segment selects only one resolution among K resolutions.



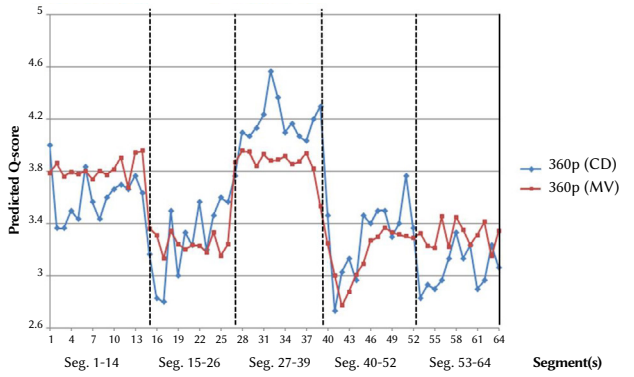
(a) Video resolution 1080p



(b) Video Resolution 720p



(c) Video resolution 480p



(d) Video resolution 360p

Fig. 11. Comparison of predicted Q-scores calculated with the click density (CD) and motion vector (MV) methods for four different video resolutions.

The $\lambda_{i,j}$ indicates that video segment i selects the j^{th} resolution and has a choice of 0 or 1.

$$T = \begin{bmatrix} \lambda_{1,1} & \lambda_{2,1} & \cdots & \lambda_{N,1} \\ \lambda_{1,2} & \lambda_{2,2} & \cdots & \lambda_{N,2} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{1,K} & \lambda_{2,K} & \cdots & \lambda_{N,K} \end{bmatrix} \quad (3)$$

If the j^{th} resolution of segment i is chosen, the Q-score gain, $q_{i,j}$, is accumulated in $Q(T)$. Because the objective is to maximize the average Q-score gain, $Q(T)$, the target function is expressed by (4):

$$\text{Maximize } Q(T) = \sum_{i=1}^{N} \sum_{j=1}^{K} q_{i,j} \times \lambda_{i,j} \quad (4)$$

When each video segment is selected, there are K resolutions to choose from, and each resolution has its own Q-score. However, under different network bandwidth conditions, R, the selection is subject to two restrictions:

$$\sum_{j=1}^{K} \lambda_{i,j} = 1, 1 \le i \le N \quad (5)$$

$$R \times (L + (i-1) \times \Delta T) - \sum_{x=1}^{i} |s_x| \ge 0, 1 \le i \le N \quad (6)$$

Equation (5) is the first restriction when playing a video, and only one resolution is selected for each video segment, $\sum_{j=1}^{K} \lambda_{i,j} = 1$. Equation (6) presents the second restriction when the resolution suitable for video playback is selected in the video segment. Additionally, the entire amount of data transmitted, $\sum_{x=1}^{i} |s_x|$, are prevented from exceeding the total data size the given network bandwidth is able to transmit, $R \times (L + (i-1) \times \Delta T)$. We used the integer programming method and proposed the OSA algorithm to determine the best solution for video playback with a video length of 64 segments. With limited network bandwidth, the OSA algorithm can be used to optimize video playback.

IV. Experimental Results

To compare the video playback schedules that were edited using the different algorithms, we set up a video test platform to provide subjects with video satisfaction tests. BL is a popular scheduling algorithm for MPEG-DASH streaming applications [7]. Therefore, we applied this BL algorithm to test the effectiveness of our proposed OSA algorithm to improve the QoE of the user. The steps of the experiment were as follows. Two test videos were used: the first was scheduled using the OSA algorithm and the other using the BL algorithm.

Although the two videos contained the same movie scenes, the video segments of different resolutions were arranged in different ways in the video playback. Subjects who were participating in the experiment viewed the two videos in randomly generated order to prevent the subject from having an established impression. After the first video had finished playing, to prevent the subject from forgetting the current feeling of satisfaction with the video, we asked the user to provide a rating regarding the feeling. Moreover, the rating was referenced by ACR, the five-level rating scale for different QoE levels, and the second video was shown using the same process as before. Finally, once each subject had completed the two different video tests, we asked the subject to again indicate which one they were more satisfied with, to confirm the consistency of the answers. There were three options to choose from: the first video, the second video, or no difference. The user satisfaction tests were conducted with four different resolutions and with playback bitrates of 3000 (1080p), 1500 (720p), 500 (480p), and 200 (360p) kbps. The playback duration of each video segment was 2 s, and the total number of segments in each video was 64. Therefore, the video length was 2 min 8 s, and the videos did not include sounds or subtitles. For a fair comparison, the startup latency for running each algorithm was set to 2 s with a constant network bandwidth, and none of the videos froze. To distinguish which QoE metrics applied to the playback schedule, OSA_C indicates that the OSA algorithm applied QoE metrics from click density (CD) experiments and OSA_M represents that the OSA algorithm applied QoE metrics from the motion vector (MV) method.

A. Comparison of OSA_C and BL Algorithms

In this section, we compare the user satisfaction results of the BL algorithm with a 30s buffer with those obtained with the OSA_C algorithm. Furthermore, the results obtained with a network bandwidth of 900 kbps are presented. The experimental results for the different videos are shown in Fig. 12. We demonstrate that, among the two algorithms, the OSA_C algorithm has the largest amount of accumulated data and an average bitrate of 896.875 kbps, followed by the BL algorithm with a 30s buffer and an average bitrate of 628.125 kbps. The OSA_C algorithm processes the playback schedule based on the QoE metrics from the click density (CD) experiments. The results in Fig. 12 show that the OSA_C algorithm allocates high-resolution video segments starting from segments 15 and 46, because the gap in user satisfaction becomes wider. Therefore, the amount of accumulated data increased sharply between video segments 15 and 46. Conversely, the BL algorithm processes the resolution adaption only on the buffer occupation, and it does not adapt the video resolution in these time slots.

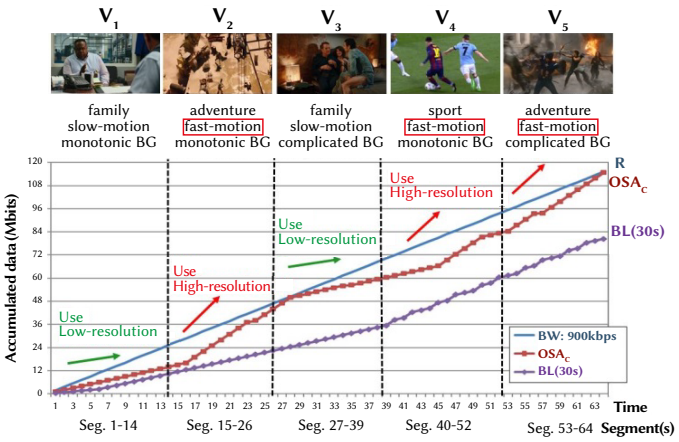


Fig. 12. Two playback schedules processed by the OSA_C and BL algorithms with a network bandwidth constraint of 900 kbps.

The data presented in Fig. 13 were obtained from 42 test subjects. The distributions can be divided into three parts: above the line, below the line, and on the line. The line represents the score at which the results for the OSA_C and BL algorithms were the same. The distributions above and below the line correspond to subjects who preferred the video playback schedules edited by either the BL or the OSA_C algorithm, respectively. User satisfaction was rated from 1 to 5 points, and the different colors represent the situations the subjects selected when reconfirming their satisfaction status. Finally, the size of each pie chart represents the number of people who chose the score in that case. For example, in the case of the largest pie chart, which represents eight subjects, five subjects chose the BL algorithm with two points and the OSA_C algorithm with three points. Upon reconfirmation, this result indicated that user satisfaction with the OSA_C algorithm was higher than that with the BL algorithm. The three different situations represented in the bar chart in Fig. 14 correspond to the number of subjects who assigned higher scores to either the OSA_C or BL algorithm or assigned the same score to both. Additionally, the subjects chose which playback schedule they preferred. Eighteen users scored the schedule derived using the OSA_C algorithm higher than

that produced by the BL algorithm, and they considered the former algorithm to yield superior results. The experimental results shown in Fig. 14 indicate that most of the subjects felt that the video playback scheduled using the OSA_C algorithm provided a more satisfactory visual experience.

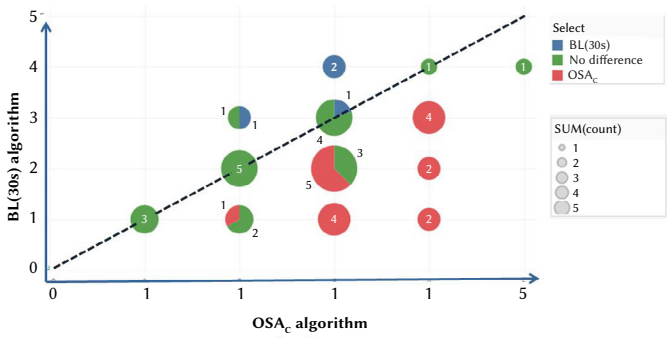


Fig. 13. Pie chart of user satisfaction with the OSA_C and BL algorithms (buffer = 30s).

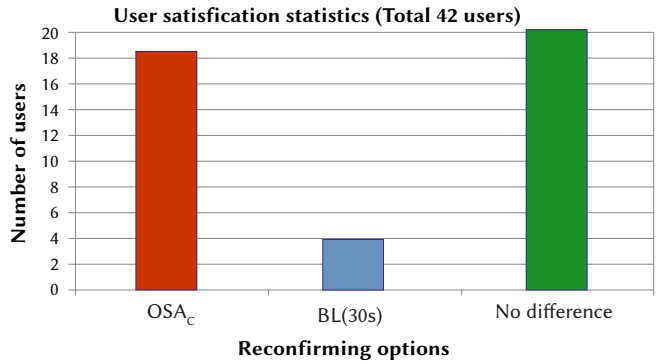


Fig. 14. User satisfaction statistics for the OSA_C and BL algorithms (buffer = 30s).

Fig. 14 shows that the satisfaction counts obtained by applying the OSA_C algorithm are higher than those obtained by applying the BL algorithm with a 30-s buffer. To confirm whether the two different video playbacks were distinct, we conducted a paired sample t-test to analyze the significance of the satisfaction scores. The null hypothesis was that the effect of the two algorithms was insignificant. Table XIII demonstrates that the differences between the experiments were significant at the $\alpha = 0.01$ level. The p-values from the first and second experiments were both less than 0.01, refuting the null hypothesis. Thus, the video playback scheduled using the OSA_C algorithm provides a more satisfactory user experience.

TABLE XIII. T-TEST OF SATISFACTION SCORE: OSA_C VS. BL (BUFFER = 30S)

Algorithm	Sum	Average	Variance	p-value
OSA_C	121	2.88	0.8391	3.39×10^{-5}
BL	91	2.17	0.9228	

B. Comparison of OSA_M and BL Algorithms

The experimental results presented in Section IV-A show that video playback scheduled with QoE metrics from the CD experiment delivers higher streaming performance. In this section, we applied the QoE metrics obtained from the motion vector (MV) method and arranged a playback schedule using the OSA_M algorithm. We then performed video playback scheduling using the OSA_M algorithm and the original approach using the BL algorithm with a 30-s buffer to determine which result is more appealing to users. Therefore, we chose a network bandwidth of 900 kbps, where the video playback was composed of high-resolution and low-resolution video segments. Fig. 15 depicts the results of the two different video playback schedules with a network bandwidth of 900 kbps. The average bitrate of the OSA_M algorithm is 875 kbps, which is slightly lower than that of the OSA_C algorithm but much higher than that of the BL algorithm with a 30-s buffer. Additionally, the OAS_M algorithm uses high-resolution video segments starting from video segments 15 and 46 because of the fast motion scenes.

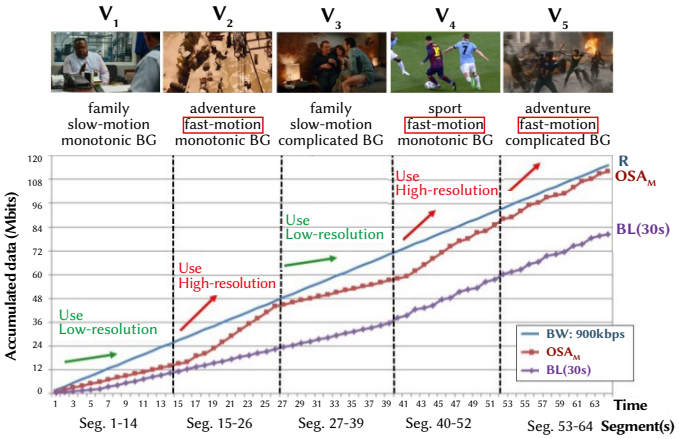


Fig. 15. Two playback schedules handled by OSA_M and BL algorithms with a network bandwidth of 900 kbps.

The user feedback, shown in Fig. 16, is that of 30 participants who participated in the test. The experimental results indicate that most users rated the OSA algorithm higher than the BL algorithm with a 30-s buffer. Moreover, the larger pie charts are mostly below the line, which means that the subjects were more satisfied with the OSA algorithm. Fig. 17 demonstrates that more than half of the subjects tended to choose the OSA algorithm instead of the BL algorithm with a 30-s buffer. Thus, with a network bandwidth of 900 kbps, the QoE gain scheduled using the OSA algorithm was significantly more satisfactory to the test users.

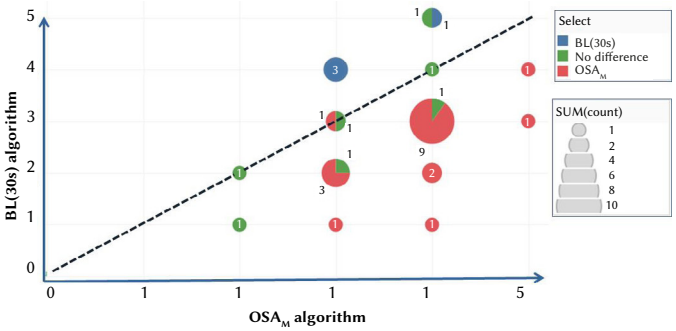


Fig. 16. Pie chart of user satisfaction with the OSA_M and BL algorithms (buffer = 30s).

Because of the mechanism of the BL algorithm, it always chooses low-resolution video segments at the beginning, and when the buffer occupancy is sufficient, the BL algorithm switches to high-resolution video segments. In contrast, the OSA_M algorithm always uses high-resolution video segments for video streaming. To confirm whether the two different video playbacks were distinct, we conducted a paired sample t-test to analyze the significance of the satisfaction scores. Fig. 17 shows that the satisfaction counts obtained by applying the OSA_M algorithm are higher than those obtained by applying the BL algorithm with 30-s buffering. The null hypothesis was that the effect of the two algorithms was insignificant. The results in Table XIV demonstrate

that both experiments have significant differences at the $\alpha = 0.01$ level. The p-values from the first and second experiments were both less than 0.01, refuting the null hypothesis. Thus, the video playback scheduled using the OSA_M algorithm provides a more satisfactory user experience.

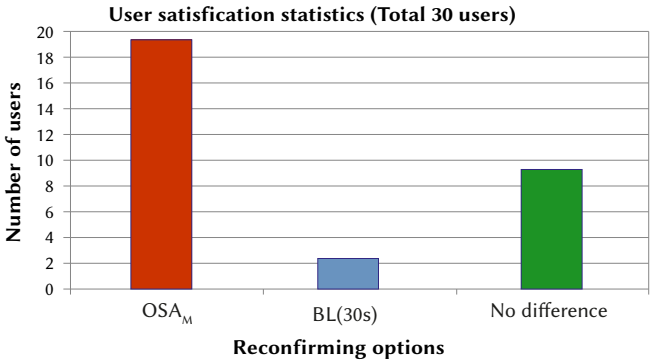


Fig. 17. User satisfaction statistics for the OSA_M and BL algorithms (buffer = 30s).

TABLE XIV. T-TEST OF SATISFACTION SCORE: OSA_M VS. BL (BUFFER = 30S)

Algorithm	Sum	Average	Variance	p-value
OSA_M	108	3.6	0.5241	4.48×10^{-4}
BL	86	2.87	1.085	

V. Conclusions and Future Work

DASH streaming is a popular method for video streaming over the Internet. The QoE score measures the extent to which the user is satisfied with the viewing experience but is difficult to evaluate, and the factors influencing the QoE, including the user's subjective feeling and sound environment, are complicated. The current DASH streaming scheduling algorithms only consider the network constraints, and the primary objective of these algorithms is to provide stable video playback, instead of focusing on the QoE. To evaluate the factors influencing the QoE, we designed a series of click density (CD) experiments to collect unsatisfactory click counts in different scenes with different video resolutions. The click distributions in the CD experiment indicated that the test subjects were usually unsatisfied with fast motion scenes and relatively satisfied with slow motion scenes at the same video resolution. In other words, the use of high-resolution video segments in high-motion scenes would significantly improve the QoE. Additionally, we observed that the difference between the two levels of resolution was greater in high-motion scenes than in low-motion scenes. Therefore, we defined an ACR five-level Q-score for rating the quality of different QoE levels, applied integer programming, and proposed the OSA algorithm to generate video playback schedules that maximized the Q-score gain with network bandwidth constraints. Then, we designed a subjective experiment to test user satisfaction with our proposed OSA algorithm and the most popular DASH streaming algorithm, the BL algorithm. In addition to converting clicks into Q-scores, we applied multiple regression to derive the correlation of the motion vector variations with different resolutions with the Q-scores. The experimental results show that the playback videos generated using the OSA algorithm with the motion vector model also increased user satisfaction. In this study, we only used fast or slow motion to investigate user satisfaction during different scenes of different video resolutions, where motion speed is an influencing factor. Thus, we aim to incorporate other influencing factors and modify our proposed Q-score prediction algorithms for video playback scheduling in the near future.

References

[1] A. C. Begen, T. Akgul, and M. Baugher, "Watching video over the web: Part 1: Streaming protocols," *IEEE Internet Computing*, vol. 15, no. 2, pp. 54–63, 2010.

[2] I. Sodagar, "MPEG-DASH standard for multimedia streaming over the internet," *IEEE Multimedia*, vol. 18, no. 4, pp. 62–67, 2011.

[3] T. Stockhammer, "Dynamic Adaptive Streaming over HTTP - Standards and design principles," *in Proceedings of the 2nd Annual ACM Conference on Multimedia Systems*, pp. 133–144, 2011.

[4] Pantos and May, Apple, Inc., "HTTP live streaming," IETF RFC 8216, July, 2017. [online] Available: http://tools.ietf.org/html/draft-pantos-http-live-streaming-23

[5] A. M. Tekalp, "Digital Video Processing," *2nd edition, Prentice Hall Press Upper Saddle River*, NJ, USA, 2015.

[6] U. Reiter, K. Brunnström, K. D. Moor, M. C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, "Factors influencing quality of experience," *Quality of Experience, Advanced Concepts, Applications and Methods*, pp. 55–72, 2014.

[7] C. Müller, S. Lederer and C. Timmerer, "An Evaluation of Dynamic Adaptive Streaming over HTTP in Vehicular Environments," *in Proceedings of the 4th ACM International Multimedia Conference on Mobile Video*, pp. 37–42, 2012.

[8] A. Zambelli, "IIS smooth streaming technical overview," *Microsoft documentation, Microsoft Inc.*, vol. 3, 2009.

[9] M. Levkov, "Video encoding and transcoding recommendations for HTTP dynamic streaming on the Adobe Flash platform," *White Paper, Adobe Systems, Inc.*, 2010.

[10] A. Schrijver," Theory of Linear and Integer Programming," *John Wiley & Sons Inc.*, ISBN 978-0-471-90854-8, 1986.

[11] R. Sakamoto, T. Shobudani, R. Hotchi, and R. Kubo, "QoE-Aware Stable Adaptive Video Streaming Using Proportional Derivative Controller for MPEG-DASH," *IEICE Transactions on Communications*, vol. E104.B no. 3 pp. 286–294, 2021.

[12] C. Cetinkaya, K. Herguner, C. Hellge, and M. Sayit, "Segment-aware dynamic routing for DASH flows over software-defined networks," *International Journal of Network Management*, vol. 30, issue 4, 2020.

[13] Y. Gao, X. Wei, L. Zhou, "Personalized QoE improvement for networking video service," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2311–2323, 2020.

[14] Bentaleb, Abdelhak, A. C. Begen, and R. Zimmermann. "SDNDASH: Improving QoE of HTTP adaptive streaming using software defined networking," *in Proceedings of the 24th ACM International Conference on Multimedia*, pp.1296–1305, 2016.

[15] C. L. Li, L. Toni , J. Zou, H. Xiong, and P. Frossard, "QoE-driven mobile edge caching placement for adaptive video streaming," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 965–984, 2017.

[16] S. Zhao, Z. Li, D. Medhi, P. L. Lai, and S. Liu, "Study of user QoE improvement for dynamic adaptive streaming over HTTP (MPEG-DASH)," *in Proceedings of the IEEE International Conference on Computing, Networking and Communications (ICNC)*, pp. 566–570, 2017

[17] S. H. Lee, E. Lee, and H. W. Lee, "Quality adaptation scheme for improving QoE of MPEG DASH," *in Proceedings of IEEE International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 368–370, 2016.

[18] Y. M. Cao, X. Q. You, J. Wang, and L. Song, "A QoE Friendly Rate Adaption Method for DASH," *in Proceedings of IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pp. 1–6, 2014.

[19] I. R. Alzahrani, N. Ramzan, S. Katsigiannis, and A. Amira, "Use of Machine Learning for Rate Adaptation in MPEG-DASH for Quality of Experience Improvement," *in Proceedings of the 5th International Symposium on Data Mining Applications*, pp. 3–11, 2018.

[20] W. W. Huang, Y. P. Zhou, X. Y. Xie, D. Wu, M. Chen, and E. Ngai, "Buffer State is Enough: Simplifying the Design of QoE-Aware HTTP Adaptive Video Streaming", *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 590–601, 2018.

[21] X. Q. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP," *in Proceedings of the ACM SIGCOMM* , pp. 325–338, 2015.

[22] W. Xin, P. Chuan, L. Zhou and Y. Qian, "QoE Oriented Chunk Scheduling

in P2P-VoD Streaming System," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 8012–8025, 2019.

[23] K. T. Chen, C. C. Tu, and W. C. Xiao, "OneClick: A framework for measuring network quality of experience," *in Proceedings of IEEE INFOCOM*, pp. 702–710, 2009.

[24] T. Yue, H. Wang, S. Chen, and J. Shao, "Deep learning based QoE evaluation for internet video," *Neurocomputing*, vol. 386, pp.179–190, 2020.

[25] G. Dimopoulos, I. Leontiadis, P. B. Ros, and K. Papagiannaki, "Measuring video QoE from encrypted traffic," *in Proceedings of the Internet Measurement Conference*, pp. 513–526, 2016.

[26] T. Zhao, Q. Liu, and C. W. Chen, "QoE in video transmission: A user experience-driven strategy," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 285–302, 2017.

[27] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J. N. Antons, K. Y. Chan, N. Ramzan, and K. Brunnstro, "Psychophysiology-based QoE assessment: A survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 6–21, 2016.

[28] S. H. Hu, L. F. Sun, C. X. Xiao, and C. Gui, "Semantic-aware adaptation scheme for soccer video over MPEG-DASH," *in Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 493-498, 2017.

[29] A. Barjatya, "Block matching algorithms for motion estimation," *IEEE Transactions on Evolution Computation.* vol. 8. pp. 225–239, 2004.

[30] Hosur, P. Irappa, and K. K. Ma. "Motion vector field adaptive fast motion estimation." *in Proceedings of the 2nd International Conference on Information, Communications and Signal Processing (ICICS'99)*, pp. 7–10, 1999.

[31] A. M. Tourapis, O. C. L. Au, and M. L. Liou, "Predictive motion vector field adaptive search technique (PMVFAST): enhancing block-based motion estimation," *Visual Communications and Image Processing*, vol. 4310, pp. 883–892, 2001.

[32] S. Arora, K. Khanna, and N. Rajpal," A Novel Hybrid Approach for Fast Block Based Motion Estimation," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 6, pp. 24–30, 2017.

[33] S. Kamble, N. Thakur, A. Samdurkar, and A. Patharkar, "Object Detection and Tracking using Modified Diamond Search Block Matching Motion Estimation Algorithm," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 1, pp. 73–85, 2018.

[34] D. K. Krishnappa, D. Bhat, and M. Zink, "DASHing YouTube: An analysis of using DASH in YouTube video service," *in Proceedings of the 38th Annual IEEE Conference on Local Computer Networks*, pp. 407–415, 2013.

[35] YouTube genres and categories. [online] Available:https://developers.google.com/youtube/v3/live/guides/encoding-with-dash

Shin-Hung Chang

Shin-Hung Chang received his Ph.D. and M.S. degrees in Computer Science and Information Engineering from National Taiwan University in 2005 and 1998, respectively. He received his B.S. degree in Computer Science and Information Engineering from Fu Jen Catholic University in 1996. He joined the Institute of Information Science (IIS), Academia Sinica as a research assistant from 1998 to 2005 and as a post-doctoral fellow from 2005 to 2006. Dr. Chang is responsible for patent matters in the Computer System and Communication Laboratory (CSCL) of the IIS, and he holds one United States patent and four Taiwanese patents in computer screen recording technology. In the IIS, Dr. Chang was also responsible for transferring several technologies to industry. After his postdoctoral work, Dr. Chang joined an American company, MagnetoX Co. Ltd., as an R&D team manager. He led several multimedia application development projects and handled patent matters from 2006 to 2007. Dr. Chang joined Fu Jen Catholic University as a professor in 2007. Dr. Chang coached students to attend many ICPC programming contests and hosted several Ministry of Science and Technology (MOST) research projects during his teaching period at Fu Jen Catholic University. Additionally, Dr. Chang is an IEEE member and IEEE computer society member. Dr. Chang's research interests cover the integration of theory and application, including machine-learning algorithms, network protocol design, multimedia applications, media streaming, media codec algorithms, slot machine game model construction, and real-time strategy (RTS) game AI.

Min-Lun Tsai

Min-Lun Tsai received his M.S. and B.S. degrees in Computer Science and Information Engineering from Fu Jen Catholic University in 2018 and 2016, respectively. Mr. Tsai joined the Institute of Information Science (IIS), Academia Sinica, as a research intern in 2017. After earning his M.S. degree, he accepted a position as software engineer at ACTi Corporation, where he is currently employed, and develops IP camera applications, such as intelligent retail solutions, digital signage with age/gender recognition, and window applications. Mr. Tsai's research interests cover media streaming, multimedia applications, stock analysis, app development, and recognition algorithms.

Meng-Huang Lee

Meng-Huang Lee received his B.S. and M.S. degrees in Electrical Engineering from National Cheng Kung University in 1987 and 1989, respectively, and his Ph.D. degree in Computer Science and Information Engineering from National Taiwan University in 1996. He is currently a professor at the Department of Information Technology and Management, Shih Chien University. His research interests include multimedia systems, IPTV, and computer networks.

