Prediction of COVID-19 Using a Clinical Dataset With Machine Learning Approaches

A. Suruliandi[1], R. Ame Rayan[1]*, S. P. Raja[2]

[1] Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, (India)
[2] School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, (India)

* Corresponding author: amerayanld@gmail.com

Received 20 August 2023 | Accepted 11 July 2024 | Early Access 29 January 2025

Abstract

COVID-19 is an infectious disease that spreads quickly from person to another. The pandemic, which spread worldwide over time, presents huge risks in terms of blood clotting, breathing problems and heart attacks, sometimes with fatal consequences if not detected early. The PCR test, CT scans, X-rays, and blood tests are methods commonly employed to detect the disease, though the PCR test is, without question, considered the gold standard. The American Center for Disease Control and Prevention (CDC) reports that the PCR has an 80% accuracy rate. An alternative to the PCR is clinical data, which is less expensive, easy to collect, and offers better accuracy. Machine learning, with its rich feature selection and classification methods, helps detect COVID-19 at the earliest stages, using clinical test results. This research proposes a clinical dataset and offers a comparative analysis of feature selection and classification algorithms for detecting COVID-19. Filter-based feature selection methods such as the ANOVA-F, chi-square, mutual information and Pearson correlation, along with wrapper-based methods such as Recursive Feature Elimination (RFE) and Sequential Forward Selection (SFS) were used to choose a subset of features from the feature set. The selected features were thereafter applied to the Support Vector Machine (SVM), Naïve Bayes, K-NN (K-Nearest Neighbor) and Logistic Regression(LR) classification algorithms to detect Coronavirus Disease. The experimental results of the comparative study show that the clinical dataset provides better accuracy at 94.8%, with mutual information and the SVM classifier.

Keywords

Blood Samples, Classification, COVID-19 Prediction, Feature Selection, Machine Learning.

DOI: 10.9781/ijimai.2025.01.003

I. Introduction

In December 2019, the novel coronavirus caused a public health crisis which spread rapidly worldwide. The disease is transmissible, striking healthy individuals who come in contact with droplets from an infected individual [1]. The infected person is sometimes asymptomatic, while others develop symptoms like cough, fever, shortness of breath and body pain, as well as loss of taste and smell. The Reverse Transcription Polymerase Chain Reaction (RT-PCR) has been followed as the gold standard in diagnosing COVID-19 [2]. Notwithstanding its popularity, the test has certain intrinsic flaws in that it is consumes more time, expensive, requires specially designed laboratory devices, and has a false negative rate of 20% [3]. Blood tests, X-rays, CT scans and breath sound analysis have been used as alternative procedures in COVID-19 diagnosis. Even though positive results are obtained using chest X-ray and CT scans images based on machine learning [4], the downside of these tests is exposure to high doses of radiation. Given that recent studies have shown that the blood features of COVID-19 patients change dramatically [5]–[10], hence early detection of the virus can be done by recognizing and working with these parameters. The blood tests results are ready in quick time and are relatively cheaper than other tests.

A decision support system is most useful in predicting COVID-19 using clinical data in the early stages so appropriate decisions can be made in good time. Clinical data includes biochemical parameters, obtained through blood tests that are made easily available in little time. The parameters include C-reactive protein (CRP), lymphocytes, DC:Neutrophils and D-Dimer, among others, which show changes due to coronavirus infection. As a result, the most common clinical findings, such as biochemical and hematological parameters, play an important role in COVID-19 preliminary screening [11].

Researchers in Artificial Intelligence use machine learning as a tool to assist healthcare workers diagnose disease. Machine learning classification and clustering algorithms give the best results when it comes to building such decision support systems. Machine learning provides algorithms that handle large datasets in a minimum runtime by selecting appropriate attributes. Further, it provides excellent

Please cite this article as:
A. Suruliandi, R. Ame Rayan, S. P. Raja. Prediction of COVID-19 Using a Clinical Dataset With Machine Learning Approaches, International Journal of Interactive Multimedia and Artificial Intelligence, (2024), http://dx.doi.org/10.9781/ijimai.2025.01.003

detection methods [12]. Machine learning models speed up information analysis and help make efficient disease prediction decisions. A model designed using machine learning recognizes patterns in blood samples and uses them to diagnose COVID-19. The objectives of this paper are (i) to propose a new clinical dataset to predict whether the person is affected by COVID-19 or not, and (ii) to find an optimal feature selection method and classifier for COVID-19 prediction.

II. Related Work

An analysis of the literature highlights similar work on COVID-19 prediction using blood test datasets. A tool was designed by Wu et al. [13] using the random forest algorithm to predict COVID-19. A total of 253 samples of data were collected for this purpose from 169 suspected patients. Each instance of data had 49 parameters, with 24 and 25 of those relating to the hematological and biochemical, respectively. In all, 11 parameters were extracted with the help of random forest algorithm. The overall performance of the tool in terms of accuracy in COVID-19 prediction was measured at 95.95%. Bastug et al. [14] undertook a comprehensive analysis of laboratory and clinical attribute for detecting COVID-19. The severity of the illness is predicted by training the model with the information from 191 coronavirus affected patients, admitted at an Ankara city hospital. In all, 29 blood routine parameter features were statistically analyzed. Kolmogorov-Smirnov test was used to check the normality of variables and to predict disease severity binary logistic regression was applied.

Brinati et al. [15] developed two machine learning models for COVID-19 detection. The study collected 279 blood samples from 177 COVID-19 positive and 102 COVID-19 negative individuals. The missing values were handled using Multivariate Imputation by Chained Equation (MICE) and feature Importance was used to select the best features. Five different machine learning algorithms, including the extremely randomized trees, logistic regression, decision tree, k-nearest neighbors, Naïve Bayes and random forest were compared to detect COVID-19. Of these, the two models designed using random forest algorithm outperformed with the accuracy of 82% and 86% respectively. Kukar et al. [16] used data from the University Medical Center in Ljubljana, Slovenia, to train a machine learning model to detect COVID-19. Blood samples from 5333 patients with viral and bacterial infections and 160 COVID-19 positive patients were included in the dataset. Feature selection was carried out using the feature importance scoring feature of the XGBoost machine learning algorithm, and the predictive model designed using the algorithm yielded an AUC of 0.97 in detecting COVID-19. Chadaga et al. [17] used data from Brazil's Albert Einstein Hospital to build a model for COVID-19 diagnosis. This model used SMOTE balancing technique to balance the dataset through oversampling, along with correlation analysis and feature importance to select the best features. The random forest, k-nearest neighbors, logistic regression, and XGBoost classifiers were compared using this dataset, and the best accuracy (92%) was produced by random forest algorithm.

Aljame et al. [18] proposed the "ER-CoV" machine learning model to predict the incidence of COVID-19 using hematological and demographic parameters. Data collected from 5644 patients of the Albert Einstein Hospital, Brazil, were preprocessed using the KNNImputer algorithm to handle null values and the SMOTE to balance the dataset. The SHAP technique was used to select 18 features from a total of 108. The proposed model has two level classifiers. The first level had the random forest, logistic regression and extra trees classifiers, and the output from this level was given as input to the second-level extreme gradient boosting classifier to detect COVID-19. The proposed model achieved 99.88% overall accuracy. The authors of [19] proposed a model using five ML algorithms such as gradient boost trees, SVM, logistic regression, neural networks, and random forest for the diagnosis of COVID-19. A dataset was created using the data collected from Brazil's Albert Einstein Hospital. The dataset contains 235 blood samples with 102 confirmed cases