Jan-Ming Ho

Jan-Ming Ho received his Ph.D. degree in electrical engineering and computer science from Northwestern University in 1989. He received his M.S. degree at the Institute of Electronics of National Chiao Tung University in 1980 and his B.S. degree in electrical engineering from National Cheng Kung University in 1978. Dr. Ho joined the Institute of Information Science, Academia Sinica, as an Associate Research Fellow in 1989, and was promoted to Research Fellow in 1994. During the years 2000-2003, he served as the Deputy Director of the institute. In 2004-2006, he served as Deputy Director of IIS, Academia Sinica in 2000-2003, and as Director General of the Division of Planning and Evaluation, National Science Council. During his term as Deputy Director of the IIS, he was responsible for the planning and operation of the National Digital Archive Program, organized by the Summer Institute on Bioinformatics in 2001. He also founded the Open Source Foundry and organized a series of events promoting open sources. He co-founded Foresight Taiwan and the Germination Initiative and Functional Units with Dr. Eugene Wong in 2007-2012. He served as a reviewer and think tank member of the Germination Initiative ever since. He also served as an advisor to the Ministry of Education and its Division of IT education in 2012-2016. He served as Chair of the Office of Science & Technology Program Executive Review Board in 2017-2018. Dr. Ho's research interests cover the integration of theory and applications, including combinatorial optimization, information retrieval and extraction, multimedia network protocols, bioinformatics, and digital library and archive technologies. Dr. Ho also published results in the field of VLSI/CAD physical design.

Pulmonary Nodule Classification in Lung Cancer from 3D Thoracic CT Scans Using `fastai` and `MONAI`

Satheshkumar Kaliyugarasan[1,3**], Arvid Lundervold[2,3], Alexander Selvikvåg Lundervold[1,3 ** *]

[1] Dept. of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen (Norway)
[2] Dept. of Biomedicine, University of Bergen (Norway)
[3] Mohn Medical Imaging and Visualization Centre, Department of Radiology, Haukeland University Hospital, Bergen (Norway)

[**] These authors contributed equally to the current work

Received 9 November 2020 | Accepted 31 March 2021 | Published 4 May 2021

Abstract

We construct a convolutional neural network to classify pulmonary nodules as malignant or benign in the context of lung cancer. To construct and train our model, we use our novel extension of the `fastai` deep learning framework to 3D medical imaging tasks, combined with the `MONAI` deep learning library. We train and evaluate the model using a large, openly available data set of annotated thoracic CT scans. Our model achieves a nodule classification accuracy of 92.4% and a ROC AUC of 97% when compared to a "ground truth" based on multiple human raters subjective assessment of malignancy. We further evaluate our approach by predicting patient-level diagnoses of cancer, achieving a test set accuracy of 75%. This is higher than the 70% obtained by aggregating the human raters assessments. Class activation maps are applied to investigate the features used by our classifier, enabling a rudimentary level of explainability for what is otherwise close to "black box" predictions. As the classification of structures in chest CT scans is useful across a variety of diagnostic and prognostic tasks in radiology, our approach has broad applicability. As we aimed to construct a fully reproducible system that can be compared to new proposed methods and easily be adapted and extended, the full source code of our work is available at https://github.com/MMIV-ML/Lung-CT-fastai-2020.

Keywords

Convolutional Neural Networks, Fastai, Lung Cancer, Thoracic CT.

DOI: 10.9781/ijimai.2021.05.002

I. Introduction

Using convolutional neural networks is well-known to result in powerful tools to analyse medical images, across a variety of important applications [1], [2]. This approach to medical image analysis can lead to valuable insights and assistance in imaging diagnostics. The path from research to clinical practice is however slow and arduous, perhaps more so than is generally thought [2], [3]. But the number of software solutions on the market, with regulatory approval and aimed at diagnostic support, is growing, along with their adoption in hospital workflows.

In radiology, the computed tomography (CT) imaging modality is currently experiencing the highest impact of deep learning-based solutions. CT uses computer-processed combinations of many X-ray measurements taken from different angles to produce cross-sectional digital images (virtual slices) of specific regions or organs within the human body. This allows for non-invasive inspection of disease processes or lesions. Another prominent and widespread imaging modality is magnetic resonance imaging (MRI). It is based on quite different physical principles (nuclear spins in magnetic fields, spin excitation by application of radio-frequency pulses, magnetic resonance, and tissue specific and disease-related magnetization and relaxation phenomena) and enables exploitation of a large collection of measurement techniques and contrast mechanisms. Compared to CT, MRI examinations are generally more expensive, more time-consuming and less available. The signal properties are also more complex and typically multi-parametric, and proper interpretation puts high demands on radiologists' specialized training and experience. This partly explain why CT is more heavily used in daily routine radiology, and also why it is a popular target for the med-ical machine learning community [4].

Identifying and assessing structures in the lung from thoracic CT scans (chest CT) is a crucial task across multiple diseases involving the lungs and upper abdomen, e.g. lung cancer, chronic lung disease and pneumonia. Computer-aided diagnostic tools addressing chest CT is therefore an important area in medical imaging[1].

The diagnosis and follow-up of lung cancer patients using chest CT requires the identification of malignant tumors appearing as

* Corresponding author.
E-mail addresses: sathiesh.kumar.kaliyugarasan@hvl.no (S. Kaliyugarasan), arvid.lundervold@uib.no (A. Lundervold), allu@hvl.no (A.S. Lundervold).

[1] An area of particular relevance at the time of writing is the viral pneumonia caused by SARS-CoV-2 ([5], [6]).

pulmonary nodules (i.e. spots on the lungs). Distinguishing benign and malignant nodules is difficult, as the differences can be subtle and the malignancy potential is highly variable [7], but such assessment forms an important source of information for diagnosis and evaluation of progression and treatment responses. Indications of lung cancer can also appear as incidental findings on CT scans. As chest CT is widely used across a range of diseases and injuries, this represents an additional challenge for radiologists.

II. Related Work

Multiple studies have investigated how CNNs can be used in the context of lung cancer. Two recent and quite comprehensive reviews are [8], [9]. Below we highlight two illustrative examples of recent, related work.

In [10], the authors constructed an end-to-end system based on three 3D CNNs for the localization and categorization of lung cancer risk, using low-dose CT images as inputs. They achieved a test set ROC AUC of 94.4% using data from the National Lung Cancer Screening Trial (NLST), and a ROC AUC of 95.5% on an independent data set collected at Northwestern Medicine. A retrospective reader study was conducted, in which their model outperformed six experienced US board-certified radiologists. Their system had four main components: (i) a Mask R-CNN for instance segmentation used to produce lung segmentation masks; (ii) a 3D RetinaNet CNN trained to output ROIs around possible cancer lesions; (iii) a 3D version of Inception V1 trained to predict cancer diagnosis within one year directly from CT volumes; (iii) a CNN classifier trained on features extracted from the detected ROIs as well as features extracted from the volume model, outputting malignancy scores for each ROI. Their study was based on a combination of publicly available data from LUNA, LIDC and NLST, in combination with a large data set sourced from Northwestern Medicine that is not publicly available. The source code used in their work is not publicly available.

In [11], the authors construct *DeepLung*, a "cancer diagnosis system" based on two 3D CNNs that perform lung nodule detection and binary classification (benign vs. malign), respectively. For nodule detection they constructed a 3D Faster R-CNN with dual-path blocks and a similar encoder-decoder structure to the U-Net of [12], obtaining a FROC (Free Response Operating Characteristic) score of 84.2% on the LUNA16 data set [13] using a 10-fold patient-level cross-validation split. Their nodule classification model consisted of a 3D dual-path network extracting classification features, and a gradient boosting machine trained on the extracted features combined with raw nodule CT pixels and nodule size. They achieved a classification accuracy of 90.44%on the LIDC-IDRI data set using the same cross-validation approach as in LUNA16. The source code is available at https: //github.com/wentaozhu/DeepLung.

III. Main Contributions

Motivated by a lack of a common set of training data for machine learning models for lesion malignancy classification in the literature and what we see as important missing elements in how most CNNs for 3D medical imaging tasks are trained, our objectives are the following: (i) bring a set of techniques for training CNNs that have been shown to be highly impactful for 2D image classification to 3D by extending and incorporating ideas from the popular `fastai` library, and (ii) to provide a reproducible setup of data and model evaluation that can be used by other researchers aiming to train models to perform lung nodule classification. Our main contributions are:

1. We preprocessed and prepared the comparably large and well-annotated LIDC-IDRI data set (Section IV) for use in a binary malignancy prediction task, taking care to set aside a separate test set consisting of particularly well-characterized patients.
2. We constructed and trained a three-dimensional CNN using our novel extension to 3D of the `fastai` [14] deep learning library, combining it with features from `MONAI` (https://github.com/Project-MONAI/MONAI[2]), obtaining results comparable to the state-of-the-art in nodule classification and patient-level cancer diagnoses for the LIDC-IDRI data set.
3. We investigated the malignancy predictions by integrating a 3D version of gradient-weighted class activation mapping (Grad-CAM) [16] in our framework, enabling some element of *explainable AI* [17].
4. To ensure reproducibility and to ease further extensions or adaptions of our approach, we have made the source code openly available under a permissive open source license at https://github.com/MMIV-ML/Lung-CT-fastai-2020, in a tutorial-like Jupyter Notebook [18] that step through the process from data loading to result interpretations.

IV. Methods and Materials

A. Data Set

Using supervised learning with CNN models requires large amounts of labelled training data. For pulmonary nodule analysis, the data is typically obtained by manually labelling nodule locations and outlining lesions on CT images, a costly and hard to scale process hampered by intra- and inter-rater variability. Nevertheless, reasonably large annotated data sets with benign and malignant pulmonary nodules have been made openly available for researchers, reducing the entry price and increasing the pace of new research.

We used the Lung Image Database Consortium image collection (LIDC-IDRI), consisting of diagnostic and clinical lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions [19][3]. The images were extracted from the picture archiving and communication systems (PACS) of seven different institutions and anonymized in accordance with HIPAA guidelines. The data collection was approved by the local IRBs of the seven participating LIDC-IDRI institutions. To each image there is associated the results of a two-stage anno-tation process involving four experienced thoracic radiologists. First, in a blinded-read phase, each radiologist independently reviewed the CT scans, marking lesions belonging to one of three categories (*nodule ≥ 3 mm*, *nodule < 3 mm*, and *non-nodule ≥ 3 mm*), where the concept of "nodule" refers to a focal abnormality[4]. Then each radiologist (among a total of 12 radiologists coming from altogether five LIDC-IDRI institutions) assessed independently and subjectively each *nodule ≥ 3 mm* for characteristics such as subtlety, internal structure, spiculation, lobulation, shape (sphericity), solidity, margin, and likelihood of malignancy. Each such nodule, having (by its size) a greater probability of malignancy than lesions in the other two categories, was marked regardless of presumed histology, e.g. a primary lung cancer, metastatic disease, a noncancerous process, or indeterminate in nature.

By design, reader consistency studies are not possible with the LIDC-IDRI data set as the order of the readers varies from instance to instance. However, the marks from up to four readers for a given

[2] Originally, we developed our extension of `fastai` and `MONAI` for 3D MRI of the head, as a tool for the estimation of brain age from MRI recordings (unpublished work and [15]) indicating our framework's general utility.

[3] See also https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

[4] Some radiologists will argue that these three lesion categories could be somewhat artificial relative to clinical practice.

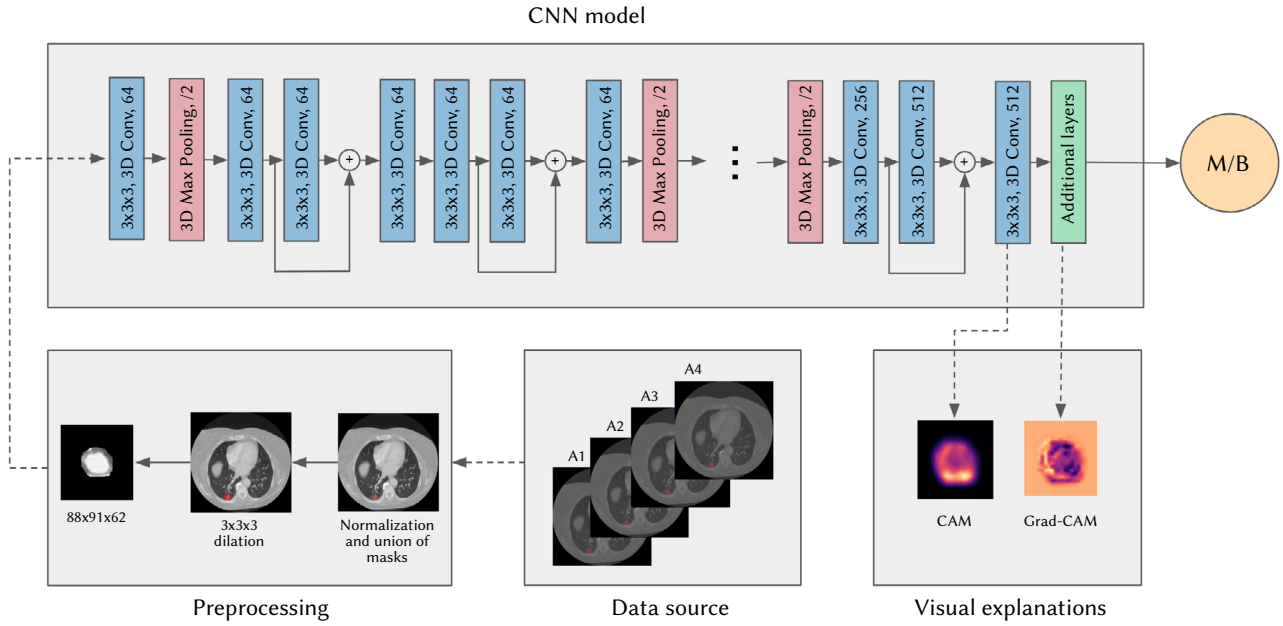


Fig. 1. The annotated images from our data source, LIDC-IDRI, are preprocessed by extracting 3D regions of interests around each of the nodules by taking the union of all the masks provided by expert annotations (e.g. A1−A4), before dilating the image slightly to capture some of the nodule surroundings. Using the expert assessment of malignancy, the resulting nodule images are used to train a 3D CNN model. This results in a nodule classification model with binary output: malignant (M) or benign (B). Our 3D implementation of class activation maps provides a visual explanation, here shown as a pair of 2D slices, indicating areas impacting our model's nodule classification decision. For further details, see the text and the accompanying code repository: https://github.com/MMIV-ML/Lung-CT-fastai-2020

lesion, using a five-point scale (a low score denoted likely benign nodule, a high score likely malignant), makes it possible to assess different degrees of reader agreement. Assessing inter-rater variability is very important to gauge the performance of systems aiming to automate the process. We therefore made an analysis of inter-rater variability regarding the "likelihood of malignancy" characteristic using the *Krippendorff's alpha* coefficient [20].

In our study, we have used a total of 2662 annotated nodules that were annotated as *nodule* $\geq$ *3 mm* by at least one radiologists, collected from clinical thoracic CT scans of 1018 patients in the LIDC-IDRI data set.

B. Preprocessing

The voxels in a 3D CT recording are displayed in terms of relative radiodensity. More specifically, the signal intensities or attenuations in CT are expressed in Hounsfield units (HU). This is based on a linear transformation of the original attenuation coefficients in which the radiodensity of distilled water has HU = 0 and the radiodensity of air is set to HU = −1000. According to this HU scale, lung parenchyma is in the range [−700, −600], fat is [−120, −90], lymph nodes [+10, +20], and blood [+13, +50], to mention a few relevant tissue types. In our CT data we considered voxels within a HU-range of [−1200, +600], and voxel values were normalized to the interval [0, 1] according to the transformation $x'' \mapsto x'' : x' = (x + 1200)/(1200 + 600)$; $x'' = 0$ if $x' < 0$, $x'' = 1$ if $x' > \ldots 1$, else $x'' = x'$.

For each CT scan of a subject, we collected all the radiologists segmentation masks. To ensure that we captured entire nodules we took the union of the masks. To make some of the surrounding context of each nodule available for the classification model, we dilated the resulting mask by adding 3 voxels to its boundary. The data set used to construct and evaluate our models was the constructed by applying the masks to the corresponding normalized CT and cropping to a cube containing the nodules. This gave us a total of 2662 3D images containing nodules. See Fig. 1 for an illustration of the preprocessing process.

We extracted each of the radiologists' subjective assessments of malignancy likelihood and computed the median scores across the readers for each nodule. If the median score for a nodule was < 3 we marked it as *benign*, if $> \ldots 3$ as *malignant*. The nodules with median score 3 (indeterminant) were dropped from our data set. This gave us a total of 1106 benign nodules and 525 malignant.

C. Our `fastai` Extension and the 3D CNN Architecture

Our work is based on a combination of the `MONAI` deep learning framework and our own extension of the powerful `fastai` library built on top of PyTorch [14]. We have added functionality to support the construction, training and evaluation of three-dimensional convolutional neural networks, tailored for medical imaging-specific problems and file formats. In short, we have extended `fastai` to support 2D and 3D MRI and CT images by constructing new data loaders and data augmentation capabilities, and enabled the use of custom 3D CNNs while still supporting the highly impactful training techniques of `fastai`. This includes the learning rate finder [21] to find the optimum learning rate and the one-cycle learning rate policy (i.e. specific learning rate changes during the training, related to the concept of super-convergence [22], [23]).

The architecture of our 3D CNN is shown in Fig. 1. Each convolutional layer in our network consists of $3 \times 3 \times 3$ convolutions, followed by a batch normalization layer [24] and a rectified linear unit (ReLU) layer [25]. We add residual connections after each second convolutional layer. Each down-sampling block has a two-stride $2 \times 2 \times 2$ max-pooling layer.

To enable *discriminative learning rates*, i.e. different learning rates for different parts of the network, we divide the network into two layer groups: convolutional layers and additional layers. This also allow us to do gradual unfreezing, and eases the potential re-use of trained weights from the early layers for other tasks (i.e. *transfer learning*).

D. Training and Evaulation

To evaluate and get a robust estimate of our model's performance, we selected all the subjects in the LIDC-IDRI data set that have corresponding patient-level diagnoses as our test set (99 subjects, 238 nodules). The remaining data were divided into a training set (526 subjects, 1140 nodules) and a validation set (90 subjects, 255 nodules), using stratified sampling and no patient overlap between the sets. In order to deal with imbalanced classes in the training set (802 benign, 338 malignant), we over-sampled the malignant class by duplicating each sample.

Before feeding the images into the network, each image was padded to have the same volume dimension as the largest volume data × a scaling factor. We used data parallelism to train our model on four NVIDIA Tesla V100 32GB GPUs. Our training process was composed of two phases:

- Training a model on $44 \times 46 \times 31$ volumes, with weights randomly initialized (He initialization [26]).
- Training a final model on $88 \times 91 \times 62$ volumes, with weights initialized by copying the weights of the previous model.

This approach is known as progressive image resizing [27], a technique used to both reduce training time and to increase model performance. In our case, we found that it improved the accuracy on the validation set by almost two percentage points.

Our model was trained end-to-end in mixed precision [28] using the Adam optimizer [29]. The base value for the cyclic learning rate in the final model was set to 6×10^{-4} for frozen layers and 5×10^{-5} after unfreezing the layers, with learning rates for earlier layers scaled down by a factor of 20. We trained the model using a batch size of 128. For data augmentation we used random scaling with a factor from 1.0 to 1.1 and random rotation by an angle in the range [-35, 35]. As the geometry of the nodules can contain information about their malignancy, we only used shape-preserving morphisms. For regularization, we used a weight decay rate of 0.01 and a dropout ratio of 0.4, selected based on the performance on the validation data. Our final model was trained on the combined training and validation data for a few epochs, with a small cyclic learning rate, to also make use of the information contained in the validation data and its labels during model training.

E. Explainable AI and Class Activation Maps

As deep learning models are highly complex hierarchical objects with enormous amounts of parameters, there is an inherent "black-boxiness" to them. As they are increasingly being implemented across the medical imaging and decision making domains, this raises both technical challenges (how to open the black box?) and ethical conundrums (when is it OK to use predictions you cannot fully understand?). Using our extension of `fastai` we can produce what are called class activation maps (CAM) [30] and gradient-weighted class activation mapping (Grad-CAM) [16]. These are heat maps that can be used to indicate the importance of regions of an image for the model's classification, providing a relatively simple way to gain some explainability for image classification models, and potentially also to gain useful insights into the data used to construct the model.

CAM generates heat maps from the adaptive pooling layer, where the average of each cell across every channel is calculated. On the other hand, Grad-CAM uses the gradient information flowing into the last convolutional layer to produce heat maps, making it applicable to any CNN architecture.

A problem with these methods is that the resolution of the heat maps are the same size as the final convolutional layer. This means that we have to upsample them to the same size as the input images to highlight class-specific image regions. To mitigate this problem one can remove the pooling layers, but this will require more computational power due to larger spatial dimensions. In addition, overfitting is more likely to occur, which might reduce the performance of the network.

V. Experimental Results

Our test set consisted of 238 nodules from 99 subjects, 146 benign and 92 malignant. There were no overlap among train and test subjects. In addition to predicting nodule malignancy, we further investigated the models predictive capabilities by using the ground truth labels of patient diagnosis available in the LIDC-IDRI data set. The 99 patients in our test set were all diagnosed as either *malignant* or having *benign or non-malignant disease*. If one or more nodules from a patient was predicted to be *malignant*, we predicted malignant, else *benign or non-malignant disease*.

The results are displayed in Table I, Fig. 2 and Fig. 3.

TABLE I. Performance Metrics of Our Binary Classifier Predicting Single Nodules (N=238) and Patient Cases (N=99) in the Test Data Set: Accuracy (ACC), Precision (PREC) and Recall (REC). For the Patient Predictions We Give Performance Values Separately for Those Obtained by Our Model (CNN) and for Those Obtained by the Median Radiologist Assessments (Rad)

Classification task						
Nodule classification (%)			Patient classification (%)			
ACC	PREC	REC	Source	ACC	PREC	REC
92.4	85.6	96.7	CNN	75	86.8	78.7
			Rad	70	88.1	69.3

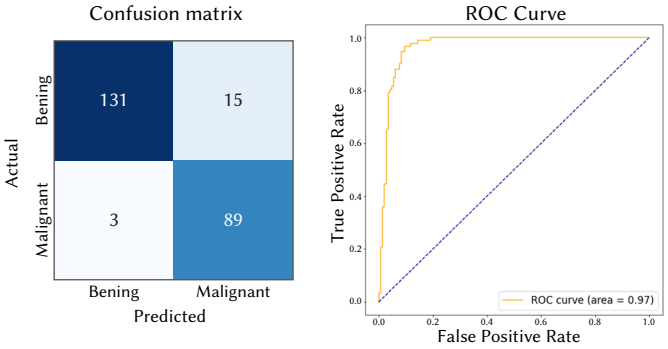


Fig. 2. Predicting the "likelihood of malignancy" in the test set of 238 nodules. (a) Confusion matrix. (b) ROC curve.

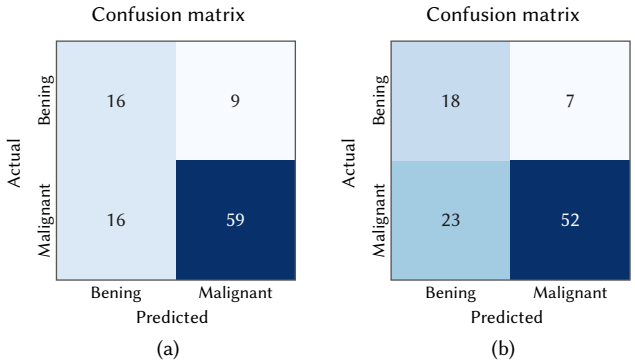


Fig. 3. Confusion matrices: (a) for the CNN predictions, (b) for the median malignancy scores by the radiologists. Note the additional cancer diagnoses captured by our CNN.

The mean score assigned to each nodule classified correctly as benign was 1.91 (SD 0.56) and as malignant 4.18 (SD 0.56). The nodules

misclassified as benign had a mean score of 3.5 (SD 0.0) and those misclassified as malignant had a mean score of 2.23 (SD 0.4).

To assess the inter-rater variability and how the model compares to the human raters, we calculated the *Krippendorff's alpha* coefficient [20] for the 238 nodules. Krippendorff's alpha applies to any measurement level, can handle various number of raters and is invariant to the permutation and selective participation of raters. It also ignores missing data entirely. The independent and interchangeable rater panel per unit consisted of one to five radiologists using scores $s \in \{1$ (*most likely benign*), $2, \ldots, 5$ (*most likely malignant*)$\}$.[5] We note that the agreement on these subjective assessments were not very high. For the Krippendor's $\alpha \in [0, 1]$, $\alpha = 0$ is absence of agreement, and $\alpha = 1$ is perfect agreement. For the "likelihood of malignancy" we found Krippendorff's $\alpha = 0.49$, $\mathrm{CI}_{.025,.975} = [0.43, 0.54]$ (obtained by bootstrapping), indicating poor agreement among the raters.

The Krippendorff's alpha coefficient (in this case equivalent to Cohen's Kappa score) comparing the model's rating to the ground truth (determined by the median radiologist rating) was 0.84, $\mathrm{CI}_{.025,.975} = [0.78, 0.91]$.

The Krippendorff's alpha of the binary assessments of malignancy among the radiologists was $\alpha = 0.58$. By including the independent, CNN-based rater we obtained an increased alpha score to 0.68, indicating the usefulness of including this rater in the assessment of each nodule.

We applied our class-activation map approach described in Section IV.E to a selection of test nodules and CNN predictions. In general, getting better insight into CNN behavior and model predictions, both in cases where it classifies correctly and in cases where it fails, is of interest for several reasons. The class activation maps can provide discriminative information in image regions or part of the lesion being used by the model to predict the class label for the particular instance. This ability can at best introduce interpretability and trust in the model, or facilitate exploration and discovery of new features (image biomarkers) that might have a mechanistic relation to the disease process or disease state. In the present study, we did not fully explore the CAM approach or its potential by involving radiologists or pathologists, and the CAM results are anecdotal and not rigorously validated.

Some of the generated heat maps from our CNN model are presented in Fig. 4. By examining the malignant nodules (nodule 1 and nodule 2) and their corresponding heat maps, we can see that the lesion brims are highlighted, indicating that these regions are most important for the predictions. This might reflect typical malignant tumor growth characterized by central necrosis and viable tumor cells in a well-vascularized periphery. Another interesting finding was nodule 4, a nodule rated benign but classified as malignant by our model. This nodule was assessed by two radiologists deciding malignancy likelihood 2 and 3, respectively (i.e. towards benign), whereas the biopsy done on this nodule concluded that it was a malignant primary lung tumor.

VI. Discussion and Perspectives

We have addressed an important field of oncological radiology: the use of 3D CT scans to characterize focal lung lesions as benign or malignant. Using a large multi-center collection of well-organized CT examinations we constructed and trained a 3D CNN model to perform nodule malignancy classification.

[5] The "likelihood of malignancy" characteristic is particularly subjective since the radiologists were not provided with any clinical information about the patients. As a general scaling guide, the likelihood of malignancy was rated under the assumption that the lesion was associated with a 60-year-old male smoker.

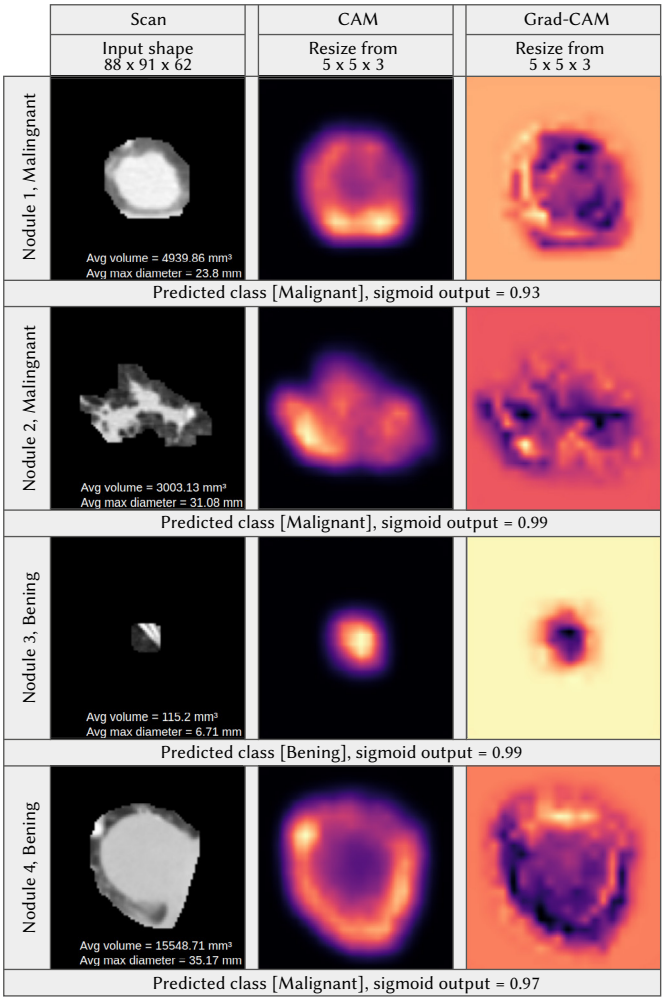


Fig. 4. Examples of CAMs and Grad-CAMs for our model and the corresponding predictions and sigmoid outputs for the respective classes on a selection of four test set nodules.

Because CNNs automatically extract features from data, both interpretation and troubleshooting are more difficult compared to traditional machine learning models. For domains like medical diagnosis, where decision confidence is crucial, it is important to make sure that the results make sense. Otherwise, these models can easily end up performing worse than expected when used for real-world decision making. CAMs and Grad-CAMs generated from CNN models can be valuable for developers to gain some visual insights into models decision processes, helpful to identify data leakage, structural bias and for more comprehensive performance evaluation. In addition, the heat maps have the potential to detect local features that can be used as a biomarker for identifying malignant nodules. We implemented and explored these simple "explainable AI" techniques, assessing successful and unsuccessful nodule predictions.

Our model had a test set accuracy of 92.4% on the per-nodule malignancy classification task. On the patient-level malignancy classification task, our model had an accuracy of 75%. This gave an indication of the network's ability to pick up patterns corresponding to real nodule malignancy. As shown in Fig. 4, class activation maps can highlight regions of particular relevance for the nodule classifications, further indicating that the reasonableness of the features picked up by our CNN model.

In further work we will use the present system as a component in a detection + classification framework, obviating the need for manual annotation steps. We will test the system in the established

radiology research workflow at our hospital, through our "research PACS and RIS" system, enabling us to run arbitrary algorithms on locally recorded images. Such real-world testing is crucial to uncover and surmount the many technical obstacles faced when attempting to bring deep learning-based systems into practice [3]. Especially as it facilitates prospective investigations of the effect of combining the algorithm's predictions with radiologists' expertise, arguably the most interesting next step for research into applications of deep learning in medicine.

Acknowledgment

This work was supported by the Trond Mohn Research Foundation, grant number BFS2018TMT07.

References

[1] A. S. Lundervold, A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[2] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The lancet digital health*, vol. 1, no. 6, pp. e271–e297, 2019.

[3] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. Ioannidis, G. S. Collins, M. Maruthappu, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 368, 2020.

[4] M. Brown, P. Browning, M. W. Wahi-Anwar, M. Murphy, J. Delgado, H. Greenspan, F. Abtin, S. Ghahremani, N. Yagh-mai, I. da Costa, *et al.*, "Integration of chest ct cad into the clinical workflow and impact on radiologist efficiency," *Academic radiology*, vol. 26, no. 5, pp. 626–631, 2019.

[5] C. Bao, X. Liu, Z. H., Y. Li, J. Liu, "Coronavirus Disease 2019 (COVID-19) CT Findings: A Systematic Review and Meta-analysis," *J Am Coll Radiol*, vol. Mar 25, pp. 1–9, 2020.

[6] G. D. Rubin, C. J. Ryerson, L. B. Haramati, N. Sverzellati, P. Kanne, S. Raoof, N. W. Schluger, A. Volpi, J.-J. Yim, B. Martin, *et al.*, "The role of chest imaging in patient management during the covid-19 pandemic: a multinational consensus statement from the fleischner society," *Chest*, vol. 158, no. 1, pp. 106–116, 2020.

[7] P. de Groot, B. Carter, G. F. Abbott, C. C. Wu, "Pitfalls in chest radiographic interpretation: blind spots," in *Seminars in roentgenology*, vol. 50, 2015, pp. 197–209, WB Saunders Ltd.

[8] D. Li, B. Mikela Vilmun, J. Frederik Carlsen, E. Albrecht-Beste, C. Ammitzbøl Lauridsen, M. Bachmann Nielsen, Lindskov Hansen, "The performance of deep learning algorithms on automatic pulmonary nodule detection and classification tested on different datasets that are not derived from LIDC-IDRI: a systematic review," *Diagnostics*, vol. 9, no. 4, p. 207, 2019.

[9] A. Halder, D. Dey, A. K. Sadhu, "Lung Nodule Detection from Feature Engineering to Deep Learning in Thoracic CT Images: a Comprehensive Review," *Journal of Digital Imaging*, pp. 1–23, 2020.

[10] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[11] W. Zhu, C. Liu, W. Fan, X. Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 673–681, IEEE.

[12] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241, Springer.

[13] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge," *Medical image analysis*, vol. 42, pp. 1–13, 2017.

[14] J. Howard, S. Gugger, "fastai: A Layered API for Deep Learning," *Information*, vol. 11, no. 2, p. 108, 2020.

[15] S. Kaliyugarasan, A. Lundervold, A. Lundervold, *et al.*, "Brain age versus chronological age: A large scale mri and deep learning investigation," 2020, European Congress of Radiology-ECR 2020.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[17] D. Gunning, "Explainable Artificial Intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, 2017.

[18] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, *et al.*, "Jupyter Notebooks — a publishing format for reproducible computational workflows," *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, p. 87, 2016.

[19] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[20] K. Krippendorff, "Reliability in Content Analysis: Some Common Misconceptions and Recommendations," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004.

[21] L. N. Smith, "No more pesky learning rate guessing games," *CoRR, abs/1506.01186*, vol. 5, 2015.

[22] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[23] L. N. Smith, N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, 2019, p. 1100612, International Society for Optics and Photonics.

[24] S. Ioffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[25] V. Nair, G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[26] K. He, X. Zhang, S. Ren, J. Sun, "Delving Deep Into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[27] T. Karras, T. Aila, S. Laine, J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[28] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, *et al.*, "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.

[29] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

Satheshkumar Kaliyugarasan

Satheshkumar Kaliyugarasan is a doctoral researcher at the Mohn Medical Imaging and Visualization Center focusing on machine learning in radiological imaging. In 2019 he completed his MSc degree in soft-ware engineering at the Western Norway University of Applied Sciences, Norway.

Arvid Lundervold

Arvid Lundervold is a professor of medical information technology at the University of Bergen and head of the Neuroinformatics and Image Analysis Laboratory in the Neural Networks Research Group, and co-leader of the Computational Medical Imaging and Machine Learning Group at the MMIV center. His research interests are image processing and pattern recogni tion, functional imaging, image registration, quantification and visualization, and mathematical modeling. Lundervold received an MD from the University of Oslo and a PhD in physiology from the University of Bergen.

Alexander S. Lundervold

A.S. Lundervold has a PhD in mathematics from the University of Bergen, Norway. He's currently working as an associate professor at the Western Norway University of Applied Sciences, and as a senior data scientist at the Dept. of radiology, Haukeland University Hospital, Norway. He leads the Computational Medical Imaging and Machine Learning Group at MMIV, together with A.L. His expertise lies in medical data analysis, with a particular focus on medical image processing and applications of machine learning to medicine.

Modeling of Performance Creative Evaluation Driven by Multimodal Affective Data

Yufeng Wu[1], Longfei Zhang[1]*, Gangyi Ding[1], Tong Xue[1], Fuquan Zhang[2]

[1] Key Laboratory of Digital Performance and Simulation Technology, Beijing Institute of Technology, Beijing,100081 (China)
[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control Minjiang University (China)

Received 10 October 2020 | Accepted 2 July 2021 | Published 4 August 2021

Abstract

Performance creative evaluation can be achieved through affective data, and the use of affective features to evaluate performance creative is a new research trend. This paper proposes a "Performance Creative—Multimodal Affective (PC-MulAff)" model based on the multimodal affective features for performance creative evaluation. The multimedia data acquisition equipment is used to collect the physiological data of the audience, including the multimodal affective data such as the facial expression, heart rate and eye movement. Calculate affective features of multimodal data combined with director annotation, and defined "Performance Creative—Affective Acceptance (PC-Acc)" based on multimodal affective features to evaluate the quality of performance creative. This paper verifies the PC-MulAff model on different performance data sets. The experimental results show that the PC-MulAff model shows high evaluation quality in different performance forms. In the creative evaluation of dance performance, the accuracy of the model is 7.44% and 13.95% higher than that of the single textual and single video evaluation.

Keywords

Performance Creative Evaluation, Multimodal Affective Feature, Multimedia Acquisition, Data-driven, Affective Acceptance.

DOI: 10.9781/ijimai.2021.08.005

I. Introduction

THE cultural industry is prospering, and the performing arts, as an important branch of the cultural industry, highlights the core aesthetic values Lee [1]. Performance evaluation research has very important academic and application value, which helps to promote people's cognition and exploration of performing arts. After a long period of development, performing arts have evolved into many different forms of performance, such as stage performance, dramatic performance, large-scale event performance (opening and closing ceremonies of sports games, etc.) and multi-media interactive performance. Performance creativity includes the director or choreographer's novel ideas and clever designs in stage design, content structure, audience interaction, and the use of multimedia technology. A good performance creative can not only improve the performance quality, but also go deep into the hearts of the audience and affect the emotional and aesthetic cognition of the audience. So, how do we determine whether the creative of a performance is success? At present, the performance creative evaluation is mainly guided by professional aesthetic experience. In this context, it is difficult to make a true and objective evaluation of the performance creative, and it even brings a negative impact on the performance itself. In the field of performing arts, performance creative evaluation has begun to show its research value and practical significance.

* Corresponding author.
E-mail address: longfeizhang@bit.edu.cn

Performance creativity evaluation has its particularity, which is mainly reflected in the three types of people involved in the performance: director, actor and audience. From the perspective of time and space, performances can be divided into linear performances and non-linear performances. Linear performances mainly refer to film, television, media and other linearly edited video content. Non-linear performances refer to performances dominated by live performances, such as stage performances and square performances. Performance creators can design performance content through stage design, scenery, costume props and other performance elements. The actors will be affected by the performance environment and audience feedback, and even the possibility of on-site improvisation may occur. This is also the essential difference from linear performance, and it is also the main elements that affect performance creative. The performance creative evaluation proposed in this paper is mainly for non-linear performance.

In recent years, with the improvement of computer hardware and computing power, powerful and automatic feature extraction capabilities can effectively determine better features, thereby improving the efficiency of the model recognition system. Significant progress has been made in the research of computer vision [2]-[4], speech recognition [5]-[6], and natural language processing [7]-[9]. For example, an image recognition system based on deep learning has been described by [2], which uses various image preprocessing algorithms to perform grayscale processing and banalization on training data to enhance the training data, and then uses the GoogLeNet model to train these preprocessed images and test its recognition result. The method proposed in this study provides a new idea for future medical image-

assisted diagnosis. An attention segmental recurrent neural network (ASRNN) that relies on a hierarchical attention neural semi-Markov conditional random fields (semi-CRF) model has been proposed by [7] for the task of sequence labeling.

The improvement of computing power has promoted the arrival of the era of sensing intelligence [10]. Digital audio-visual technology, video stage design, and interactive performance content provide performance creators with broad ideas and promote the diversified development of performance creative. Using digital and intelligent methods to evaluate performance creative is the current main research direction. Performance evaluation uses these digital methods to promote the realization of new evaluation ideas and methods. Cross-topic research integrating machine learning, affective computing, artificial intelligence and other fields is gradually forming. In the complex performance environment, many factors such as performance behavior, stage setting, viewing angle, audience perception, etc. have had a great impact on performance evaluation. Only considering a single modal feature as a research object cannot meet the needs of performance evaluation. The use of multi-modal data features for modeling to solve complex systemic problems such as performance creative has become the focus of attention of researchers.

In order to solve the above-mentioned problems in the evaluation of performance creative, this article starts from the perspective of intelligent multimedia analysis, and integrates the physiological signal data such as audience evaluation text, audience facial expressions, audience heart rate, and eye movement data, to extract multimodal affective features and to evaluate performance creative. This paper argues that the core issue of performance creativity evaluation is to obtain the true emotions of performance works and the true evaluation of performance content, which poses new challenges to the methods and means of performance creativity evaluation.

The contributions of this work are presented as follows:

- This article proposes a "Performance Creative— Affective Acceptance (PC-Acc)" to evaluate the quality of performance creative, trains and builds a "Performance Creative—Multimodal Affective (PC-MulAff)" model, which can evaluate creative for different performance forms.
- This paper proposes a new "Performance Creativity-Multimodal Evaluation Data Set", which is composed of performance video data, audience evaluation text data and audience physiological data, which makes up for the problem of insufficient description of affective features by a single data type.
- Based on the establishment of a multi-modal evaluation data set, the correlation analysis between the audience's multi-modal physiological signals and the emotional dimension of the "Director Label" is realized. This work plays a decisive role in the evaluation of performance creativity.

The structure of this document is as follows: section II reviews the related works, section III details the methods proposed in our research, section IV presents the experiment results, section V details the discussion, and section VI conclusions and look forward to the future work.

II. Related Works

A. Performance Creative Evaluation (PCE)

In 1994, Abbé Decarroux put forward the importance of performing arts quality assessment and the challenges of analyzing it [11]. Frieder [12] proposed the concept of introducing computer technology into the field of fine arts, transforming the creative process into a computable process as early as 2007. **Evaluation from the perspective of the creative process:** Yamada [13] proposed the use of weighted sum of wavelet coefficients to generate creative dance movements, although implemented on the algorithm, it is difficult to describe an action that does not exist in words, and it is difficult to empathize with the result of this creative at the emotional level. Chang [14] proposed a calculation on integrating creative operation into music creation process. Although calculations and extensions have been made on musical characteristics such as timbre, there is still the problem of how to evaluate the calculated results. Gove [15] proposed that the coordination and precise interaction of players in musical performance should be regarded as the main constituent parameters of the creative process.

Evaluation from the perspective of the creative methods: Cabral [16] used interactive digital media to visualize the ideas of choreography, this paper proposed to use video annotator to annotate, analyze and evaluate the creative process of dance performance. However, this kind of interactive method will decompose the actors' attention and produce irreparable interference to the performance taking place. Therefore, the accuracy of the evaluation conclusion obtained from this method is difficult to be guaranteed. Cisneros [17] put forward the creative potential of VR and how to provide creative process for choreographers and dancers. This kind of creative method relying on new interactive tools may have a certain impact on the surface form, but this article believed that this has little ability to solve the structural problems of artistic creative and the problem of creative evaluation.

Evaluation from the perspective of the creative cognition: Christensen [18] analyzed the creative process and the creative evaluation process from the cognitive dimension. Pegah [19] proposed and evaluated a new method for stimulating creative in a common design system. The semantic similarity of creative expression and the structural similarity of hand-drawn sketch were calculated respectively. The creative level of the volunteers ware divided by the three intervals of high similarity, existing similarity and low similarity, so as to stimulate and transform the creativity of the volunteers, but the thresholds of these three intervals were not quantified and set. Tiffany [20] and Richardson [21] put forward the idea of solving the creative strategy and evaluation problems from the perspective of cognition, but it also needs to solve the quantitative analysis and modeling of the attention of creative works, which cannot be well realized in a short time.

Evaluation from the perspective of the creative modeling: Kyu [22] put forward the Computational Thinking Pattern Analysis (CTPA). It used CTPA to calculate differences in three different learning conditions as an index for measuring creativity. The differences were calculated by CTPA, and the creativity itself was mapped to 9 kinds of high-dimensional cosine space as divergent elements, which were used as indicators to measure the creativity. However, these indicators were only tested on the creative ideas of the two game designs, and there was no reasonable explanation for the on-site or off-site creative behaviors. Ajit [23] constructed a novel computing model combining visual and conceptual features to quantitatively represent and analyze feedback on creativity. Jonas [24] proposed a strategy of evaluating ideas through crowdsourcing feedback. Although directors can obtain feedback information from online audience groups, the behavior of group feedback is easy to have a guiding influence on individual subjective judgment.

Evaluation from the perspective of the management: Abfalter [25] recognized the importance of leading creative teams and creative environments in the performing arts context. The influence of leadership and organizational structure on performance and evaluation is elaborated, which can be regarded as a new perspective to solve problems, but there is no practical verification.

Above, we can see that the researchers to try various **perspectives (creative process, creative methods**, **creative cognition**, **creative modeling**, **management)** to evaluate the performance creative. However, the existing methods of performance creative evaluation have subjective limitations, and the feedback strategy of evaluation is complicated, which have different effects on the accuracy and effectiveness of the digital evaluation of creative performance.

B. Multimedia Computing in PCE

From the perspective of sentiment analysis and machine learning, the core problem of performance evaluation is to solve audience emotion detection and how to build an evaluation model.

1. Affective Computing Based on Semantic Features

It is a common research method to obtain and analyze audience emotion through semantic network. The Lee [26] proposed a solution to build big data for dance performance and develop a big data creative analysis model system suitable for dance research, 25 kinds of high-frequency words were classified according to dance theme, characters and themes. Min [27] analyzed 20,776 dance research data texts accumulated in South Korea from 1958 to 2016 by using text mining and degree concentration based on semantic network, and obtained the core creative ideas and emotional themes of Korean dance performances in different historical periods. Ryeon [28] determined the determinants of Korean dance performance through tree analysis based on data mining. Choi [29] aimed to systematically investigate the knowledge structure of modern dance research through text mining, and establish the emotional cognitive system of modern dance performance in the future. Kyung [30] Choihyojin [31]and Kimhayeon [32] analyzed the thematic emotional trend of Korean dance performances in the past 20 to 30 years through text mining. Zhou [33] proposed a method to integrate acoustic features and text features to calculate emotions from large-scale Internet voice data. Liang [34] proposed a UAM proposed a universal affective Model (UAM) to calculate the potential emotion of short text in social media. Hung [35] introduced and discussed the classification methods for MuSe-Topic sub-challenges, as well as the data and results. For topic classification, Hung integrated two language models, ALBERT and RoBERTa, to predict 10 topic categories. In order to classify valence and arousal, SVM and random forest are combined with feature selection to enhance performance.

Although this method can quickly establish the semantic network of the text, it cannot guarantee whether the semantic of the text truly reflects the psychological and emotional of the audience. This method relies heavily on the performance cognition of the audience.

2. Affective Computing Based on Physiological Signal Features

Sowden [36] analyzed the influence of different emotions on creative. Corness [37] described the research of extracting the audience's experience in the performance scene to evaluate the audience's empathy experience in the performance process. Coursaris [38] developed and tested a cognitive model of cognitive user satisfaction with high explanatory power, which was used to assess the direct impact of cognitive and emotional dimensions on satisfaction. Altuwairqi [39] proposed an emotional model and a new process to test students' engagement in learning, six core silver factors affecting the model were analyzed by statistical methods (strong, high, medium, low, disengagement). Rahdari [40] introduced a multi-modal emotion recognition system based on two different modalities, namely emotional speech and facial expressions. For emotional speech, common low-level descriptors including prosodic and spectral audio features are extracted. Loprinzi [41] proposed a cognitive emotional model to assess physical activity based on experience, to evaluate the exercise habit and intensity of adults. The emotional changes can be seen after the acute exercise, it is related on the exercise intensity and exercise cycle of the volunteers. At the same time, personal health will also have an impact on emotional changes, which shows that this measurement method has inevitable flaws in its universality. A method of using emotional calculation to evaluate football players is proposed by Liu [42], which combines the text information of the post-match report and emotional calculation to measure the performance quality of the players. The author established a player performance evaluation model based on LSTM, However, this method still has some problems to be solved. For example, it is difficult to achieve accurate statistics and quantification of specific behaviors in the game, which has a greater impact on the collection and evaluation of key information. Wei [43] proposed a new method for extracting emotional features from facial expression images using multi-modal strategies is proposed. The basic idea is to combine low-level experience features and high-level self-learning features into multi-modal features. The convolutional neural network was used to extract the two-dimensional coordinates of the key points of the face as low-level experience features, and the convolutional neural network is used to extract high-level self-learning features. In order to reduce the free parameters of CNNs, small filters are used in all convolutional layers. Chen [44] used sLORETA to analyze the significant difference between the active source area and frequency band of the EEG reconstruction source based on emotion, and selected 26 Brodmann regions as regions of interest (ROI). On this basis, the support vector machine was used to extract the time-frequency domain features of 6 important activity regions and frequency bands, and to classify different emotions. Choi [45] proposed an emotional response generation model based on emotional feature extraction is proposed. Deepika [46] identified three bases for speech emotion recognition system: database, feature extraction and various classification methods. The performance of speech emotion recognition system was discussed. Features were divided into basic, prosodic and spectral features. Wei [47] proposed a new method for extracting emotional features from facial expression images using multi-modal strategies. The basic idea was to combine low-level experience features and high-level self-learning features into multi-modal features. The convolutional neural network was used to extract the two-dimensional coordinates of the key points of the face as low-level experience features, and the convolutional neural network was used to extract high-level self-learning features.

It can be seen that although researchers have proposed various affective models, there are still many problems in the data collection and input of the models. It is a relatively superior research method to establish an emotional evaluation mechanism by obtaining the physiological information and cognitive status of the audience. Usually this method is called implicit sentiment measurement. Radbourne [48] proposed to monitor the emotional experience of the audience in live performances, and emphasized the importance of interaction firstly. Radbourne [49] proposed that the ability to stimulate the audience's emotions is the key to accurate performance evaluation, and the role of the audience in the performance is gradually changing. In 2011, Latulipe [50] used GSR to monitor the emotional arousal of 6 spectators in dance performances. The research divided the performance attributes into two categories: LH scale and ER scale, which represent the audience's degree of affection for the performance and the degree of arousal of the audience by the performance content. The results showed that the GSR value is positively correlated with the degree of ER arousal. Wang [51] used GSR signals to monitor the emotional state of 15 volunteers in live performances. The research conducted a cluster analysis on the audiences through the GSR values, and found that the data of 10 volunteers were closely related. Martella [52] analyzed the audience's feedback in live performances, and used a three-axis accelerometer to calculate the acceleration of the audience's body movements, which integrated complex emotional experiences

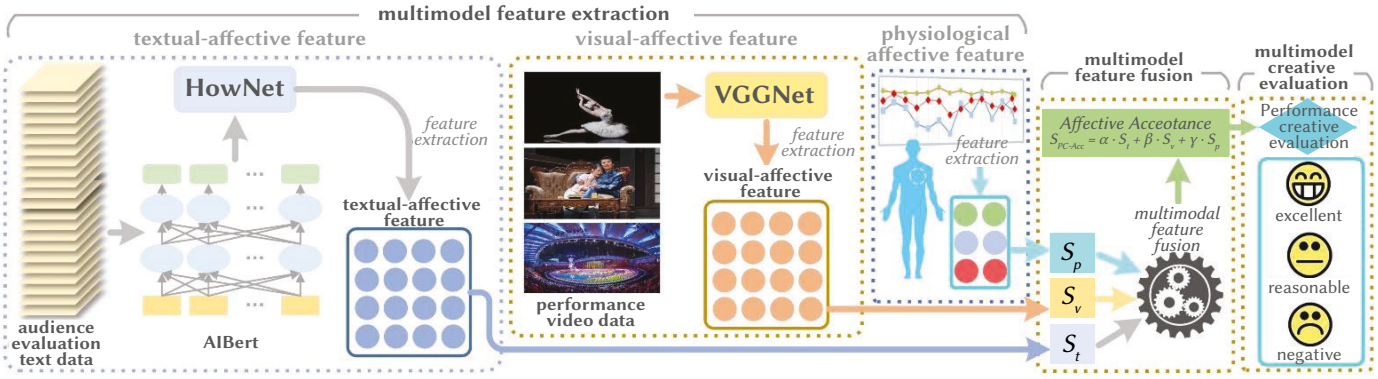


Fig. 1. Research framework.

such as "Enjoyment" or "Immersion" for quantification. By calculating the dynamic changes of acceleration, the research has reached nearly 90% accuracy in predicting whether the audience is in a "enjoying state". Psychological research has confirmed that the relationship between emotion and EEG signals has an important impact on human cognitive processes [53] [54].

It has confirmed that when users watch highly aroused content, the EEG coherence between the hemispheres has increased significantly [55]. The interaction of paired EEG electrode amplitudes has been confirmed to be correlated with positive arousal in emotional states [56]. According to the four-quadrant theory of emotions proposed by Koelstra [57], 32 performance emotion-inducing materials have been divided into "high arousal high valence" (HAHV), "high arousal low valence" (HALV), and "low arousal high valence" (LAHV) and "Low Excitation Low Valence" (LALV) four emotional level categories. Pinto [58] processed electrocardiogram, electromyography and dermal electrical activity to find a physiological model of emotion. Using samples of 55 healthy subjects, Pinto used single-peak and multi-peak methods to analyze which signals or combinations of signals can better describe emotional responses.

Zhang [59] proposed a multi-modal emotion recognition method using deep autoencoders for facial expressions and EEG interactions. The decision tree is used as the target feature selection method. Then, based on the facial expression features recognized by the sparse representation, the solution vector coefficients are analyzed to determine the facial expression category of the test sample. After that, the bimodal depth autoencoder was used to fuse EEG signals and facial expression signals.

We can clearly see that the use of human physiological data characteristics to calculate audience emotions, thereby realizing the evaluation of performing arts, has become a new research focus. Based on the affective computing method of physiological signals, this paper designed and collected the physiological signal data of the audience while watching the performance. For the first time, the method of "Director Label" was used to label the performance videos and physiological signals. Based on the above mentioned, "Performance Creative - Multimodal Evaluation Dataset" was developed for performance evaluation.

III. Method

A. Research Framework

The architecture of the PC-MulAff model is proposed in this paper, as in Fig. 1. The model consists of three parts: multimodal feature extraction module, multimodal feature fusion and multimodal creative evaluation. The multimodal feature extraction includes textual-affective feature, visual-affective feature and physiological affective feature unit. First of all, the audience evaluation text data in textual-affective feature unit get score S_t, and then, the performance video data get score S_v, and similarly the physiological data get score S_p. Finally, the S_{PC-Acc} of was calculated, and the performance creativity was evaluated according to the value of $S_{PC\text{-}Acc}$.

B. Multimodal Feature Extraction

This paper extracts the affective features of the audience and the visual affective features of the performance video, and conducts analysis and evaluation of performance creative through the quantified scores after the fusion of multimodal features.

1. Textual-Affective Feature Extraction

In the depth of the traditional learning method, usually using Word2Vec extracted feature, such as the Skip-gram, Continuous Bag of Words (CBOW), but they catch the embedded part of the training, this part of the parameter is less, if continue to downstream text processing tasks you will need to add a lot of parameters, and from the training, increase a lot of training data and the training time, and Word2Vec emotional analysis of the context is limited by the length of the context, with the result of the classification of emotional impact. In order to reduce training data to complete the downstream tasks of natural language processing, researchers began to learn the embedding of general text through a large corpus: two-way LSTM M. Peters [60] combined embedding for forward and backward propagation, and BERT using Transformer model for two-way encoding, decoding and pre-training A.Radford [61] J. Devlin [62]. AlBert Lan [63], which adopted the full-network pre-training algorithm, adopted the parameter sharing mechanism, which could share most of the parameters with subsequent processing tasks, not only saving calculation time, but also avoiding the problem of limited context length. Therefore, in this study, AlBert is selected to extract the emotional features of the text.

$$Attention(Q,\ K,\ V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

The core algorithm of AlBert's feature extraction is Transformer. Transformer USES multi-head self-attention mechanism. It projects H different Queries(Q), Keys(K) and Values(V) respectively, and the mapped dimensions are ALL D_k and D_v as the Equation.1 Ashish [64].

2. Visual-Affective Feature Extraction

In the classification of visual emotions, the description of features is particularly important. A new spatial representation is proposed to solve the limitation of image compression domain, which provides a fast quasi-spatial transformation and effectively integrates the histograph-based method to solve the shortcomings of image enhancement in compression domain [65]. This method not only

achieves excellent visual quality, but also provides new possibilities in the acquisition of image features. Zhao [66] proposed Principles of Arts-based emotion features, and Xu [67] proposed three emotions for deep convolutional neural network to predict images: Positive emotion, negative emotion and neutral emotion, Jou [68] cross residual neural network is proposed for image emotion classification, Gao [69] visual attention and circulation is put forward combining the neural network to capture the key content, Zha [70] put forward the methods of extracting video features of visual attention, combined with the research in recent years, we will introduce the concept of visual attention to enhance the visual characteristics of expression ability, we will VGGNet as visual emotional feature extraction model as the Equation.2 Simonyan [71]. After obtaining the feature representation we input it into Softmax's full connection layer for visual emotion classification.

$$F_I = CNN_{VGGNet}(I) \tag{2}$$

3. Physiological-Affective Feature Extraction

We refer to Zhang [59] physiological feature extraction method to collect and calculate the physiological data of the audience. Instead of a single mode, this method uses a fusion dual mode depth autoencoder (BDAE) to integrate EEG and facial expression data to obtain an emotional model. In the process of multi-modal emotion recognition, Zhang adopts Restricted Boltzmann Machine (RBM) model. All collected data correspond to the visible layer of the model, and extracted features correspond to the hidden layer of the model. There is no connection between the nodes in the layer but the edge of the layer has. The variables $v \in \{0, 1\}^M$ in the visible layer and $h \in \{0, 1\}^N$ in the hidden layer are defined as follows:

$$E(v, h; \theta) = -\sum_{i=1}^{M}\sum_{j=1}^{N} W_{i,j}\, v_i h_j - \sum_{i=1}^{M} b_i v_i - \sum_{j=1}^{N} a_j h_j \tag{3}$$

C. Multimodal Feature Fusion

1. Score of Textual-Affective Feature

HowNet is mainly divided into Chinese and English parts. There are 3730 Chinese positive evaluation words, 3116 Chinese negative evaluation words, 836 Chinese positive emotion words and 1254 Chinese negative emotion words. There are 3594 positive evaluation words, 3563 positive evaluation words, 769 positive emotion words and 1011 negative emotion words in English. We define the emotional characteristic score of the text as S_t, as the Equation.4.

$$S_t = \sum_{i}^{n} HowNet\,(T_i) \tag{4}$$

Where, n is the number of affective words, T_i is the score value of the i th word in the affective word set in HoeNet, when $S_t > 1$ is positive affective, and when $S_t < -1$ is negative affective.

2. Score of Visual-Affective Feature

After extracting the affective features through VGGNet visual affective feature extraction model, we obtained the affective classification result S_v in the Softmax classifier, as the Equation.5. Where, V_{-1}represents the probability of negative affective, V_0 represents the probability of neutral affective, and V_1 represents the probability of positive affective.

$$S_v = \begin{cases} V_{-1} \\ V_0 \\ V_1 \end{cases} \tag{5}$$

3. Multimodal Feature Fusion

After obtaining the textual—affective feature score S_t, the visual—affective feature score S_v, and the physiological—affective feature score S_p. The weighted sum method is used to fuse the three features to obtain the final emotional score S_{PC-Acc}, which is defined as the affective acceptance, as the Equation.6.

$$S_{PC-\mathrm{Acc}} = \alpha \cdot S_t + \beta \cdot S_v + \gamma S_p \quad (\alpha + \beta + \gamma = 1) \\ \alpha \in [0,1]\ \beta \in [0,1]\ \gamma \in [0,1] \tag{6}$$

In this paper, principal component analysis (PCA) is used to determine the weights of α, β and γ. When $S_{PC-Acc} > 0$, it means that the audience acceptance is positive, so the performance creative is excellent. When $S_{PC-Acc} < 0$, it means that the audience acceptance is negative, so the performance creative is a failure. When $S_{PC-Acc} = 0$, it means that the audience acceptance is neutral, so the performance creative is reasonable.

IV. Result

This paper evaluates the creative of three different performance forms and verifies the validity and rationality of the PC-MulAff model. In this paper, dance performance, drama performance and square performance were selected for the experiment. All the experiments were carried out in the Linux environment on a PC computer with core I7 processor and 512GBRAM. All methods are implemented in Python.

A. Data Description

Lisetti [72] used the audience's physiological signals to recognize the emotion of the movie and divide the emotion into six discrete categories. Sun Kai [73] used the IMDB scoring system based on bayesian statistical algorithm to select the movie emotional content data set and establish the movie emotional space model. MAHNOB HCI [74] movie emotion database was established. Currently there is still a lack of emotional data sets for performances. Our research builds performance evaluation data set with emotional features based on the audience's evaluation text and physiological signals annotated in three common representative performance types (dance performances, drama performances, and square performances).

This study recruited 50 volunteers (25 men and women), age range 22-45 years old, their professional background is computer science (39%), digital media art (46%), dance performance (15%). EEG signal acquisition uses the whole brain induction head-mounted device Emotiv, which consists of 14 channel sensors and 2 bipolar reference electrodes, samples at a rate of 128 Hz, and is directly connected to the computer via Bluetooth. It has better wearing comfort for the audience, which is conducive to the collection of EEG signals. The data collection of the audience's facial expressions uses a conventional high-definition camera with an image resolution of 1920*1080. Considering that the use of wearable eye tracking devices will adversely affect the audience's viewing experience and may cause noise to the EEG signal. Therefore, we add a single camera for the collection of eye movement data and the collection of facial expressions. Eye movement frequency mainly includes eye tracking and saccade. Because eye tracking can clearly describe the audience's interest points in the performance, the saccade movement can well reflect the audience's continuous attention. These two key parameters provide important for the subsequent "Director Label". While the audience is watching the performance videos, we use a smart bracelet to collect the audience's heart rate. The device uses a PPG heart rate sensor, a three-axis acceleration sensor and a three-axis gyroscope, which can more accurately sample and record the audience's heart rate. In this article, we set the heart rate collection interval with 1 minute, and the heart rate detection can assist in verifying the audience's excitement. Each data sample includes 1

piece of performance video, 50 pieces of audience evaluation text and corresponding length of audience physiological signal data (facial expression, EEG, heart rate and eye movement frequency). The form of the data set as in Fig. 2.

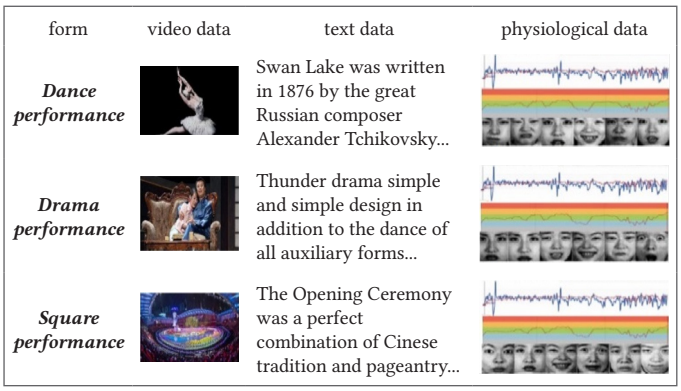


Fig. 2. Performance creative evaluation data set.

The performance creativity evaluation data set including 215 Chinese and Foreign famous dance works, 182 drama works and 213 large-scale square performances (such as opening and closing ceremony performances). The total amount of data is 610 pieces of performance video data. We have performed manual editing for each performance content to remove noise images that are not related to the performance content, and the time for intercepting the performance content is controlled within 30-50 minutes. 30,500 pieces of audience evaluation data are collected. We divide the training set and the test set in a ratio of 7:3, (see Table I).

TABLE I. Experimental Data Set

Dataset	Mode	Amount	TrainSet	TestSet
Dance	performance video	215	150	65
	evaluation text	10750	7500	3250
Drama	performance video	182	127	55
	evaluation text	9100	6350	2750
Square	performance video	213	149	64
	evaluation text	10650	7450	3200
Total	performance video	610	426	184
	evaluation text	30500	21300	9200

B. Results Analysis

1. Comparative Experiment of Singlemodal and Multimodal Affective Features

This article uses a supervised model training method to label the training set and the test set. In view of the professionalism and particularity of the performance data in this article, the research uses the "director label" method on the data set to achieve the most effective training result. Based on the creative intention and aesthetic experience of the performance work, the director defined and marked the high-point and low-point of the performance creative. Among them, high-point corresponds to positive and wonderful ideas, and low-point corresponds to negative failed creative, the others corresponds to general reasonable creative, and the same annotation mode is used for the same text data set evaluated by the audience. The director label as in Fig. 3.

In this paper, the collected physiological data is associated with the director's annotations to form a complete annotated multimodal emotional data set. Annotation system for multimodal affective data sets as in Fig. 4. This part of the experiment compares the creative evaluation methods of singlemodal and multimodal features. First, we only use a single textual feature for creative evaluation. The accuracy of the creative evaluation model training results is shown in Fig. 5.

The x-axis in the figure represents the number of the test set, and the y-axis represents the accuracy rate. It can be seen from the figure that the accuracy rate obtained by the evaluation method of a single textual feature can improve faster. It can be seen from the figure that the accuracy rate obtained by the evaluation method of a single text feature can reach more than 80.56%, indicating that the audience's evaluation quality is relatively objective, but the degree of the audience's emotional fluctuations in the creative is relatively stable, which reflects the evaluation has a greater impact on the audience's artistic background and aesthetic experience. Then, we use a single visual feature for creative evaluation, and the accuracy of the creative evaluation model training results in Fig. 6.

It can be seen from the figure that the accuracy rate obtained by the evaluation method of a single visual feature improve slowly. It can be seen that the evaluation method of visual features has great emotional fluctuations, and the visual effects including the change of light and shade, the richness of color and whether the picture is in a wonderful moment have a greater impact on the evaluation results. Finally, we adopt the creative evaluation method of multimodal feature fusion, and the accuracy of the creative evaluation model training results in Fig. 7.

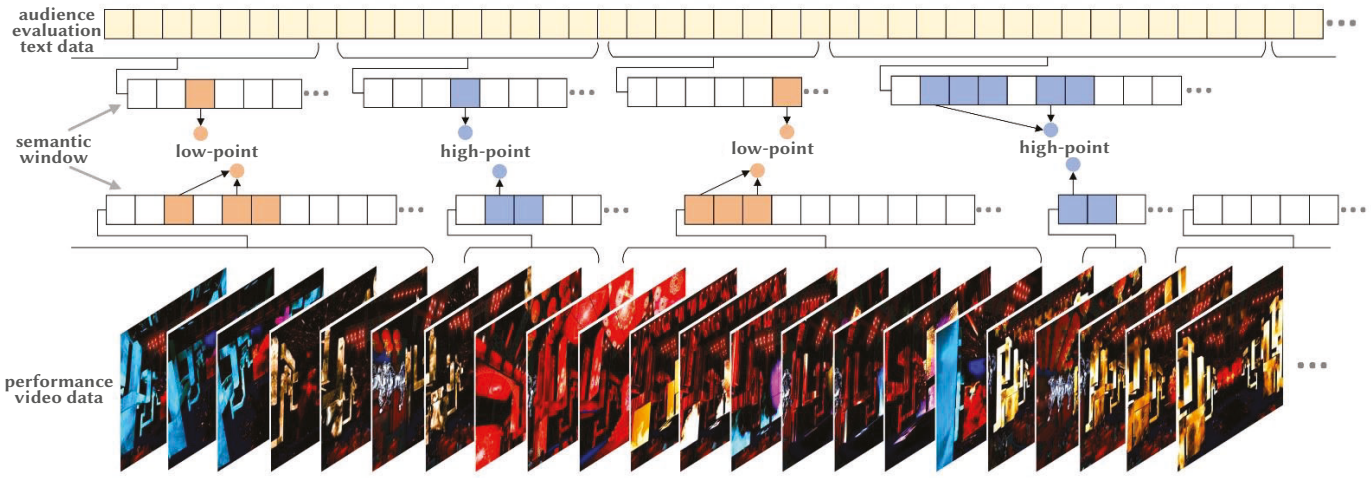


Fig. 3. Director label.

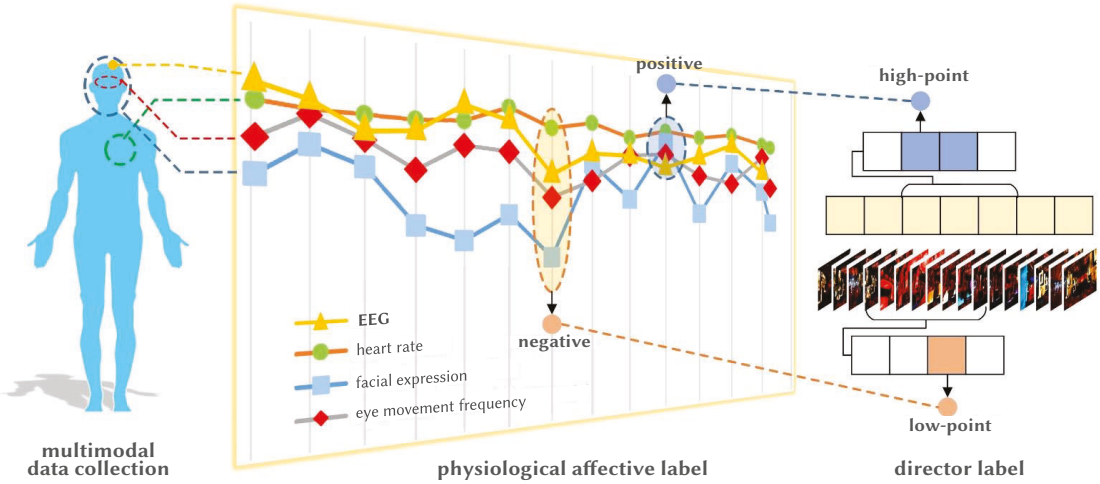


Fig. 4. Annotation system for multimodal affective data.

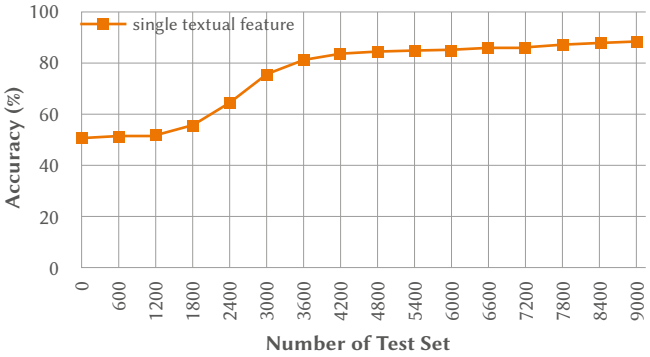


Fig. 5. Single mode evaluation accuracy in single textual feature.

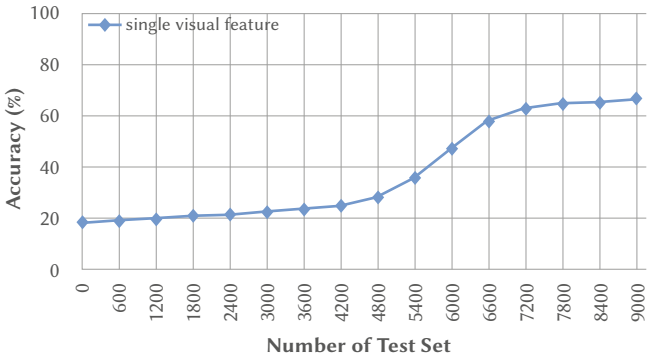


Fig. 6. Single mode evaluation accuracy in single visual feature.

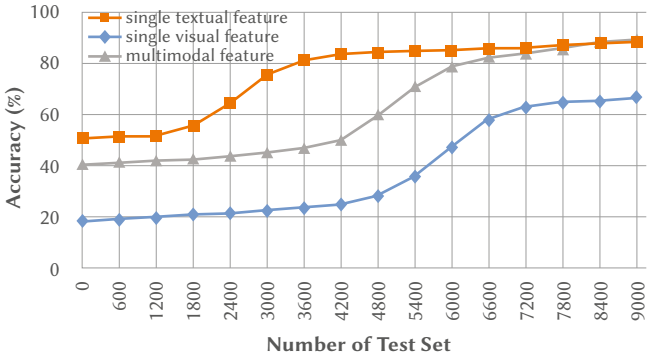


Fig. 7. Comparison of evaluation accuracy between single and multimodal.

From the figure, we can see that the multi-modal method makes up for the shortcomings of the single feature evaluation method. While detecting the objective evaluation of the audience, it also expresses the degree of emotional fluctuation well. Although the accuracy of the multi-modal method before the test data volume of 5400 has not increased faster than the accuracy of the text feature evaluation method, this precisely shows that the multi-modal evaluation method is more stable than the single modal. The multi-modal evaluation model minimizes the impact of noise in the preliminary calculations. After the test data volume is 5400, the accuracy of the multi-modal starts to rise rapidly and begins to exceed the single text feature evaluation when the test data volume is 8400. It conforms to the expectations of the study in this paper, and further verifies the stability and accuracy of the multimodal evaluation method.

2. Comparative Experiments in Three Performance Forms with Singlemodal and Multimodal Affective Features

By experiment method is the textual—affective feature, visual—affective feature and PC-MulAff model performed in three different kinds of performance creative evaluation experiment. The validity of PC-MulAff model in creative performance evaluation is verified. Based on the Accuracy(A), Precision(P) and Recall(R) and F-measure (F) to evaluate the three different performance creative evaluation model. F-measure is the geometric average Charu [57] of accuracy and recall rate, as the Equation.6.

$$F1-score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

The experimental comparison of three performance creative evaluation modes in “Dance Performance” data set (see Table Ⅱ).

TABLE Ⅱ. Experimental Comparison of Three Performance Creative Evaluation Modes in Dance Performance Data Set

Dataset	Mode	Accuracy	Precision	Recall	F1-score
Dance	S_t	0.7674	0.7545	0.7830	0.7684
	S_v	0.7023	0.6909	0.7169	0.7036
	S_{PC-Acc}	0.8418	0.8272	0.8584	0.8386

The experimental comparison of three performance creative evaluation modes in “Dance Performance” data set (see Table Ⅲ).

TABLE III. EXPERIMENTAL COMPARISON OF THREE PERFORMANCE CREATIVE EVALUATION MODES IN DRAMA PERFORMANCE DATA SET

Dataset	Mode	Accuracy	Precision	Recall	F1-score
Drama	S_t	0.6428	0.6272	0.7419	0.6797
	S_v	0.5659	0.5636	0.6666	0.6107
	S_{PC-Acc}	0.6978	0.6727	0.7956	0.7290

The experimental comparison of three performance creative evaluation modes in "Dance Performance" data set (see Table IV).

TABLE IV. EXPERIMENTAL COMPARISON OF THREE PERFORMANCE CREATIVE EVALUATION MODES IN SQUARE PERFORMANCE DATA SET

Dataset	Mode	Accuracy	Precision	Recall	F1-score
Square	S_t	0.6197	0.6090	0.6380	0.6231
	S_v	0.6760	0.6636	0.6952	0.6790
	S_{PC-Acc}	0.7605	0.7454	0.7809	0.7627

It can be seen that the evaluation mode of PC-MulAff model has achieved good expected effects in different evaluation of performance creative in Fig. 8.

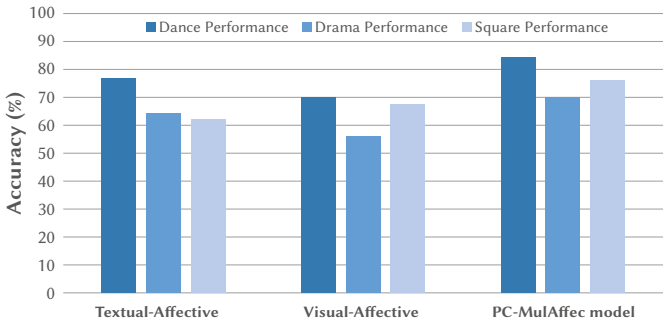


Fig. 8. Comparison of accuracy of three evaluation modes in different performance data.

V. DISCUSSION

Through the comparative experiment 1, we can find the audience's evaluation text data can reach high accuracy in a short time, but the accuracy of performance video data has always been slower to improve for a long period of time. Neither of these two methods can objectively evaluate the accuracy of the model. The multimodal evaluation model makes up for the defects of single mode. Not only the accuracy rate is relatively stable, but also the high accuracy rate of text evaluation can be achieved. Through the comparative experiment 2, we also find the evaluation mode of PC-MulAff model is greatly improved compared with both textual—affective evaluation mode and visual—affective evaluation mode. Especially in accuracy and precision than visual—affective evaluation mode, the recall rate and F1 on increased 0.1395, 0.1363, 0.1415, 0.1350 respectively. It can also be seen from the table that the audience's evaluation effect of textual—affective on dance performance is better than visual—affective, indicating that the audience's evaluation text is richer than visual affective.

The evaluation mode of PC-MulAff model is significantly improved compared with the other two evaluation modes in drama performance. Especially than visual—affective evaluation mode in the accuracy, precision and recall rate and F1 on increased 0.1319, 0.1091, 0.1290, 0.1183 respectively. It can also be seen that the audience's evaluation text is richer than the visual affective features in the drama performance. The evaluation mode of PC-MulAff model is more advanced than the other two evaluation modes in square performance. Especially than textual—affective evaluation mode in accuracy, precision and recall rate and F1 on increased 0.1408, 0.1364, 0.1429, 0.1396 respectively. It can be seen from the table that visual—affective features are more abundant than textual—affective features in large square performances. This shows that in large square performances, the audience is limited by their visual field, so the overall effect of square performances is not strong. The creative effect of square performances is more suitable to be shown through video shots.

From the experimental results, the performance creative evaluation method proposed in this paper is effective and has reached the research expectations. In the perspective of the data set, we create a performance evaluation data set that integrates a variety of physiological signals from the audience and has a "Director Label" for the first time. This multi-modal data set corresponds the director's experience with the audience's physiological feedback for the first time. The establishment of this correspondence has played a key role in verifying the effectiveness and accuracy of the evaluation method. [76] present a Brain-Adaptive Digital Performance (BADP) was designed to measure and analysis of audience engagement level. Only to detect changes in the audience's engagement through the monitoring of EEG signals. First of all, in terms of the types of performances tested of stage performances are given, and there is no clear explanation of the types of performances and problem boundaries of participation monitoring. Second, the article only conducted monitoring experiments in a virtual performance environment. However, the audience can adapt more to the artistic perception ability and cognitive level of the real performance scene. The lack of experience in watching virtual performances and the freshness and cognition of visual perception will affect the accuracy of EEG signals. The authors did not give a method to remove the noise signal. The focus of our research is to use EEG signals as the attributes of the evaluation data set, and there are other physiological signal data as mutual verification of emotional features, which not only ensures the accuracy of emotional features, but also enhances the adaptability of the evaluation model. A multi-modal emotion recognition framework called EmotionMeter is proposed [77], which combined brain waves and eye movements, and verified that the modal fusion of multi-modal deep neural networks is higher than the performance of a single modal. However, the training data set mentioned in the article has differences and instabilities in cross-modal feature distribution. In the research of our paper, the method of "Director Label" is adopted to effectively avoid this defect. A Conditional Generative Adversarial Network (cGAN) is proposed to establish emotion-related EEG data [78], and two components that constitute an emotional EEG signal (YEEG) are defined: emotion-related (YE motion) and emotion-related (NOthers). While, the accuracy of labeling with facial expression images needs to be further improved. And the current coarse-grained labeling has limitations in emotion recognition and analysis.

We can also further conclude that in the evaluation model of dance performances the accuracy rate is the highest from Fig. 8, which can reflect that the audience's aesthetic feelings and perceptions of dance performances are higher than those of drama performances and square performances. It also confirms that the dance performances that the audience are exposed to are more than other performances. In the evaluation model of drama performance, the evaluation of text features is more accurate than the evaluation of visual features, which also shows that the literary of drama performance is stronger, and the audience's drama and literature accomplishment is higher than that of the square performance. In the evaluation model of square performances, visual feature evaluation is more accurate than text

feature evaluation, which shows that the creative core of large-scale square performances mainly revolves around the performance of visual effects. These viewpoints demonstrated from experiments not only show the effectiveness of the performance creative evaluation model proposed in this article, but also highlight the creative core of different performance forms. This helps directors to carry out effective and reasonable performance creative, and improves the performance creative efficiency and level. At the same time, the actor and audience's aesthetic perception of performance is improved. Therefore, it can be seen that our research has high potential value and practical benefits.

VI. Conclusion

Experiments show that the PC-MulAff model is effective, especially in the comparative experiments for three different performance modes, PC-MulAff model has achieved good results. The purpose of this article is to evaluate performance creative through multimodal affective features. In order to achieve this goal, the affective features of the audience and the visual features of the performance video are extracted respectively, and the performance creative is analyzed through the quantified score after multimodal feature fusion. The main contributions of this article: 1). this paper proposes a PC-Acc to evaluate the quality of performance creative, trains and builds a PC-MulAff model, which can evaluate creative for different performance forms. And also 2). we propose a new "Performance Creativity-Multimodal Evaluation Data Set", which is composed of performance video data, audience evaluation text and audience physiological data. It not only makes up for the insufficient description of features by a single data type, but also provides a performance evaluation data set type with multiple physiological signal emotional features. This work provides a standardized verification basis for performance evaluation and fills the gap in the direction of the verification in performance evaluation data set. 3) Based on the establishment of multi-modal evaluation data, the correlation analysis between the audience's multi-modal physiological signals and different performance types has been realized. For the first time, the audience's emotional features and performance creative has been mapped through the method of "Director Label". This work plays a decisive role in the evaluation of performance creative. Digital and intelligent technical provide directors with scientific evaluation methods and verification basis.

Based on the limitations of currently collected performance works and experimental scenes, although the research methods proposed in this paper have been substantively verified, we believe that the existing evaluation framework can still be further optimized in future work , In particular, research on related algorithms for precise extraction of performance content based on emotion classification, and optimization of experimental scenes, to provide volunteers with a more immersive environment to extract more accurate physiological data and improve multimedia evaluation data sets . We hope that the next step will continue the ideas of this paper, and make corrections and adjustments to the score calculation including algorithm parameter adjustments of multi-modal special fusion.

The multimodal data-driven performance creative evaluation method proposed in this paper is effective. The model not only provides a multi-dimensional analysis for the performance creativity evaluation, but also proposes to solve the interpretability and data set support of creative evaluation of performing arts.

References

[1] I. S. Lee, "Performing arts in the age of transmedia," *Journal of acting studies*, vol. 17, pp. 17–32, 2020.

[2] K.-W. Huang, C.-C. Lin, Y.-M. Lee, Z.-X. Wu, "A deep learning and image recognition system for image recognition," *Data Science and Pattern Recognition*, vol. 3, no. 2, pp. 1–11,2019.

[3] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[5] J. S. Chung, B.-J. Lee, I. Han, "Who said that?: Audio-visual speaker diarisation of real-world meetings," *arXiv preprint arXiv:1906.10042*, 2019.

[6] J. Wu, Y. Xu, S.-X. Zhang, L.-W. Chen, M. Yu, L. Xie, D. Yu, "Time domain audio visual speech separation," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 667–673, IEEE.

[7] J. C.-W. Lin, Y. Shao, Y. Djenouri, U. Yun, "Asrnn: a recurrent neural network with an attention model for sequence labeling," *Knowledge-Based Systems*, vol. 212, p. 106548, 2020.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[9] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[10] J. C.-W. Lin, G. Srivastava, Y. Zhang, Y. Djenouri, M. Aloqaily, "Privacy preserving multi-objective sanitization model in 6g iot environments," *IEEE Internet of Things Journal*, 2020.

[11] F. Abbé-Decarroux, "The perception of quality and the demand for services: Empirical application to the performing arts," *Journal of Economic Behavior & Organization*, vol. 23, no. 1, pp. 99–107, 1994.

[12] F. Nake, "Computer art: creativity and computability," in *Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition*, 2007, pp. 305–306.

[13] K. Yamada, T. Taura, Y. Nagai, "Design and evaluation of creative and emotional motion," in *Proceedings of the 8th ACM conference on Creativity and cognition*, 2011, pp. 239–248.

[14] C.-y. Chang, Y.-p. Chen, "Fusing creative operations into evolutionary computation for composition: From a composer's perspective," in *2019 IEEE Congress on Evolutionary Computation (CEC)*, 2019, pp. 2113–2120, IEEE.

[15] L. Goves, "Multimodal performer coordination as a creative compositional parameter," *Tempo*, vol. 74, no. 293, pp. 32–53, 2020.

[16] D. Cabral, J. G. Valente, U. Aragão, C. Fernandes, N. Correia, "Evaluation of a multimodal video annotator for contemporary dance," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 2012, pp. 572–579.

[17] R. E. Cisneros, K. Wood, S. Whatley, M. Buccoli, M. Zanoni, A. Sarti, "Virtual reality and choreographic practice: The potential for new creative methods," *Body, Space & Technology*, vol. 18, no. 1, 2019.

[18] B. T. Christensen, L. J. Ball, "Dimensions of creative evaluation: Distinct design and reasoning strategies for aesthetic, functional and originality judgments," *Design Studies*, vol. 45, pp. 116–136, 2016.

[19] P. Karimi, N. Davis, M. L. Maher, K. Grace, L. Lee, "Relating cognitive models of design creativity to the similarity of sketches generated by an ai partner," in *Proceedings of the 2019 on Creativity and Cognition*, 2019, pp. 259–270.

[20] T. Knearem, X. Wang, J. Wan, J. M. Carroll, "Crafting in a community of practice: Resource sharing as key in supporting creativity," in *Proceedings of the 2019 on Creativity and Cognition*, 2019, pp. 83–94.

[21] M. Richardson, F. Hernández-Hernández, M. Hiltunen, A. Moura, M. Fulková, F. King, F. M. Collins, "Creative connections: The power of contemporary art to explore european citizenship," *London Review of Education*, 2020.

[22] K. H. Koh, "Computing indicators of creativity," in *2011 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, 2011, pp. 231–232, IEEE.

[23] A. Jain, "Measuring creativity: Multi-scale visual and conceptual design analysis," in *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*, 2017, pp. 490–495.

[24] J. Oppenlaender, "Supporting creative workers with crowdsourced

feedback," in *Proceedings of the 2019 on Creativity and Cognition*, 2019, pp. 646–652.

[25] D. Abfalter, "Authenticity and respect: Leading creative teams in the performing arts," *Creativity and innovation management*, vol. 22, no. 3, pp. 295–306, 2013.

[26] J. Lee, E. Jun, J. Chae, "Big data analysis for dance studies using text mining," *The Journal of Dance Society for Documentation & History*, vol. 42, p. 191—212, 2016.

[27] L. J. Min, "An analysis of semantic relations in knowledge information in dance research data in korea from 1958 to 2016," *The Korean Journal of Arts Studies*, no. 16, p. 215—237, 2017.

[28] K. H. Ryeon, "Exploring the determinants of korean dance recognition and importance: Application of decision tree analysis based on data mining," *Dance Research Journal of Dance*, vol. 77, no. 1, p. 17—29, 2019.

[29] Choi, Hyo-jin, "Previous study research on korean contemporary dance using text mining," *The Korean Journal of Dance Studies*, vol. 76, no. 4, p. 97—111, 2019.

[30] K. Woo-Kyung, J.-Y. Yoo, "Analysis on the trends of research themes of the korean dance using text mining," *Journal of the Korea Entertainment Industry Association*, vol. 13, no. 5, p. 215—228, 2019.

[31] choihyojin, "Analysis of korean contemporary dance research trends using text mining," *Korean Journal of Arts Education*, vol. 17, no. 4, p. 103—118, 2019.

[32] Kimhayeon, "Analysis on the international contemporary dance research trend using text mining," *Korean Journal of Arts Education*, vol. 18, no. 1, p. 171—192, 2020.

[33] S. Zhou, J. Jia, Y. Wang, W. Chen, F. Meng, Y. Li, J. Tao, "Emotion inferring from large-scale internet voice data: A multimodal deep learning approach," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, 2018, pp. 1–6, IEEE.

[34] W. Liang, H. Xie, Y. Rao, R. Y. Lau, F. L. Wang, "Universal affective model for readers' emotion classification over short texts," *Expert Systems with Applications*, vol. 114, pp. 322–333, 2018.

[35] H. M. Hung, H.-J. Yang, S.-H. Kim, G.-S. Lee, "Variants of bert, random forests and svm approach for multimodal emotion-target sub-challenge," arXiv preprint *arXiv:2007.13928*, 2020.

[36] P. T. Sowden, L. Dawson, "Creative feelings: the effect of mood on creative ideation and evaluation," in *Proceedings of the 8th ACM Conference on Creativity and Cognition*, 2011, pp. 393–394.

[37] G. Corness, K. Carlson, T. Schiphorst, "Audience empathy: a phenomenological method for mediated performance," in *Proceedings of the 8th ACM conference on Creativity and cognition*, 2011, pp. 127–136.

[38] C. K. Coursaris, W. Van Osch, "A cognitive-affective model of perceived user satisfaction (campus): The complementary effects and interdependence of usability and aesthetics in is design," *Information & Management*, vol. 53, no. 2, pp. 252–264, 2016.

[39] K. Altuwairqi, S. K. Jarraya, A. Allinjawi, M. Hammami, "A new emotion–based affective model to detect student's engagement," *Journal of King Saud University-Computer and Information Sciences*, 2018.

[40] F. Rahdari, E. Rashedi, M. Eftekhari, "A multimodal emotion recognition system using facial landmark analysis," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 43, no. 1, pp. 171–189, 2019.

[41] P. D. Loprinzi, S. Pazirei, G. Robinson, B. Dickerson, M. Edwards, R. E. Rhodes, "Evaluation of a cognitive affective model of physical activity behavior," *Health Promotion Perspectives*, vol. 10, no. 1, p. 88, 2020.

[42] W. Liu, X. Xie, S. Ma, Y. Wang, "An improved evaluation method for soccer player performance using affective computing," in *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, 2020, pp. 324–329, IEEE.

[43] W. Wei, Q. Jia, Y. Feng, G. Chen, M. Chu, "Multi-modal facial expression feature based on deep-neural networks," *Journal on Multimodal User Interfaces*, vol. 14, no. 1, pp. 17–23, 2020.

[44] G. Chen, X. Zhang, Y. Sun, J. Zhang, "Emotion feature analysis and recognition based on reconstructed eeg sources," *IEEE Access*, vol. 8, pp. 11907–11916, 2020.

[45] H.-J. Choi, Y.-J. Lee, "Deep learning based response generation using emotion feature extraction," in *2020 IEEE International Conference on Big Data and Smart Computing (Big- Comp)*, 2020, pp. 255–262, IEEE.

[46] C. Deepika, "Speech emotion recognition feature extraction and classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 1257–1261, 2020.

[47] W. Wei, Q. Jia, Y. Feng, G. Chen, M. Chu, "Multi-modal facial expression feature based on deep-neural networks," *Journal on Multimodal User Interfaces*, vol. 14, no. 1, pp. 17–23, 2020.

[48] J. Radbourne, K. Johanson, H. Glow, T. White, "The audience experience: Measuring quality in the performing arts," *International journal of arts management*, pp. 16–29, 2009.

[49] Radbourne, Jennifer, "The quest for self actualization meeting new consumer needs in the cultural industries," *In ESRC Seminar Series Creative Futures-Driving the Cultural Industries Marketing Agenda*, vol. 6, 2007.

[50] C. Latulipe, E. A. Carroll, D. Lottridge, "Love, hate, arousal and engagement: exploring audience responses to performing arts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 1845–1854.

[51] C. Wang, E. N. Geelhoed, P. P. Stenton, P. Cesar, "Sensing a live audience," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014, pp. 1909–1912.

[52] C. Martella, E. Gedik, L. Cabrera-Quiros, G. Englebienne, H. Hung, "How was it? exploiting smartphone sensing to measure implicit audience responses to live performances," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 201–210.

[53] R. Adolphs, D. Tranel, A. R. Damasio, "Dissociable neural systems for recognizing emotions," *Brain & Cognition*, vol. 52, no. 1, pp. 61–69, 2003.

[54] A. R. Damasio, T. J. Grabowski, A. Bechara, H. Damasio, L. L. Ponto, J. Parvizi, R. D. Hichwa, "Subcortical and cortical brain activity during the feeling of self-generated emotions," *Nature neuroscience*, vol. 3, no. 10, pp. 1049–1056, 2000.

[55] J. Radbourne, K. Johanson, H. Glow, T. White, "The audience experience: Measuring quality in the performing arts," *International journal of arts management*, pp. 16–29, 2009.

[56] M. Wyczesany, S. J. Grzybowski, R. J. Barry, J. Kaiser, A. M. Coenen, A. Potoczek, "Covariation of eeg synchronization and emotional state as modified by anxiolytics," *Journal of Clinical Neurophysiology*, vol. 28, no. 3, pp. 289–296, 2011.

[57] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[58] G. Pinto, J. M. Carvalho, F. Barros, S. C. Soares, A. J. Pinho, S. Brás, "Multimodal emotion evaluation: A physiological model for cost-effective emotion classification," *Sensors*, vol. 20, no. 12, p. 3510, 2020.

[59] H. Zhang, "Expression-eeg based collaborative multimodal emotion recognition using deep autoencoder," *IEEE Access*, vol. 8, pp. 164130–164143, 2020.

[60] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[61] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, "Improving language understanding by generative pretraining," 2018.

[62] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[63] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942,* 2019.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[65] C.-M. Kuo, N.-C. Yang, J.-Y. Wu, S.-C. Chen, "Histogrambased image enhancement in quasi-spatial domain for compressed image,"

[66] S. Zhao, Y. Gao, X. Jiang, H. Yao, T.-S. Chua, X. Sun, "Exploring principles-of-art features for image emotion recognition," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 47–56.

[67] C. Xu, S. Cetintas, K.-C. Lee, L.-J. Li, "Visual sentiment prediction with deep convolutional neural networks.," *arXiv preprint arXiv:1411.5731,* 2014.

[68] B. Jou, S.-F. Chang, "Deep cross residual learning for multitask visual recognition," in *Proceedings of the 24th ACM international conference on*

Multimedia, 2016, pp. 998–1007.

[69] L. Gao, Z. Guo, H. Zhang, X. Xu, H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[70] N. Zhao, H. Zhang, R. Hong, M. Wang, T.-S. Chua, "Videowhisper: Toward discriminative unsupervised video feature learning with attention-based recurrent neural networks," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2080–2092, 2017.

[71] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[72] C. L. Lisetti, F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP Journal on Advances in Signal Processing* vol. 2004, no. 11, p. 929414, 2004.

[73] S. un Kai, Y. Junqing, "Audience oriented personalized movie affective content representation and recognition," *Journal of Computer-Aided Design &Computer Graphics*, vol. 22, no. 1, pp. 136–144, 2010.

[74] M. Soleymani, J. Lichtenauer, T. Pun, M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42–55, 2011.

[75] C. C. Aggarwal, C. Zhai, Mining text data. Springer Science & Business Media, 2012.

[76] S. Yan, G. Ding, H. Li, N. Sun, Z. Guan, Y. Wu, L. Zhang, T. Huang, "Exploring audience response in perfor ming arts with a brain-adaptive digital performance system," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 7, no. 4, pp. 1–28, 2017.

[77] W.-L. Zheng, W. Liu, Y. Lu, B.-L. Lu, A. Cichocki, "Emotionmeter: A multimodal framework for recognizing human emotions," *IEEE transactions on cybernetics*, vol. 49, no. 3, pp. 1110–1122, 2018.

[78] B. Fu, F. Li, Y. Niu, H. Wu, Y. Li, G. Shi, "Conditional generative adversarial network for eeg-based emotion finegrained estimation and visualization," *Journal of Visual Communication and Image Representation*, vol. 74, p. 102982.

Yufeng Wu

Yufeng Wu obtained B.E. degree from Yantai University, China in 2011, and he is currently working toward the PhD in the School of Computer Science and Technology, Beijing Institute of Technology. His research interests are digital performance, computer simulation, deep learning and computer vision. His main research includes: intelligent generation methods for performance creativity, construction of the framework and theoretical system of performance creativity; and performance data analysis methods and creativity evaluation modeling based on graph neural networks and deep reinforcement learning. He is also working on developing a highly dynamic and intelligent creative evaluation platform based on the human-machine collaboration.

Longfei Zhang

Longfei Zhang obtained Ph.D. from School of Computer Science and Engineering, Beijing Institute of Technology, China in 2005. He is an associate professor in School of Computer Science and Technology at Beijing Institute of Technology. He went to Carnegie Mellon University as a visiting scientist from 2009 to 2011. His main research focuses on "Analysis, Prediction and Construction of 3D Behaviors of Video Personnel", "Cross Media Knowledge Base and Common Sense Library", "Intelligent Performance Creation and Evaluation", "Sports Content Understanding and Intelligent Directing".

Gangyi Ding

Gangyi Ding received the B.E. degree from Peking University, China in 1988, and Ph.D. at Beijing Institute of Technology, China in 1993. He is a professor with the Key Laboratory of Digital Performance and Simulation Technology of the Beijing Institute of Technology. He joined the faculty at the Beijing Institute of Technology in 1993. His research mainly involves training simulation, large-scale crowd simulation, environment simulation, digital performance and creative simulation. He provided simulation technical support for the arrangement of large-scale events such as the opening and closing ceremonies of the Beijing 2008 Olympic Games, and the Beijing 8-Minutes of the Pyeongchang Winter Olympics.

Tong Xue

Tong Xue received her B.E. degree from Communication University of China in 2016. She is currently working toward the PhD degree in School of Computer Science and Technology, Beijing Institute of Technology. She is a joint PhD student at Distributed and Interactive Systems, Centrum Wiskunde & Informatica (CWI). Her research interests lie in human-computer interaction and affective computing.

Fuquan Zhang

Fuquan Zhang received the PhD degree in School of Computer Science & Technology, Beijing Institute of Technology, China in 2019. Now he is a professor of Minjiang University, China. He is now a member of the National Computer Basic Education Research Association of the National Higher Education Institutions, a member of the Online Education Committee of the National Computer Basic Education Research Association of the National Institute of Higher Education, a member of the MOOC Alliance of the College of Education and Higher Education Teaching Guidance Committee, ACM SIGCSE, CCF member, CCF YOCSEF member, director of Fujian Artificial Intelligence Society.

Modified YOLOv4-DenseNet Algorithm for Detection of Ventricular Septal Defects in Ultrasound Images

Shih-Hsin Chen[1], Chun-Wei Wang[1], I-Hsin Tai[2]*, Ken-Pen Weng[3]*, Yi-Hui Chen[4,5]*, Kai-Sheng Hsieh[6,7]*

1 Department of Information Management, Cheng Shiu University, Kaohsiung City 83347, Taiwan (R. O. C.)
[2] Department of Pediatric Cardiology, China Medical University Chidren's Hospital, Taichung City 40447, Taiwan (R. O. C.)
[3] Congenital Structural Heart Disease Center, Department of Pediatrics, Kaohsiung Veterans General Hospital, No.386, Dazhong 1st Rd., Zuoying Dist., Kaohsiung City, Taiwan (R. O. C.)
[4] Department of Information Management, Chang Gung University, Taoyuan 33302, Taiwan (R. O. C.)
[5] Kawasaki Disease Center, Kaohsiung Chang Gung Memorial Hospital, Kaohsiung 83301, Taiwan (R. O. C.)
[6] Department of Pediatrics, Shuang-Ho Hospital-Taipei Medical University, New Taipei City, 23561, Taiwan (R. O. C.)
[7] Taipei Heart Institute, Taipei Medical University, Taipei City, Taiwan (R. O. C.)

Received 30 October 2020 | Accepted 31 March 2021 | Published 8 June 2021

Abstract

Doctors conventionally analyzed echocardiographic images for diagnosing congenital heart diseases (CHDs). However, this process is laborious and depends on the experience of the doctors. This study investigated the use of deep learning algorithms for the image detection of the ventricular septal defect (VSD), the most common type. Color Doppler echocardiographic images containing three types of VSDs were tested with color doppler ultrasound medical images. To the best of our knowledge, this study is the first one to solve this object detection problem by using a modified YOLOv4–DenseNet framework. Because some techniques of YOLOv4 are not suitable for echocardiographic object detection, we revised the algorithm for this problem. The results revealed that the YOLOv4–DenseNet outperformed YOLOv4, YOLOv3, YOLOv3–SPP, and YOLOv3–DenseNet in terms of metric mAP-50. The F1-score of YOLOv4-DenseNet and YOLOv3-DenseNet were better than those of others. Hence, the contribution of this study establishes the feasibility of using deep learning for echocardiographic image detection of VSD investigation and a better YOLOv4-DenseNet framework could be employed for the VSD detection.

Keywords

Ventricular Septal Defect (VSD), Doppler Echocardiographic Images, Object Detection, Deep Learning, YOLOv4.

DOI: 10.9781/ijimai.2021.06.001

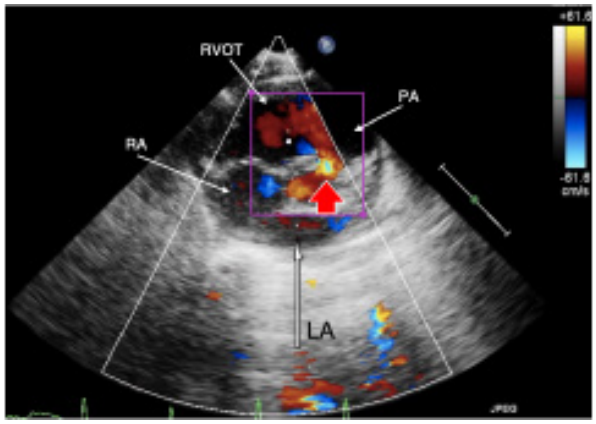
I. Introduction

Wu et al. [1] studied the 7-year data of the Taiwan health insurance database and determined that, on average, 13 out of 1000 newborn infants have congenital heart disease (CHD) annually. The 5-year relative mortality rate of infants with CHDs is 5% [2]. With the advances in medical technology, CHDs can be detected using ultrasound images obtained at 18–22 weeks of gestation [3]. Echocardiographic images do not involve radioactivity, cause minimal stress on fetuses and newborns, and are cost effective [4]–[6]. Most parents and doctors begin medical planning at 18–22 weeks. Therefore, it is critical to detect CHDs by using ultrasound images in the early stage.

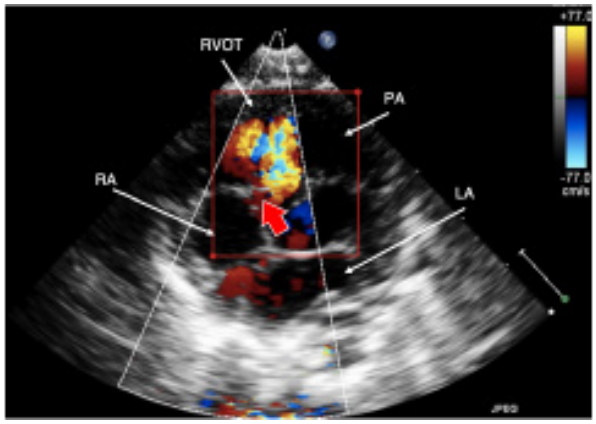
Two types of ultrasound images are possible, namely black-and-white and color doppler echocardiographic images. The color doppler echocardiographic images can provide critical information on velocities, accelerations, direction of the heart's blood flow (denoted by the red and blue color, respectively), flow rate, and whether the blood pressure is diastolic or systolic [7]. The red and blue colors represent blood flowing toward and leaving the ultrasound probe, respectively. A golden-yellow color indicates a rapid blood flow. Doppler ultrasound images provide essential features of CHD. Therefore, doppler echocardiographic imaging was used in this study.

The ventricular septal defect (VSD) is the most common CHD and accounts for upto 30% of all CHDs [1]. Therefore, the VSD recognition problem was investigated in this study. Furthermore, three subtypes of VSD, namely Type 1, Type 2, and Type 4, were included in this study. Because type 3 VSD usually is associated with other congenital cardiac anomaly (so called endocardial cushion defect or atrial-ventricular canal defect) and it is very rare to have isolated type 3 VSD; thus, we exclude type 3 VSD to study. Doppler echocardiographic images of the three images are depicted in Fig. 1. Both VSD Type 1 and Type 2 typically involve the parasternal short-axis view's aortic root section. The aortic valve is at the center of the echocardiographic image. Fig. 1a in Fig. 1 shows that if the blood flow of the hole is between 11 and 1 o'clock, then the VSD may be Type 1. If the hole blood flow is between 9 and 11 o'clock, then the VSD is Type 2. The echocardiographic image of a patient with VSD Type 2 are presented in Fig. 1b of Fig. 1.

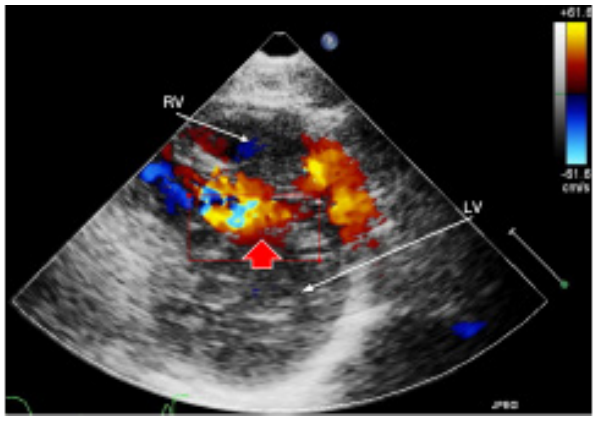
* Corresponding author.
E-mail addresses: ifaithgrace@icloud.com (I. H. Tai), kpweng@vghks.gov.tw (K. P. Weng), kshsieh@hotmail.com (K. S. Hsieh), cyh@mail.cgu.edu.tw (Y. H. Chen).



(a) The short axis imaging at cardiac base show-ing Type 1 VSD color flow jet (arrow-head)



(b) The short axis imaging at cardiac base show-ing Type 2 VSD color flow jet (arrow-head)



(c) The short axis imaging near cardiac apex showing Type 4 VSD color flow jet (arrow-head)

Fig. 1. Doppler echocardiographic of VSD Type 1, Type 2 and Type 4 in the parasternal short-axis view (LV: Left Ventricle, RV: Right Ventricle, RVOT: Right Ventricular Outflow Tract, LA: Left Atrium, RA: Right Atrium, AV: Aortic Valve, PA: Pulmonary Artery).

For identification of Type 4 VSD, doctors evaluate the horizontal section of the mitral valve of the parasternal short-axis view and determine the doppler flow in the left ventricle from 9 to 1 oâ€™clock. In practice, more than one hole of Type 4 VSD may oc-cur. Therefore, the recognition of this VSD type is difficult. In Fig. 1c, a spot was observed in the 12 o'clock position. To the best of our knowledge, this study is the first to focus on the VSD object detection problem. There is much room to study this problem.

Different views and angles may lead to difficulties in identifying the characteristics of the three VSD types. Pézard et al. [8] stated that ultrasound image detection of CHDs involved challenges such as the ability and experience of physicians or radiologists [8], ambiguous images, and the nature of defects, which may affect the outcome of the judgment [9]. Furthermore, doctors may capture many ultrasound images when they check a patient; however, determining appropriate image data is time con-suming. Deep learning (DL) algorithms used for automatic detection and segmentation of complex cardiac echocardiographic structures may detect the region of interest (ROI) rapidly, thus reducing time and effort required for the process [5], [10]-[15].

You only look once (YOLO)v3 [16], YOLOv4 [17], RetinaNet [18], and Faster RCNN [19] are some popular algorithms. YOLOv4 is a state-of-the-art algorithm [17] that can improve the quality and efficiency of detection. Compared with YOLOv3, YOLOv4 integrates numerous methods. However, according to our pilot experiments, we determined that some methods may not apply to our studied problem. A modified YOLOv4 was used in the study.

YOLOv4 comprises a cross-stage partial (CSP) network [20] and DarkNet [21]. Because DenseNet [22] can extract more features than DarkNet, DarkNet was replaced with DenseNet121 in YOLOv3, which provided better results [23], [24]. Therefore, we replaced CSPDarkNet with CSPDenseNet121 in YOLOv4. The model was named YOLOv4-DenseNet121 or YOLOv4-DenseNet. To the best of our knowledge, this study is the first to attempt such modification. The proposed algorithm, YOLOv4-DenseNet, is the other major contribution of this study.

The rest of this paper is organized as follows. Section II provides detailed steps of data collection and automatic organization of the data set. The patients' ID, name, or birthday, were removed to ensure privacy and avoid information leak. In Section III, we present the YOLOv4-DenseNet algorithm. Next, we compare the revised YOLOv4-DenseNet with the unmodified YOLOv4 and some variants of YOLOv3. We present a comparison of the proposed algorithms in Section IV. Finally, we present our conclusions in Section V.

II. Medical Image Collection and Automation of Data Set Arrangement

The ultrasound images used in this study were provided by Kaohsiung Veterans General Hospital[1]. Videos were transformed into frames/figures. Doctors identified each figure from the collected images to ensure the correctness of the data set. Sorting the correct classification and the required bounding boxes can be time-consuming. This study developed a standard operating procedure in Fig. 2 considering the professional ability of the doctor, protection of the patient's information, and correctness of the data. The step-by-step procedure is as follows.

1. Echocardiographic videos were provided by Kaohsiung Veterans General Hospital. The study protocol was approved by the Institutional Review Board (IRB) of the hospital [IRB number is 19-CT8-10(190701-2)]. Patients with VSD diagnosis were selected for the study.
2. We extracted each frame from the echocardiographic videos. The principal image resolution was 800×600.
3. Privacy: We removed images that contained the patient's personal information, such as name, case number, and date of birth. Thus, we removed any identifying information from each figure. The final resolution of the images was 706 × 532.
4. The images were converted to the PNG format to ensure compression without loss of information. The file name of the

[1] https://eng.vghks.gov.tw/

image was encrypted using the original video file name, and the custom private key of the project was encrypted using the advanced encryption standard algorithm combined with the video and frame number.

5. We used a few data sets to train the DL model. Then, the initial model was used to classify the images that were not arranged.
6. We then examined the classification results and labeled the images.
7. At least one cardiologist verified the classification and ROIs. Before implementing Step 1, the coauthors obtained the required licenses from the relevant human research ethics committee. Steps 3–6 were executed after Kaohsiung Veterans General Hospital provided the required medical images. The final steps in processing these data still relied on professional doctors' judgment to provide adequate training quality.

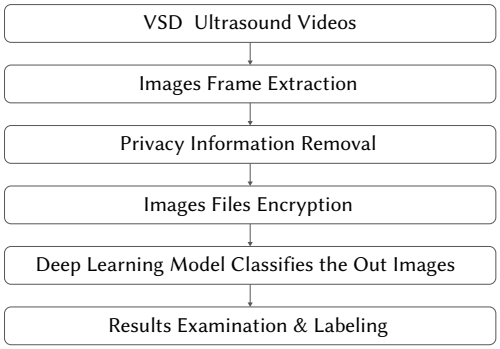


Fig. 2. Dataset arrangement procedures.

III. Methods

The YOLOv4 algorithm is described in Section III.A. The differences between YOLOv4 and the modified YOLOv4-DenseNet are described in Section III.B. The metrics applied to evaluate the performance of the proposed algorithm against other algorithms are presented in Section III.C.

A. Main Characteristics of the YOLOv4 Framework

State-of-the-art algorithms, such as CSP, spatial pyramid pooling (SPP), feature pyramid network (FPN), path aggregation network (PANet), Mish activation function, Mosaic augmentation, dropblock, complete IoU loss (CIoU), class label smoothing, and cosine annealing scheduler, are incorporated in YOLOv4. Fig. 3 presents the modified YOLOv4 framework, which includes three parts, namely the backbone, neck, and head.

CSPDarknet is the main characteristic of the YOLOv4 backbone. The CBM, which is a combination of the convolution layer and the batch normalization (BN) and Mish activation function, was the input of CSPDarkNet. The input resolution of the first convolution layer was 608*608. The Mish function is a self-regular non-monotonic neural activation function that allows relavant information to penetrate the neural network. The ZCRn is composed of zero padding, CBM, and CSPRn. CSPRn denotes the CSPNet framework with n number of replications. CSPNet divided the feature maps into two parts. In the first part, the gradient changes from the beginning to the end are recorded into the feature map, which reduces the number of calculations and memory costs and ensures high accuracy. The second includes the ResNet skip connections. Finally, the first part is concatenated with the second part's feature maps. The output resolutions of ZCR1, ZCR2, and ZCR8 were 76 × 76, 38 × 38, and 19 × 19, respectively.

In the neck area, the FPN and PANet are used in YOLOv4, whereas only the FPN is used in YOLOv3. The FPN performs the upsampling from a smaller resolution to larger resolutions and then concatenates with the large-size ZCRn. The PANet framework employs the bottom-up path augmentation with prior local convolution layers through the upsampling operation to shorten the information path between high- and low-resolution features.

In the head area, YOLOv4 and YOLOv3 use the same head. The output resolution with the number of feature maps was 76 × 76/256, 38 × 38/512, and 19 × 19/1024. The only change is the loss function. The CIoU is used as the loss function of YOLOv4 to measure the difference between the ground truth and predicted box.

B. Modified YOLOv4-DenseNet Algorithm

The results of our pilot experiments revealed that not all approaches of YOLOv4 suited our problem. In particular, the performance of the mosaic data augmentation, SPP, and cosine annealing scheduler was not satisfactory. For example, although the mosaic augmentation method is the major cause of the superior performance of YOLOv4 in small object detection, the size of the CHD detection objects is large, and characteristics of CHD are located at a specific area. The cosine annealing scheduler did not yield superior performance. Therefore, the algorithm was suitably modified to address the aforementioned characteristics. Because of the GPU memory limitation, our input resolution decreased to 416 × 416 instead of 608 × 608 used in YOLOv4.

To enhance the performance of the YOLOv4 algorithm, this study replaced CSPDarkNet with CSPDenseNet because DenseNet extracts more information than DarkNet does. In the backbone area of Fig. 4, the ZCRn block was replaced with the CSPDn block. The values of n are 6, 16, and 24. Two subblocks belonged to the CSPDn block, namely the repeated Denseblocks (blk) and transition block. The number of times Dense blk is repeated is based on the value of n. In the proposed algorithm, the FPN and PANet were implemented in the neck area. The first dif-ference between the proposed algorithm and YOLOv4 is that we removed the SPP because the SPP lowers performance. Second, because the SPP was removed, the number of CBL was six instead of seven in the top branch.

Finally, in the head area, we used a small input resolution; thus, the output scales were 52 × 52, 26 × 26, and 13 × 13. Furthermore, we determined that the number of output feature maps may not be useful for our studied problem. To solve this problem, we added more feature maps in the YOLO head. The corresponding output resolution with the number of feature maps was 52 × 52/512, 26 × 26/1024, and 13 × 13/1024. The loss function of the proposed algorithm was identical to those of the original YOLOv4 algorithm.

C. Evaluation Metrics

The unlearned image was the target of the test. The threshold of intersection over union was set to 50. If the predicted box of the unlearned images intersected with the ground truth was less than 50, the prediction failed. The four conditions for classification were as follows: true positive, true negative, false positive, and false negative. We evaluated the performance of the model in terms of accuracy (Eq. 1), average precision (Eq. 2), average recall (Eq. 3), and F1-Score (Eq. 4). Finally, we used the average precision metric of Pascal VOC 2012. We calculated the classification mean to obtain the mAP.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

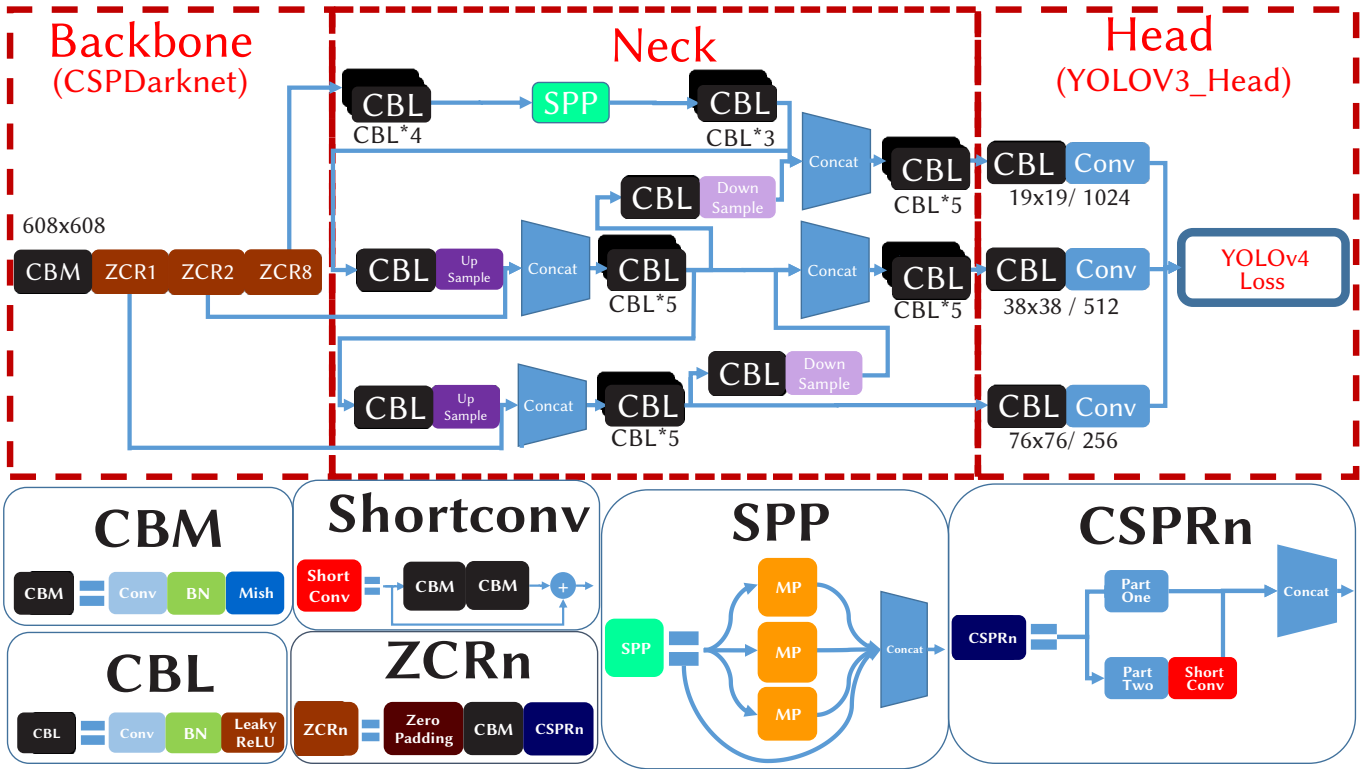


Fig. 3. YOLOv4 framework [17].

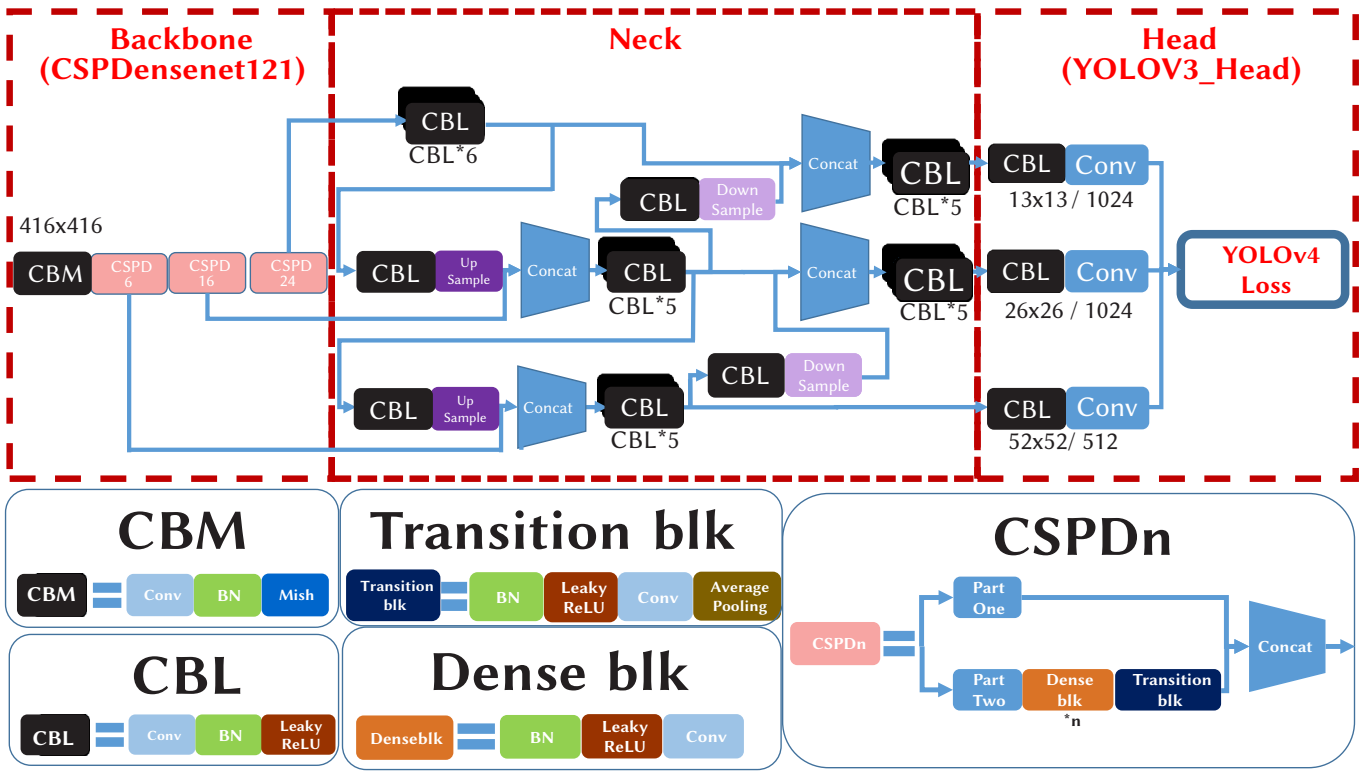


Fig. 4. Proposed algorithm: Modified YOLOv4-Densenet framework.

IV. Empirical Results

This research collected 483 images of Kaohsiung Veterans General Hospital. There are 67, 129, and 287 images for type 1, type 2, and type 4 of VSD, respectively. These figures are further divided into train, validation, and test sets. The dataset distribution is shown in Fig. 5.

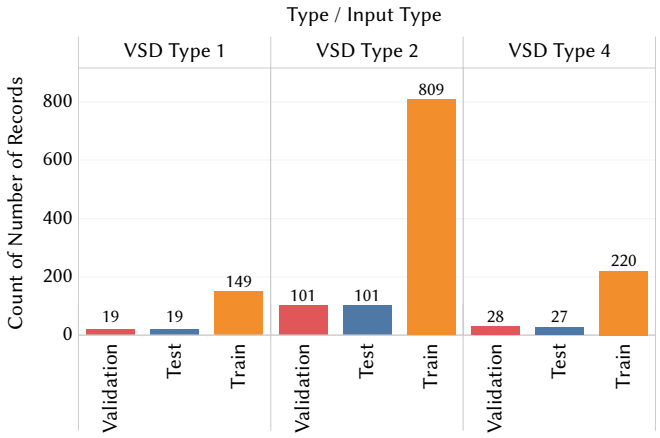


Fig. 5. Dataset arrangement PDF.

Based on a YOLOv3 project on Github[2], we code the YOLOv3-SPP [21], YOLOv3-DenseNet [23], [24], YOLOv3-DenseNet-SPP, YOLOv4 [17], and revised YOLOv4-DenseNet framework by ourselves. Each algorithm runs 1000 epochs with three replications. We executed these algorithms on Tensorflow 1.15.3 environment and nVidia RTX 2080 GPU to experiment. The parameters of the proposed revised YOLOv4-DenseNet are shown as follows. The optimization algorithm is Adam, with the learning rate 1e-4. The number of epochs is 1,000. Due to the limitation of the GPU memory, the number of batch size is set to 4 and the input resolution to be 416*416 for YOLOv4 instead of 608*608. Hence, we denote YOLOv4 to be YOLOv4' to distinguish the difference. Finally, we employed the latest model trained by each algorithm to do the following comparisons.

We list the precision, recall, F1-score, and mAP-50 of the six algorithms in Table I. When we compare the variants of YOLOv3, YOLOv3-DenseNet might be the best one, according to the F1-score and mAP-50. In particular, when we compare the YOLOv3 with YOLOv3-DenseNet, the F1-score is improved by 10%, and mAP is increased by 20%. The improvement is quite significant. Later on, SPP only improves the combination with YOLOv3 alone; however, SPP does not yield positive outcomes for YOLOv3-DenseNet-SPP because the variance of the mAP-50 values is high. That is the reason why our proposed algorithm does not include the SPP technique in the proposed algorithm.

TABLE I. Algorithm Models for the VSD

Algorithm	Precision(%)	Recall(%)	F1-score(%)	mAP-50
YOLOv3	99%	70%	81%	57.11%
YOLOv3-SPP	100%	77%	87%	59.54%
YOLOv3-Densenet	99%	84%	91%	71.56%
YOLOv3-Densenet-SPP	99%	83%	90%	70.56%
YOLOv4'	98%	71%	82%	58.42%
Revised YOLOV4-Densenet	97%	85%	91%	72.61%

When it comes to comparing the YOLOv4' and revised YOLOv4-DenseNet with the YOLOv3 variants, YOLOv4' is better than the YOLOv3. YOLOv4-DenseNet is the best one in terms of the result of F1-score and mAP-50. However, YOLOv3-DenseNet remain outperforms YOLOv4'. YOLOv4-DenseNet might be promising because this algorithm inherits the merit of YOLOv4, DenseNet [22] captures more information, and we remove some techniques that may decrease the solution quality. Most important of all, our proposed algorithm is assisted by CSPDenseNet as the backbone. This strategy enhances the prediction quality.

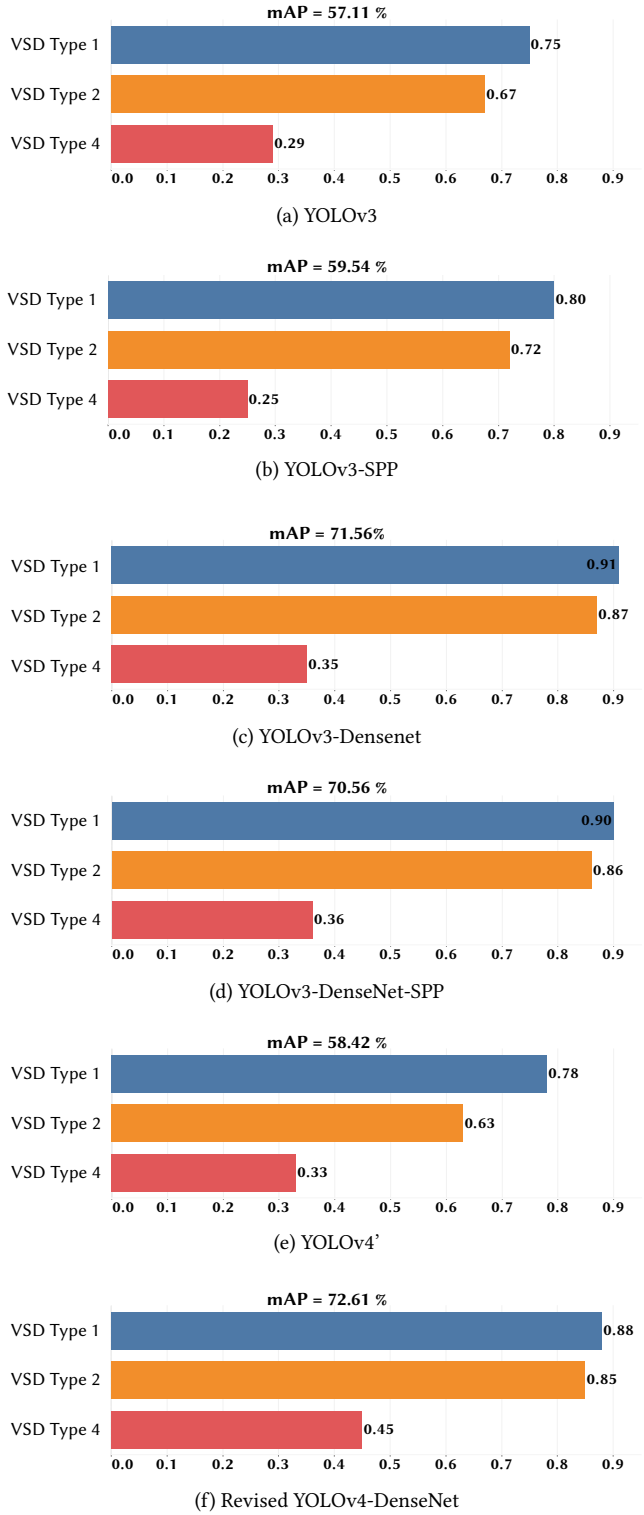
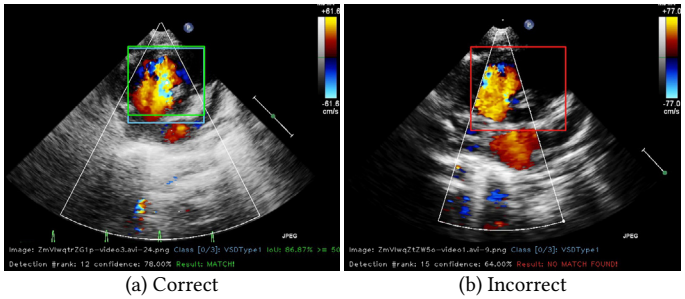


Fig. 6. The mAP performance of the compared algorithms.

[2] https://github.com/qqwweee/keras-yolo3

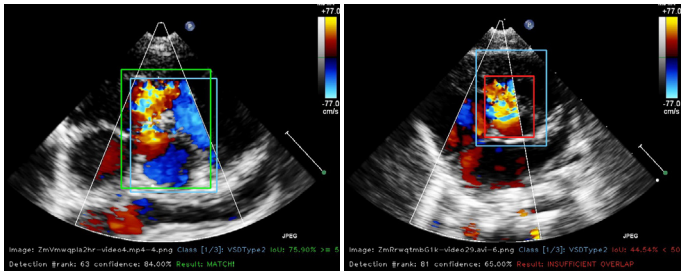
On the other hand, to explore the difficulty level of three VSD diseases, we draw the mAP details of the six selected algorithms in Fig. 6. Among all the algorithms, VSDType1 and VSDType2 achieve satisfactory results. However, VSDType4 does not perform well. For this issue, there is much room for the VSDType4 because most algorithms do not perform well. It is interesting to take a closer look at the correct and incorrect classifications.

In general, the locations of VSD Type 4 are distributed at variable sites within the muscular ventricular septum. Hence, it is good challenge for deep learning algorithms. To explain the other possible reason, we demonstrate the correct and incorrect detection in Fig. 7 to Fig. 9, which presented the three types done by our revised YOLOv4-DenseNet. The blue bounding box is the ground truth marked by us. The bounding box in green means the correct prediction done by the proposed algorithm. Otherwise, the bounding box color is in red. Except the Fig. 7b does not detect the VSDType1 at all, the bounding box in Fig. 8b and Fig. 9b might be too small. This problem may cause the result is not satisfactory. In addition, the symptoms of VSD Type 4 are discovered at varied places. Hence, it is quite necessary to increase the number of training dataset for VSD Type 4.



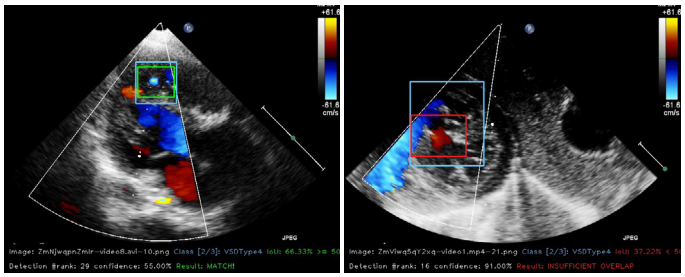
(a) Correct (b) Incorrect

Fig. 7. The correct and incorrect detection of VSDType1.



(a) Correct (b) Incorrect

Fig. 8. The correct and incorrect detection of VSDType2.



(a) Correct (b) Incorrect

Fig. 9. The correct and incorrect detection of VSDType4.

There are three ways of improving this situation. Firstly, we should increase the number of figures to train the deep learning model. Secondly, due to we set the threshold of IoU to be 50, a smaller detected bounding box yields the incorrect judgment. If we increase the IoU threshold, the mAP result should be increased. Secondly, the bounding boxes prepared by this research might be too large. As a result, we should revise the scale of the bounding boxes for the three types. In general, even though the mAP-50 result of VSDType4 is not satisfactory, it might remain useful for doctors to arrange the echocardiographic images.

V. Conclusions

This paper might be the first one to study the CHD in ultrasound image object detection problem. The revised YOLOv4-DenseNet algorithm is proposed in this paper. The reason for proposing the revised YOLOv4-DenseNet is that some features of YOLOv4 are not suitable for the ultrasound medical image, such as the mosaic data augmentation, SPP, and Cosine annealing scheduler. Later on, due to DenseNet could extract more features than DarkNet, we use CSPDenseNet as the backbone. The proposed algorithm was further compared with the origi-nal YOLOv3, YOLOV3-SPP, YOLOv3-DenseNet. We found the revised YOLOv4-DenseNet is the best in terms of the F1-Score and mAP-50. YOLOv4' is better than YOLOv3; however, YOLOv4' is not better than YOLOv3-DenseNet, and YOLOv3-DenseNet-SPP. These results indicated DenseNet as the backbone is effective.

For future research, we plan to improve the prediction quality of VSD Type 4 and study more CHDs, such as Atrial septal defect, Pulmonary stenosis, and Tetralogy of Fallot solved by our proposed algorithm. The hyper-parameters are not optimized; thus, we could use a genetic algorithm to do a global search. Finally, there are some ways of improving the medical image qualities [15], [25], [26]. when the Contrast Limited Adaptive Histogram Equalization (CLAHE) was applied on COVID-19 in chest X-Ray images, the detection accuracy was greatly improved [15]. We attempt to employ the CLAHE (supported in OpenCV) in ultrasound images without modifying the proposed algorithm.

Acknowledgment

The data used in this study are restricted by the Research Ethics Review Committee of the Kaohsiung Veterans General Hospital with the number 19-CT8-10(190701-2) to protect participant privacy. We thank the Ministry of Science and Technology for supporting this research with ID MOST 108-2221-E-230-004. We thank the co-author of YOLOv4, Dr. Chien-Yao Wang of the Institute of Information Science, Academia Sinica, Taiwan (R.O.C), clarified their proposed algorithm of CSPDenseNet Ref.-PRN [20]. Finally, we also thank Miss. Wen Mei of Kaohsiung Veterans General Hospital, Miss. Chin Yu worked on the patience data collection and medical images preparations, and Miss. Yu-Chi Lin further enhanced the Fig. 1 quality.

References

[1] M.-H. Wu, H.-C. Chen, C.-W. Lu, J.-K. Wang, S.-C. Huang, S.-K. Huang, "Prevalence of congenital heart disease at live birth in taiwan," *The Journal of pediatrics*, vol. 156, no. 5, pp. 782–785, 2010.

[2] S.-J. Yeh, H.-C. Chen, C.-W. Lu, J.-K. Wang, L.-M. Huang, S.-C. Huang, S.-K. Huang, M.-H. Wu, "National database study of survival of pediatric congenital heart disease patients in taiwan," *Journal of the Formosan Medical Association*, vol. 114, no. 2, pp. 159–163, 2015.

[3] J. Carvalho, L. Allan, R. Chaoui, J. Copel, G. DeVore, K. Hecher, W. Lee, H. Munoz, D. Paladini, B. Tutschek, *et al.*, "Isuog practice guidelines (updated): sonographic screening examination of the fetal heart," *Ultrasound in Obstetrics & Gynecology*, vol. 41, no. 3, pp. 348–359, 2013.

[4] M. Avendi, A. Kheradvar, H. Jafarkhani, "A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri," *Medical image analysis*, vol. 30, pp. 108–119, 2016.

[5] H. Chen, Y. Zheng, J.-H. Park, P.-A. Heng, S. K. Zhou, "Iterative multi-domain regularized deep learning for anatomical structure detection and segmentation from ultrasound images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 487–495, Springer.

[6] R. P. Poudel, P. Lamata, G. Montana, "Recurrent fully convolutional neural networks for multi-slice mri cardiac segmentation," in *International Workshop on Reconstruction and Analysis of Moving Body Organs*, 2016, pp. 83–94, Springer.

[7] G. Sutherland, M. Stewart, K. Groundstroem, C. Moran, A. Fleming, F. Guell-Peris, R. Riemersma, L. Fenn, K. Fox, W. McDicken, "Color doppler myocardial imaging: a new technique for the assessment of myocardial function," *Journal of the American Society of Echocardiography*, vol. 7, no. 5, pp. 441–458, 1994.

[8] P. Pézard, L. Bonnemains, F. Boussion, L. Sentilhes, P. Allory, C. Lepinard, A. Guichet, S. Triau, F. Biquard, M. Leblanc, *et al.*, "Influence of ultrasonographers training on prenatal diagnosis of congenital heart diseases: a 12-year population-based study," *Prenatal diagnosis*, vol. 28, no. 11, pp. 1016–1022, 2008.

[9] G. Hill, J. Block, J. Tanem, M. Frommelt, "Disparities in the prenatal detection of critical congenital heart disease," *Prenatal diagnosis*, vol. 35, no. 9, pp. 859–863, 2015.

[10] C. P. Bridge, C. Ioannou, J. A. Noble, "Automated annotation and quantitative description of ultrasound videos of the fetal heart," *Medical image analysis*, vol. 36, pp. 147–161, 2017.

[11] F. C. Ghesu, E. Krubasik, B. Georgescu, V. Singh, Y. Zheng,J. Hornegger, D. Comaniciu, "Marginal space deep learning: efficient architecture for volumetric image parsing," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1217–1228, 2016.

[12] A. Madani, R. Arnaout, M. Mofrad, R. Arnaout, "Fast and accurate view classification of echocardiograms using deep learning," *npj Digital Medicine*, vol. 1, no. 1, p. 6, 2018.

[13] M. Moradi, Y. Guo, Y. Gur, M. Negahdar, T. Syeda-Mahmood, "A cross-modality neural network transform for semi-automatic medical image annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2016, pp. 300–307, Springer.

[14] J. C. Nascimento, G. Carneiro, "Multi-atlas segmentation using manifold learning with deep belief networks," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, 2016, pp. 867–871, IEEE.

[15] F. A. Saiz, I. Barandiaran, "Covid-19 detection in chest x-ray images using a deep learning approach," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 6, no. 2, pp. 11–14, 2020.

[16] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[17] A. Bochkovskiy, C.-Y. Wang, H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, "Focal loss for dense object detection," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[19] S. Ren, K. He, R. Girshick, J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[20] C.-Y. Wang, H.-Y. Mark Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, I.-H. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 390–391.

[21] J. Redmon, "Darknet: Open source neural networks in c." http://pjreddie.com/darknet/, 2013–2016.

[22] H. Gao, Z. Liu, L. van der Maaten, K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[23] D. Xu, Y. Wu, "Improved yolo-v3 with densenet for multi-scale remote sensing target detection," *Sensors*, vol. 20, no. 15, p. 4276, 2020.

[24] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, Z. Liang, "Apple detection during different growth stages in orchards using the improved yolo-v3 model," *Computers and electronics in agriculture*, vol. 157, pp. 417–426, 2019.

[25] Z. Liao, M. H. Jafari, H. Girgis, K. Gin, R. Rohling, P. Abolmaesumi, T. Tsang, "Echocardiography view classification using quality transfer star generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2019, pp. 687–695, Springer.

[26] Q. Nie, Y.-b. Zou, J. C.-W. Lin, "Feature extraction for medical ct images of sports tear injury," *Mobile Networks and Applications*, pp. 1–11, 2020.

Shih-Hsin Chen

S. H. Chen received his Ph.D. in Industrial Engineering and Management from YuanZe University, Taiwan (R.O.C) in the year 2008. His research interests include medical image classification and object detection, artificial intelligence, and optimization scheduling problem. There are 30 SCI journal publications with more than 1280 citations by Google Scholar. He is currently the associate professor of the Department of Information Management and to be a group lead of Information Technology at Cheng Shiu University.

Chun-Wei Wang

He has been a software engineer for five years and is currently a master's degree student at the Department of Information Management, Cheng Shiu University. His master's degree thesis studied the ultrasound images of congenital heart diseases, including Ventricular Septal Defect, Atrial septal defect, Patent ductus arteriosus, Pulmonary stenosis, and Tetralogy of Fallot. This research goal will be completed by the end of June in the year 2021.

I-Hsin Tai

Dr. I-Hsin Tai received his medical doctor degree in 2010 from the National Defense Medical Center, Taipei. Dr. I-Hsin Tai received pediatric cardiology training in Kaohsiung Veterans General Hospital Kaohsiung Chang Gung Memorial Hospital, respectively, and continuously served in Kaohsiung Chang Gung Memorial Hospital participated in a research program of Kawasaki Disease Center. Dr. Tai had won the 2015 & 2016 original cardiology abstract price given by the Taiwan Society of Ultrasound in Medicine. He now moves to China Medical University Children's Hospital, Taichung, where he served as a pediatric cardiologist emergency physician to take care of critically ill children.

Ken-Pen Weng

Ken-Pen Weng, M.D. has completed his bachelor of medicine at the age of 25 years old from National Yang-Ming University, Taipei, Taiwan. He is the director of congenital/structural heart disease center, Kaohsiung Veterans General Hospital, Taiwan (R.O.C). He has published more than 50 papers in reputed journals.

Yi-Hui Chen

She received her Ph.D. degree in computer science and information engineering at the National Chung Cheng University. Later on, she worked at Academia Sinica as a post-doctoral fellow. Later, she worked at IBM's Taiwan Collaboratory Research Center as a Research Scientist. After that, she worked at the Department of M-Commerce and Multimedia Applications, Asia University. She is now an associate professor at the Department of Information Management, Chang Gung University. Her research interests include data mining, semantic analysis, and multimedia security.

Kai-Sheng Hsieh

Kai-Sheng Hsieh, MD, FACC, FESC, FCCM was born in Taipei, Taiwan. He received a B.S. degree in 1975 and an M.B. degree in 1978 from National Defense Medical Center, Taipei. Prof. Hsieh was trained as a clinical fellow in pediatric cardiology at Boston Children's Hospital and Harvard Medical School between 1982-1984. He was chief of pediatric cardiology at the Department of Pediatrics, Taipei Veterans General Hospital in 1985. Between 1990-2014, he served as Chief of the department of pediatrics, Kaohsiung Veterans General Hospital. Between 2014-2018, he served as General Chairman of Pediatrics, Chang Medical System, Taiwan. He is currently a professor of Pediatrics, Shuang-Ho Hospital-Taipei Medical University. Prof. Hsieh is the author of 250 articles, and 15 book chapters. He has deeply involved in clinical research and clinical teaching. He was the winner of the "Major Devotion to Medical Care of Children in Taiwan" Award in year 2013. Prof. Hsieh also is exceptionally interested in biomedical engineering. He has supervised many theses from master and doctoral degree students in bioengineering fields.

Integration of Genetic Programming and TABU Search Mechanism for Automatic Detection of Magnetic Resonance Imaging in Cervical Spondylosis

Chun-Jung Juan[1,2,3], Chen-Shu Wang[4]*, Bo-Yi Lee[5], Shang-Yu Chiang[6], Chun-Chang Yeh[7]*, Der-Yang Cho[8], Wu-Chung Shen[1,2]

[1] Department of Radiology, School of Medicine, College of Medicine, China Medical University, Taichung, Taiwan, (R.O.C)
[2] Department of Medical Imaging, China Medical University Hospital, Taichung, Taiwan, (R.O.C)
[3] Department of Medical Imaging, China Medical University Hsinchu Hospital, Hsinchu, Taiwan, (R.O.C)
[4] Department of Information and Finance Management, National Taipei University of Technology, Taipei, Taiwan, (R.O.C)
[5] Department of Management Information System, National Cheng-Chi University, Taipei, Taiwan, (R.O.C)
[6] Graduate Institute of Biomedical Electronics and Bioinformatics, National Taiwan University, Taipei, Taiwan, (ROC)
[7] Department of Anesthesiology, Tri-Service General Hospital and National Defense Medical Center, Taipei, Taiwan, (R.O.C)
[8] Department of Neurosurgery, China Medical University Hospital, Taichung, Taiwan, (R.O.C)

Received 20 December 2021 | Accepted 21 July 2021 | Published 5 August 2021

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Cervical spondylosis is a kind of degenerative disease which not only occurs in elder patients. The age distribution of patients is unfortunately decreasing gradually. Magnetic Resonance Imaging (MRI) is the best tool to confirm the cervical spondylosis severity but it requires radiologist to spend a lot of time for image check and interpretation. In this study, we proposed a prediction model to evaluate the cervical spine condition of patients by using MRI data. Furthermore, to ensure the computing efficiency of the proposed model, we adopted a heuristic programming, genetic programming (GP), to build the core of refereeing engine by combining the TABU search (TS) with the evolutionary GP. Finally, to validate the accuracy of the proposed model, we implemented experiments and compared our prediction results with radiologist's diagnosis to the same MRI image. The experiment found that using clinical indicators to optimize the TABU list in GP+TABU got better fitness than the other two methods and the accuracy rate of our proposed model can achieve 88% on average. We expected the proposed model can help radiologists reduce the interpretation effort and improve the relationship between doctors and patients.

Keywords

Cervical Spondylosis, MRI, Genetic Programming, TABU Search, Automatic Detection.

DOI: 10.9781/ijimai.2021.08.006

I. Introduction

CERVICAL spondylosis is a kind of degenerative disease which widely occurs in middle and old aged patients [1]. Besides aging issue, it may be also caused by the structural change of the cervical discs and vertebrae or spinal injury. Cervical spondylosis is a typical change in the cervical spine aging process, and is the most common reason for degenerative changes with the spinal column [2]. The bony structures of the vertebral bodies, known as osteophytes or spurs, also grow with age, which can cause compression of the spinal cord and stenosis of neuroforamina. According to statistics, more than 85% of people above 60 years old are affected by cervical spondylosis. Patients with cervical spondylosis might have symptoms including soreness, pain in the neck and shoulders, numbness, sensory change, weakness or muscular atrophy of the upper limbs as well as unstable gait caused by involvement of the lower limbs.

In addition, according to the interview results with radiologists, the cause of cervical spondylosis is also closely related to a person's lifestyle, work, and poor posture. For illustration, long-term continuous use of electronic devices in an inappropriate posture may lead to cervical spine disease, causing neck pain or muscle fatigue. The prevalence of cervical spondylosis is even higher among those who have to use computers and those who have to flex their necks for a long time [3]. The viewpoint is consist with the reported statistics that indicated the excessive use of electronic products has increased the number of patients with cervical disease [4]. Therefore, in the recent decades, cervical spondylosis not only occurs in the elderly but also develops widely in the younger population. The age distribution of cervical spondylosis patients is much lower than before. Unfortunately,

* Corresponding author.
E-mail address: wangcs@ntut.edu.tw (C. S. Wang), anes2yeh@gmail.com (C. C. Yeh).

according to a report published by eMarketer, the penetration rate of the electronic devices in Taiwan is the highest in the world(73.4%), followed by Singapore (71.8%) and Korea (70.4%) [5][6]. What is known as "smartphone zombie" or "Smombie" has accelerated the cervical spondylosis in all age groups. Smombie-related cervical spondylosis has become a new worldwide disease that threatens the human beings [7]. Therefore, how to efficiently diagnose cervical spondylosis is becoming an important issue.

The human cervical spine consists of many different structures with many nerve endings evenly distributing on it that makes cervical spine a very complex structure. Diagnosis of cervical spondylosis could be achieved by four different imaging tools, including X-rays, computed tomography (CT), myelography, and magnetic resonance imaging (MRI). Among all of them, MRI is most preferred for its noninvasiveness, high soft tissue contrast, high resolution, and free of radiation exposure. It is capable of clearly distinguishing the vertebras, disc, cerebral spinal fluid (CSF), spinal cord, and fat so that it outperforms all other aforementioned imaging tools in diagnosing cervical spondylopathy. In addition, MRI is able to detect the edema and ischemic change of the spinal cord, which helps determine the severity of spinal cord injury. Comparing to other tools mentioned above, as MRI used for the cervical spondylosis diagnosis, a false positive error (we regarded cervical spondylosis as positive which the patients are more concerned about) is more harmless than a false negative error [8].

However, in spite of all the merits, the MRI takes radiologists (or experts from other fields) a lot of time for MRI image check and interpretation. While pursuing more detailed medical information (such as advanced imaging techniques, videos, and laboratory data), the data amount increases accordingly. That is definitely a big burden for radiologists and might further increase the risk of misinterpretation due to physician's fatigue. Moreover, the shortage of medial manpower remains a problem, and training a medical expert takes rather a long time. Even not for cervical spondylosis diagnosis, the medical experts (such as clinicians and medical examiners) need more assistance for such image tool usage. We believe information technology (IT), including machine learning or data analytics prediction, can help solve the problems. For illustration, there is a research that refers improving radiologists' performance by using artificial intelligence system [9].

Also, Lin uses deep learning for automated contouring of primary tumor volumes of nasopharyngeal carcinoma MRI. It has a positive impact on tumor control and patient survival [10]. Using IT technology to help MRI interpretation is feasible.

In the past, most researches used image recognition [11][12] or feature extraction [13][14] for MRI images process. Although the MRI technology for image recognition can achieve a good accuracy rate, the detection method is like a black box and it is difficult to interpret. Therefore, how to assist clinic select appropriate feature from MRI via algorithms is an important issue.

In the literature, meta-heuristic algorithms and neural networks are used to solve decision problems, some heuristic algorithms can find good solutions, but they take more time [15]. In this research, we need to establish a flexible feature selection model based on a variety of features of cervical spine data. The model is expected to identify appropriate feature accurately and implement quickly. Therefore, Genetic Programming should be a good candidate [16], and TABU can be added to accelerate the algorithm.

In this study, we proposed a prediction model to evaluate the cervical spine condition of patients by using MRI data. The prediction model provides radiologists and neuroradiologists with prediction results of patients' cervical spondylosis severity for their diagnosis reference. Furthermore, to ensure the computing efficiency of the proposed model, we adopted a heuristic programming known as genetic programming (GP) to build the core of refereeing engine. Additionally, to guarantee the output of optimal solution, we used referencing search algorithm [16]-[19] by combining the TABU search (TS) with the evolutionary GP. Finally, to validate the accuracy of the proposed model, we implemented experiments via using actual MRI image and compared our prediction results with radiologists' diagnosis to the same MRI image. We expected the proposed model can help radiologists reduce the interpretation effort and improve the relationship between doctors and patients.

II. The Conceptual Architecture and Diagnosis Process of the Prediction Model

As indicated in the report from the National Research Council (NRC) of the United States in 2011, a successful medical analyzing system should contain two components, First, appropriate strategies are used to design and collect disease-relevant information from patients. Second, the data analytics methodologies are used to establish a practical architecture for data analysis. The clinician can then take new patients' data to test and verify such architecture [20]. To establish a prediction model for cervical spondylosis diagnosis based on the concept above, the proposed framework integrated abundant information, which was generated by disease-related inspection mechanism, and established a flexibly and comprehensively intelligent assistant model.

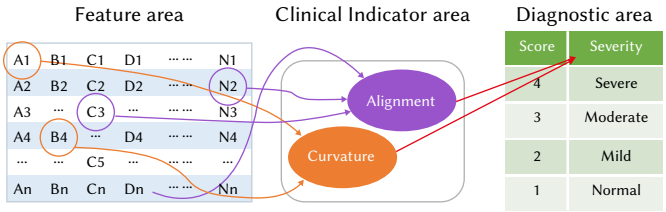


Fig. 1. The conceptual architecture of proposed predicted model.

As shown in Fig. 1, the predicted model is divided into three parts: the feature area, the clinical indicator area and the diagnostic area. In detail, the feature area collects data of the cervical magnetic resonance imaging which is measured by K-PACS. To measure the cervical spondylosis diagnosis, we adopted three indicators, including alignment, curvature and the aggregative diagnostic result which contains four severity levels. Finally, the diagnostic area reports the severity score as a reference for the doctor, enabling medical experts and patient to understand the diagnosis of cervical spondylosis. The severity score, ranging from 1 to 4, stands for the normal, mild, moderate and severe level of cervical spondylosis. To validate the proposed model, we cooperated with neuroradiologists and collected patients' MRIs of cervical spine with relative diagnosis results. The cervical vertebrae features required for the prediction model were measured under the guidance of the neuroradiologists. Then, these features were regarded as labels (named alignments, curvature, and severity) to establish and train the predictive model of cervical spondylosis. After training, the prediction model was applied to the test dataset and the accuracy with the diagnosis result of clinic was compared. The analytics process is detailed below.

A. The Analytics Process of the Proposed Prediction Model

To optimize the prediction performance, the proposed model integrated the TABU search algorithm with genetic programming (GP) to establish a prediction model. The analytics process is shown in Fig. 2.

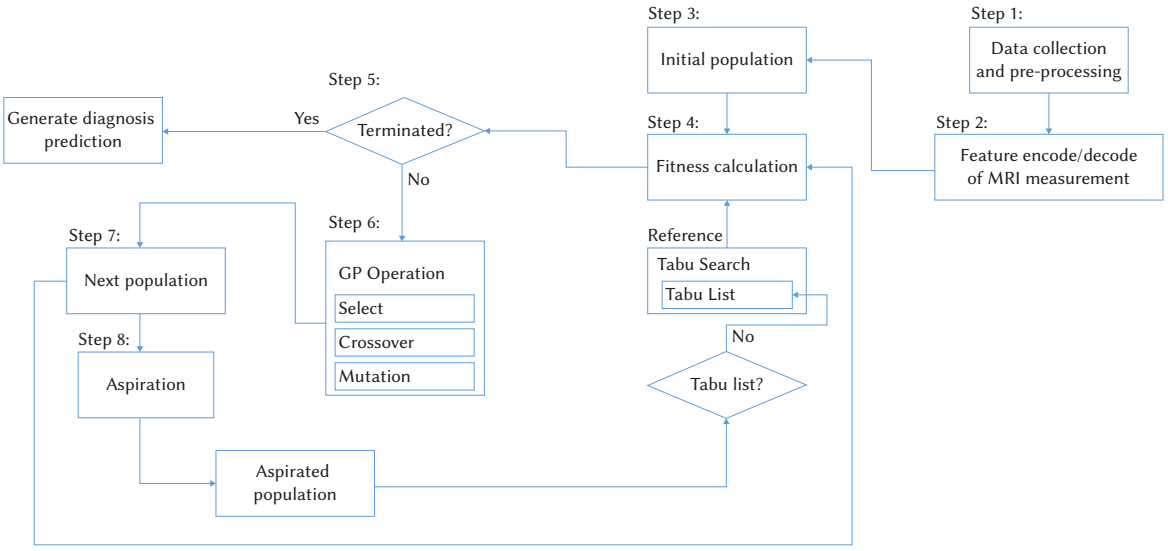


Fig. 2. Analytics process of the proposed prediction model.

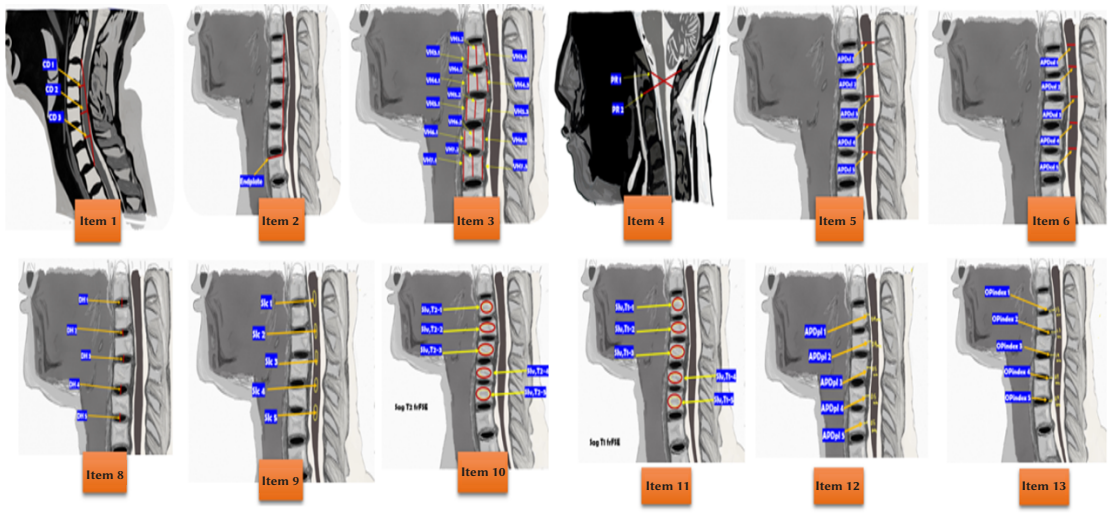


Fig. 3. Operational definition of features.

The analytics process consists of four sub-procedures, including data collection and pre-processing (step 1 in section II.B), feature encoding/decoding (step 2 in section II.C), integrating prediction model of GP (step 3 to step 6 in section II.D), with TABU search to optimize prediction result (step 6 to step 8 in section II.E). Moreover, we will further explain step 3 initial population in Fig. 7 and Fig. 8 and explain the GP process in Fig. 9. The sub-procedures are detailed from section II.B to II.E.

B. Data Collection and Pre-processing

The image data of MRI was collected from the patients' original cervical vertebrae MRIs, which were generated at the division of Magnetic Resonance Imaging in the Tri-Service General Hospital, as shown in Fig. 3 for demonstration. With high-resolution images, MRI enables radiologists to accurately diagnose cervical spine diseases and determine the extent of cervical cord damage. We adopted sagittal fast spin echo T2-weighted image and sagittal fast spin echo T1-weighted image mode to measure the cervical spine features. The images need to be clearly identified as the appropriate research data by the contours of the cervical spine. The cervical vertebrae consist of seven segments called C1 to C7. However, the atlas' shape of C1 is different from others', which have no body and are fused with the C2. Therefore, according to the suggestions from clinicians and radiologists, the measurement of cervical spine in the present study is based on C2 to C7.

In order to clearly identify the appropriate research data by the contours of the cervical spine, so that eliminated 6 images which were not clear or had missing part. The measurement features of the study are 67 features in 13 items. After image measurement and data processing, study had 147 cases of data for further analysis.

For data pre-processing, the first step is MRI image interpretation. By checking MRI, radiologists determine patients' cervical spine either normal or minor cervical spondylosis. Next, based on the definitions, we measured the cervical spine features for further model establishment. These features include curvature distance, vertebral body height and so on. However, the original images in the observation of the cervical spine may be hard to be interpreted (which may be caused by the poor resolution or angle) and each image may have different levels of definition as we measured. Therefore, we used the software to scale the images and selected the clearest MRI pictures according to the measure definition, making the collected data more accurate. The description and encoding/decoding of each feature are discussed in section II.C.

C. Feature Encoding/Decoding of MRI Measurement

To achieve higher accuracy prediction model, we collected a large number of cervical MRI images from the patients and used cervical anatomical position as the baseline when measuring. We adopted the software K-PACS to measure 67 features and then divided these features into 13 items according to their class as shown in Table I. The ways of measurement and each item's operational definition are also displayed in Table I and Fig. 3 respectively. In Table I, we could find a huge amount of information which clinicians should read, understand, make diagnosis with, and explain to patients' and their family members. Usually, to ensure the diagnosis accuracy, clinicians have to spend a lot of time dealing with this huge MRI information. Therefore, they tend to have limited time to communicate with their patients. To improve the doctor-patient relationship, clinicians need more support and assistance, such as a prediction model, to help them interpret MRI information.

TABLE I. The Description and Encode/Decode of Features of MRI Measurement

Item	Description	Encode as	Shown in Fig. 3
Item 1	Curvature distance of cervical vertebrae	CD1, CD2, CD3	(a)
Item 2	Anteroposterior diameter of superior endplate of cervical vertebrae	Endplate	(b)
Item 3	Vertebrae height	VH3.1-3、VH4.1-3、VH5.1-3、VH6.1-3、VH7.1-3	(c)
Item 4	Power ratio (The distance of cranial basis to posterior arch / The distance of occiput posterior to anterior arch)	PR1, PR2, PR3	(d)
Item 5	Antero posterior diameter of cervical canaL	APDcl1-5	(e)
Item 6	Antero posterior diameter of cervical cord	APDcd1-5	(f)
Item 7	Item 5(APDcl) / Item 6(APDcd)	APDd1-5	(g)
Item 8	Disk height	DH1-5	(h)
Item 9	Singal intensity of cervical cord	SI1-5	(i)
Item 10	Sagittal fast spin echo T2-weighted image (vetebral signal intensity)	SIvT2-1~SLVT2-5	(j)
Item 11	Sagittal fast spin echo T1-weighted image (vetebral signal intensity)	SIvT1-1~SLVT1-5	(k)
Item 12	Anteoposterior diameter of posterior longitudinal ligment	APDpl1-5	(l)
Item 13	Posterior disk herniation index	OPindex1-5	-

Additionally, for solution encoding, in this study, we adopted genetic programming (GP) as our heuristic search methodology. In the calculation process of GP, it evolves a more appropriate solution via each generation convergence. GP presented the solution as a "tree'' structure, as shown in Fig. 4. Each tree represents a predictive solution from the prediction model that represents a diagnosis path for the cervical spondylosis prediction. The tree structure has three parts: the roots (e.g., X in Fig. 4), the stems (e.g., -, /, +, or Exp in Fig. 4), and the leaves (e.g., parameters, the cervical vertebrae feature in Table I). This encoding approach is like gene sequencing that produced the best tree of prediction model after several generations of successive evolutionary processes.

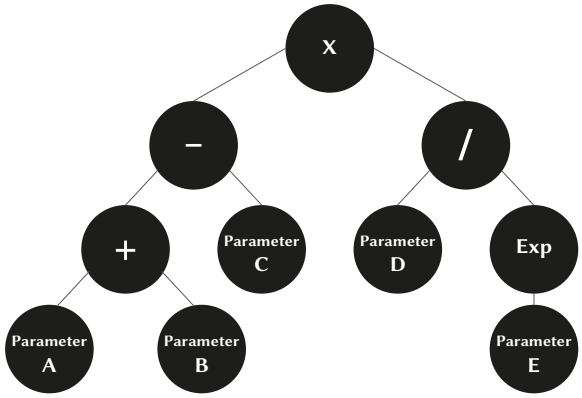


Fig. 4. The tree structure of genetic programming.

Finally, for the solution encoding and decoding, the GP method used LISP language to express the binary tree structure. As shown in Fig. 5, the tree's encoding was presented as. root : C ; left : A / C,H / 1,4,40 / ; right : E / F,C / 32,27,33 /. The tree's decoding was presented as ((1,4,C)(40,H)A)((32,F)(27,33,C)E)C. This decode way is also called postfix.

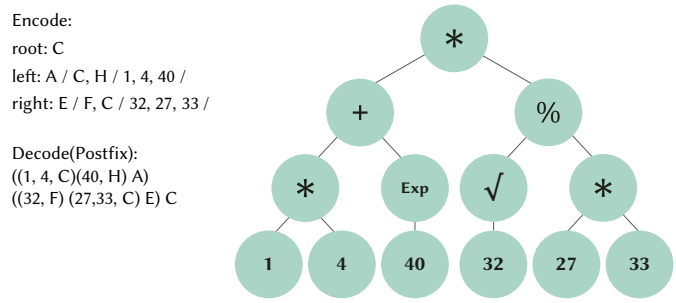


Fig. 5. The tree encoding and decoding schematic diagram.

D. Establish GP-based Forecasting Model

To ensure high accuracy of GP-based prediction model, we set the initial populations and then selected the tree with high fitness to multiply through the crossover and mutation. We repeated above process until reaching the terminal condition. Furthermore, we added TABU Search (TS) restriction and reward mechanism such as TABU list, candidate list and aspiration criterion in the process. Each node in the tree represents the terminal node (the related features of cervical magnetic resonance imaging), shown in Fig. 6A, and the function node, shown in Fig. 6B.

As the collected cervical features input, at first, the initial population was generated as a tree structure randomly. An operator (as A-E shown in Fig. 5.B) was chosen as the root node. The tree was generated from top (the roots) to bottom and left to right. Besides the operator node F to node I, for the function nod from operator code A to the operator code E, we needed two child nodes from terminal node as shown in Fig. 5A. As the depth of the tree reached the threshold from top to the bottom, it took any variable or constant from the terminal node collection (in this study it was the cervical vertebrae feature) to end the tree's expansion. The process of building the initial population is shown as Fig. 7 and the tree-expansion algorithm is shown in Fig. 8.

Parameter	Code	Parameter	Code	Parameter	Code	Parameter	Code	Parameter	Code
CD1	1	VH6.3	16	APDcd4	31	SIc4	47	APDpl4	62
CD2	2	VH7.1	17	APDcd5	32	SIc5	48	APDpl5	63
CD3	3	VH7.2	18	APDd1	34	Slv,T2-1	49	OPindex1	64
Endplate	4	VH7.3	19	APDd2	35	Slv,T2-2	50	OPindex2	64
VH3.1	5	PR1	20	APDd3	36	Slv,T2-3	51	OPindex3	65
VH3.2	6	PR2	21	APDd4	37	Slv,T2-4	52	OPindex4	66
VH3.3	7	PR3	22	APDd5	38	Slv,T2-5	53	OPindex5	67
VH4.1	8	APDcl1	23	DH1	39	Slv,T1-1	54		
VH4.2	9	APDcl2	24	DH2	40	Slv,T1-2	55		
VH4.3	10	APDcl3	25	DH3	41	Slv,T1-3	56		
VH5.1	11	APDcl4	26	DH4	42	Slv,T1-4	57		
VH5.2	12	APDcl5	27	DH5	43	Slv,T1-5	58		
VH5.3	13	APDcd1	28	SIc1	44	APDpl1	59		
VH6.1	14	APDcd2	29	SIc2	45	APDpl2	60		
VH6.2	15	APDcd3	30	SIc3	46	APDpl3	61		

(A)

Operator	Code
+	A
-	B
*	C
/	D
%	E
√	F
%100	G
Exp	H
1/x	I

(B)

Fig. 6. (A) terminal node and (B) function node for GP tree structure.

```
1  Environment parameters {
2        Input train_data (patient, diagnosis from doctor)
3        init GP operators_set
4        Set GP parameters (random level, mutation rate, crossover
         rate, population size...)
   }
5  For (i ≤ population size) {
6        Set tree root;
7        Left-Tree=Creat-Sub-Tree(Tree-Level-1);
8        Right-Tree=Creat-Sub-Tree(Tree-Level-1);
   }
```

Fig. 7. The process of building the initial population.

```
1  Create-Sub-Tree (int Tree-Level) {
2        set parameters (random level, operator)
3        if (i<level) {
4        newtree add operator as node
5        determine number of child node by operator type }
6        else if (i=level) {
7        newtree add cervical vertebrae feature as node }
8  output newtree }
   }
```

Fig. 8. The process of GP-tree expansion.

1. Fitness Function Mechanism

To ensure the calculation efficiency as dealing with large amount of MRI data, the fitness function compare the physician's diagnosis results with prediction results. The value Xn was calculated by the model (which represents the prediction result of the nth patient) and was compared with the physician-diagnosed value Yn (which represents the diagnosis result by physician of the nth patient), where n is the total patients. The prediction model regarded the comparison gap as fitness value for next generation revaluation reference. When the absolute value, resulting from Xn minus Yn, was equal to 0, the fit pulsed 1 that we try to reverse gap to fitness value. Finally, we counted the fit, divided it by n, and took the percentage as the fitness value of population. The formula is shown as follows:

$$fit(n) = |Xn - Yn| \quad if \ |Xn - Yn| = 0, \ fit + 1 \ \ else \ fit + 0 \tag{1}$$

$$Fitness = \frac{\int_1^n fit(n)}{n} \text{ x } 100 \tag{2}$$

The survival probability of individual population depended on the fitness value which was obtained by the fitness function. Therefore, according to the value of the fitness function, the higher fitness value represents, the better accuracy rate that compare with physician-diagnosed results. The correct rate was an important reference to create the next population of generation.

2. Genetic Programming Evaluation Process

GP evaluation required several kinds of parameters, including the population size, crossover rate, mutation rate and evolutionary generation, etc. The purpose of evolution was to produce a new and better solution. Thus, calculating the fitness value kept the quality of the parent generation for the next generation of evolution, and was expected to produce excellent final generation. After the fitness and diversity of the parental groups were determined, we started on the computing operations of evolution, which includes three stages: select, crossover and mutation. This evaluation process was repeated until a termination condition was reached. The detail of each stage describe as follow:

- *Select*: The purpose of this stage is survival probability of the fittest, through picking the parents with the highest fitness as elitism to stay until next generation and allow them to generate offspring with higher probability.
- *Crossover*: The purpose of this stage is increase populations diversity. A probability is generated randomly and compare to the crossover rate. As the probability is bigger than crossover rate, two selected parents' tree from population were exchanged their gene to generate two offspring as new children.
- *Mutation* : The purpose of this stage is to prevent the final solution falling into local optimization. Similar to crossover, a probability is generated randomly and compare to the mutation rate. As the probability is bigger than mutation rate, one the selected function node from the population changed the operator as a new child.

In the beginning, we sorted individual tree within population according to their fitness value from upper to lower. Using the elitism method, we reserved top 50% of these trees to survive directly in the next generation. Similar to elitism methodology, it can ensure optimization solution survival with higher probability and keep next population with certain steadily. Then, for the crossover operation, we put the selected trees into pairs. Each pair had a chance (according to mutation rate) of exchange in order to increase the evolution diversity. Crossover operator randomly selected the sub-tree to exchange (left to left, left to right, right to left and right to right) and generate new

children. Finally, similar to crossover operator, the mutation operator prevents the experiment from getting the locally optimal solution. There are two ways for mutation operation, which are node mutation and structural mutation. After the mutation, we put the evaluated tree into next generation and randomly created new tree to fulfill the size of the population which became a new generation. This evolution operation was repeated until it reached the termination condition.

3. Termination Criterion

The termination condition of this study is to reach the threshold of evolutionary generations. According to related research set 200 generations as threshold, we also found that setting more generations didn't improve accuracy, while setting less generations caused the lack of ethnic diversity. Therefore, this study observed the effect of the evolution of different populations within 200 generations. We selected the better results of the number of populations and checked the stability of different number of generations according to each number of generations after repeating ten times of experiments. After that, we unified different generations convergence effect to set the threshold for the termination of the evolution as 150 generations.

E. TS Mechanism

Genetic programming was a kind of randomly search so it didn't guarantee that the optimal solution can be obtained. For that reason, this study added TABU search as the auxiliary heuristic rule. The features of the previous generation were recorded in order to avoid making the result fall into local optimal solution. There were two ways to record. One was the short-term memory (TABU list); another was long-term memory (Candidate list). The operation of the TABU mechanism is shown in Fig. 9.

```
1  GeneticPrograming {
2      set parameter t as terminal condition
3      while (t < terminal iteration) {
4      select tree from long term memory as new tree
5      calculate fitness of new tree |
6      do Tabu record {
7      Record the poor adaptability of the features (short memory)
       Record the first optimal solution of 10 generations (long memory)
       }
8      evolution(select、crossover 、mutation)
9      Tabu list update by the relevant principles of disruption
10     t = t+1
11     Keep Tabu steps updating short and long memory
       until reached the teminal condition
12 }
```

Fig. 9. The operation of the TABU mechanism.

In a certain number of generations prohibited these populations been choosing in order to increase the diversity of evolution. TABU list had aspiration criterion to prevent the real superior solution from been prohibited. In addition, the long-term memory through establishing the candidate list to help generate a better new generation, using a more efficient way conducted a global search.

III. Experiments Design and Results Analysis

We adopted Visual C # to implement GP combined with TABU search were applied in this study. The cervical spine features with the relevant operator produced hundreds of tree-like models of each population. Repeating the evolutionary mechanism (selection, crossover, mutation) was conducted to seek the optimal solution that similar to clinical judgment. The TABU search method prevented GP from falling into the local best solution situation. Through the move, TABU lists and candidate lists increased the diversity of ethnic groups and strengthened the global search. Finally, the perdition model was summed up to assist neuroradiologists to diagnose in the future.

A. Experiment Parameters Setting

To validate the proposed model, the cervical spine MRIs were collected from Tri-Service General Hospital. At first, we selected 153 MRI cases of cervical spine disease and measure all features under the supervision of an neuroradiology's. However, it was necessary to use the Sagittal fast spin echo T2-weighted image and Sagittal fast spin echo T1-weighted image for measurement so that eliminated 6 images which were not clear or had missing part. After image measurement and data processing, we got 147 cases of data for analysis. We using picture archiving and communication system, for each patient, 67 features in 13 items were screenshotted and stored under neuroradiologist examination and correction. Furthermore, the third party (medical students) checked the measurement and confirmed the correctness.

Based on the proposed model shown in Fig. 2, the GP method was combined with TS method. By testing relevant parameter combination, including: suitable size of the populations, generations, evolution-related parameters, TS related conditions, we design two experiment series, including: evolutionary parameters combination testing (marked as experiment 1 in section 3.2) and proposed model validation (marked as experiment 2 in section 3.3).

B. Experiment 1: Evolutionary Parameters Combination Testing

We reference previous GP researches to set evolution parameters below. These parameters including: tree layers, select rate, crossover rate and mutation rate were set to 3-5, 0.5, 0.5 and 0.1 accordingly The terminal node and function node were list in Fig. 5.

Furthermore, to find out feasible population size, we tested four population size scale, such as: 50, 100, 150 and 200 as sub-experiment. Each sub-experiment was repeated for 20 generations. The average fitness value and standard deviation for each sub-experiment were shown in Fig. 10. According to these sub-experiments, we found that the highest average of the fitness value occurred when the population size was 150. And its standard deviation was also lower than the other population size which represented the evolution of P150 was stable. Therefore, we set the 150 as populations size for the following experiments. Finally, for the last parameter, generation, the maximum of 200 evolutionary generations was set. These parameters were further used for the proposed model validation in experiment 2.

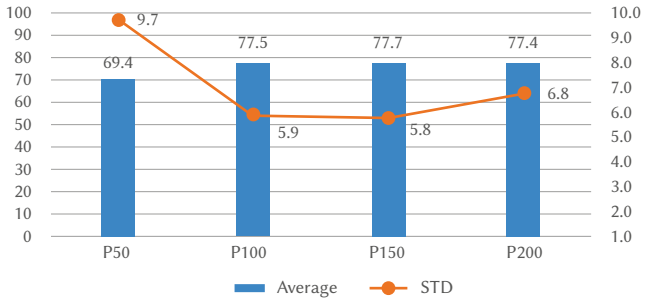


Fig. 10. The result of experienced four sizes of population.

C. Experiment 2: Cervical Curvature and Cervical Alignment Test Results

According to the proposed model in Fig. 2, we conducted a two-stage experiment. In Stage 1, we used GP method combined with TS

to predict the degree of curvature of the clinical indicators and the degree of alignment. In Stage 2, we put the relevant features in the clinical indicators into the TS mechanism and check the convergence performance that improved by TS mechanism.

We compared the genetic planning method (GP), genetic programming method combined with TABU search method (GP + TABU) and used clinical indicators to optimize the TABU list in GP+TABU (GP + TABU + Refinded_list). We compare these three experimental methods via ten tests. The cervical spine severity data was divided into training set and test set. We took 70% of the data as a training set, the remaining 30% for the test set. The experimental results were summarized in Table II. The experiment found that using clinical indicators to optimize the TABU list in GP+TABU got better fitness than the other two methods that the accuracy rate of our proposed model can achieve 88% on average.

TABLE II. Experimental Results

Method / Test times	GP	GP+TABU	GP+TABU +Refinded_list
1	79%	83%	88%
2	73%	89%	89%
3	74%	89%	88%
4	87%	82%	90%
5	87%	77%	89%
6	81%	81%	88%
7	80%	74%	86%
8	73%	76%	89%
9	71%	80%	86%
10	89%	79%	89%

Furthermore, from 67 features, we intended to find out import cervical vertebrae for diagnosis reference. According to the experimental results, we summarized the features of item statistical results as shown in Fig. 11. We found top 3 features were vertebral height, vertebral signal intensity (Sagittal fast spin echo T2-weighted image) and vertebral signal intensity (Sagittal fast spin echo T1-weighted image), showing a significant effect for determining the severity of the cervical vertebrae. Also, we found the indexes of both cervical vertebrae's curvature distance and anteroposterior diameter of posterior longitudinal ligament were 0. This represented that the two items for the diagnosis of patients with cervical spine severity had lower impact. For clinician reference, these important feature indexes may reveal more clues for future diagnosis process.

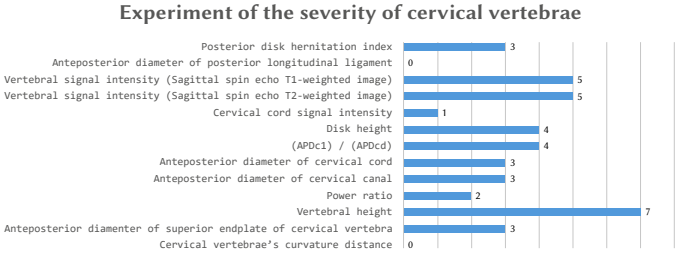


Fig. 11. The statistical results of features in experimental prediction model.

IV. Conclusions

Cervical spondylosis is a kind of degenerative disease which not only occurs in elder patients. According to previous researches, the prevalence of cervical spondylosis is even higher in those who have to use computers and those who have to flex their necks for a long time. The age distribution of cervical spondylosis patients is decreasing year by year. For now, MRI is the best tool to confirm the cervical spondylosis severity, despite of the fact of taking radiologists a lot of time for image check and interpretation.

In this study, we proposed a prediction model to evaluate the cervical spine condition of patients by using MRI data. Furthermore, to ensure the computing efficiency of the proposed model, we adopted a heuristic programming, genetic programming (GP), to build the core of refereeing engine and combined the TABU search (TS) with the evolutionary GP. Finally, to validate the accuracy of the proposed model, we implemented experiments and compared our prediction results with radiologists' diagnosis on MRI image. The experiment found that using clinical indicators to optimize the TABU list in GP+TABU would have better fitness than the other two methods and the accuracy rate of our proposed model can achieve 88% on average.

Furthermore, from 67 features, we found out import cervical vertebrae for diagnosis reference. We expected the proposed model can help radiologists reduce the interpretation effort and improve the relationship between doctors and patients. More case studies and model production can be our future works.

For future works, the production of the model could combine with some novel models for example Recurrent Neural Networks (RNN). RNN greatly relies on features and knowledge extracted from tasks [21], and its selections are motivated by what it has learned from the past or focus on the multi-target model [22] to predict. On the other hand, future works could also implement this feature extraction algorithm on medical CT images [23] .

Acknowledgment

Chen-Shu Wang received financial support of the Ministry of Science and Technology (Grant number: MOST109-2221-E027-072), Taiwan, R.O.C. Chun-Chang Yeh and Shang-Yu Chiang received financial support of Tri-Service General Hospital (Grant number: TSGH-C106-086) for this work. Chun-Jung Juan, Wu-Chung Shen and Der-Yang Cho received funding support partly from the China Medical University Hospital (Grant number: CMUH-DMR-108-056).

References

[1] B. M. McCormack, and P. R. Weinstein, "Cervical spondylosis. An update," *western Journal of Medicine*, vol. 165, no. 1-2, pp. 43, 1996, PMCID:PMC1307540.

[2] C. Wang, F. Tian, Y. Zhou, W. He, and Z. Cai, "The incidence of cervical spondylosis decreases with aging in the elderly, and increases with aging in the young and adult population: A hospital-based clinical analysis," *Clinical interventions in aging*, vol. 11, pp. 47, 2016, doi: 10.2147/CIA.S93118.

[3] A. I. Binder, "Cervical spondylosis and neck pain," *Bmj*, vol. 334, no. 7592, pp. 527-531, 2007, doi: 10.1136/bmj.39127.608299.80.

[4] L. Brain and M. Wilkinson, Ed., *Cervical spondylosis and other disorders of the cervical spine*, Oxford, United Kingdom: Butterworth-Heinemann, 2013.

[5] S. Y. Kim and S. J. Koo, "Effect of duration of smartphone use on muscle fatigue and pain caused by forward head posture in adults," *Journal of physical therapy science*, vol. 28, pp. 1669-1672, 2016, doi: 10.1589/jpts.28.1669.

[6] S. P. Cohen, "Epidemiology, diagnosis, and treatment of neck pain," *Mayo Clinic Proceedings*, vol. 90, no. 2, pp. 284-299, 2015, doi: 10.1016/j.mayocp.2014.09.008.

[7] Y. G. Kim, M. H. Kang, J. W. Kim, J. H. Jang, and J. S. Oh, "Influence of the duration of smartphone usage on flexion angles of the cervical and lumbar spine and on reposition error in the cervical spine," *Physical Therapy Korea*, vol. 20, pp. 10-17, 2013, doi: 10.12674/ptk.2013.20.1.010.

[8] M. Terry, "Campbell's operative orthopedics," *The Journal of the American Medical Association*, vol. 301, no. 3, pp. 329-330, 2009, doi: 10.1001/jama.2008.969.

[9] Y. Jiang, A. V. Edwards, and G. M. Newstead, "Artificial intelligence applied to breast MRI for improved diagnosis," *Radiology*, vol. 298, no. 1, pp. 38-46, 2021, doi: 10.1148/radiol.2020200292.

[10] L. Lin, Q. Dou, Y. M. Jin, G. Q. Zhou, Y. Q.Tang, W. L. Chen, ... and Y. Sun, "Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma," *Radiology*, vol. 291, no. 3, pp. 677-686, 2019, doi: 10.1148/radiol.2019182012.

[11] A. Paul, A. Paul, & P. B. Chanda, "Detection and Classification of Cervical Spondylosis Using Image Segmentation Techniques," *In Proceedings of Second National Conference*, Springer, Singapore, pp. 145-154.

[12] L. Zhang, , & H. Wang, "A novel segmentation method for cervical vertebrae based on PointNet++ and converge segmentation," *Computer Methods and Programs in Biomedicine*, vol. 200, 105798, 2021, doi: 10.1016/j.cmpb.2020.105798.

[13] C. S. Wang, C. J. Juan, T. Y. Lin, C. C. Yeh, and S. Y. Chiang, "Prediction Model of Cervical Spine Disease Established by Genetic Programming," *In Proceedings of the 4th Multidisciplinary International Social Networks Conference*, New York, United States, pp. 1-6.

[14] J. Ma, & X. Gao, "A filter-based feature construction and feature selection approach for classification using Genetic Programming," *Knowledge-Based Systems*, vol. 196, 105806, 2020, doi: 10.1016/j.knosys.2020.105806.

[15] N. Rokbani, R. Kumar, A. Abraham, A. M. Alimi, H. V. Long & I. Priyadarshini, "Bi-heuristic ant colony optimization-based approaches for traveling salesman problem," *Soft Computing*, vol. 25, pp. 3775–3794, 2021, doi: 10.1007/s00500-020-05406-5.

[16] R. S. Sexton, B. Alidaee, R. E. Dorsey, and J. D. Johnson, "Global optimization for artificial neural networks: A tabu search application," *European Journal of Operational Research*, vol. 106, no. 2-3, pp. 570-584, 1998, doi: 10.1016/S0377-2217(97)00292-0.

[17] T. Hou, J. Wang, L. Chen, and X. Xu, "Automated docking of peptides and proteins by using a genetic algorithm combined with a tabu search," *Protein Engineering*, vol. 12, no. 8, pp. 639-648, 1999, doi: 10.1093/protein/12.8.639.

[18] Q. Shen, W.-M. Shi, and W. Kong, "Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data," *Computational Biology and Chemistry*, vol. 32, no. 1, pp. 53-60, 2008, doi: 10.1016/j.compbiolchem.2007.10.001.

[19] X. Zhang, T. Wang, H. Luo, J. Y. Yang, Y. Deng, J. Tang, et al., "3D Protein structure prediction with genetic tabu search algorithm," *BMC systems biology*, vol. 4, S6, 2010, doi: 10.1186/1752-0509-4-S1-S6.

[20] N. R. Council, Ed., *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*, D.C., USA: National Academies Press, 2011.

[21] J. C. W. Lin, Y. Shao, Y. Djenouri, & U. Yun, "ASRNN: a recurrent neural network with an attention model for sequence labeling," *Knowledge-Based Systems*, vol. 212, 106548, 2021, doi: 10.1016/j.knosys.2020.106548.

[22] J. C. W. Lin, G. Srivastava, Y. Zhang, Y. Djenouri, & M. Aloqaily, "Privacy preserving multi-objective sanitization model in 6G IoT environments," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5340 – 5349, 2021, doi: 10.1109/JIOT.2020.3032896.

[23] Q. Nie, Y. B. Zou, & J. C. W. Lin, "Feature extraction for medical ct images of sports tear injury," *Mobile Networks and Applications*, vol. 26, pp. 404–414, 2021, doi: 10.1007/s11036-020-01675-4.

Chun-Jung Juan

M.D. at National Defense Medical Center, Taiwan, 1995, Ph.D. in Electrical Engineering at National Taiwan University, Taiwan, 2007. PhD in Computer Science and Information Engineering at National Taiwan University, Taiwan, currently in progress. Currently working as neuroradiologist and vice superintendent at China Medical University Hsinchu Hospital, and as professor and chief in Department of Radiology at China Medical University, Taiwan. Research interests are neuroradiology regarding ischemic and hemorrhagic stroke, head and neck radiology regarding cancer and post-radiation injury of salivary glands, spine radiology regarding spondylopathy, magnetic resonance imaging, computed tomography, radiography, artificial intelligence, machine learning, and deep learning.

Chen-Shu Wang

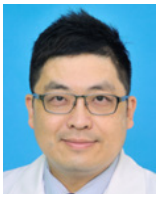
Chen-Shu Wang is now the professor of Department of Information and Finance Management at National Taipei University of Technology, Taiwan. She received Ph.D. of Department of Management Information System from Cheng Chi University in Taiwan. Her research interest is big data analytics, IT and AI applications, including: business intelligence, production analyze, and medical data analytics. Recently, Dr. Wang also devoted to MOOCs and course material development.

Bo-Yi Lee

Bo-Yi Lee is a PhD candidate in the Department of Management Information Systems at the National Cheng-Chi University, Taiwan. His research interests include IT adoption behavior, data mining, machine learning, smart home, etc.

Shang Yu Chiang

Shang Yu Chiang is a PhD student in the Biomedical Electronics and Bioinformatics from National Taiwan University. She holds a master's degree in lab" Big Data Analysis" of Information and Finance Management from National Taipei University of Technology. The main academic interests include Medical image processing and Big Data Analysis.



Chun-Chang Yeh

M.D. at National Defense Medical Center, Taiwan, 1993, Ph.D. in Biomedical Pharmaceutical Science at Fu Jen Catholic University, Taiwan, 2016. Currently working as anesthesiologist and chairman in Department of Anesthesiology and chief in Integrated Pain Management Center at Tri-Service General Hospital, and as associate professor in Department of Anesthesiology at National Defense Medical Center. Research interests are anesthesiology, critical care medicine, and interventional pain management.

Der-Yang Cho

M.D. at National Yang-Ming University, Taiwan, 1982, master's in Health Management at Asian University, Taiwan, 2005. Currently working as neurosurgeon and superintendent at China Medical University Hospital, and as professor in graduate institute of biomedical science at China Medical University, Taiwan. Research interests are neuroimmunology, basic neuroscience, cellular and molecular biology, stem cell biology, tumor immunology, pediatric neurosurgery, endoscopic neurosurgery, functional neurosurgery, epilepsy surgery, Gamma knife surgery, spinal surgery, and neurovascular surgery.

Wu-Chung Shen

M.D. at China Medical University, Taiwan, 1977. Currently working as neuroradiologist and consultant at China Medical University Hospital, as professor in Department of Radiology at China Medical University, and as director of China Medical University and medical system. Research interests are medical education regarding neuroradiology, magnetic resonance imaging, and computed tomography.

unir
LA UNIVERSIDAD
EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net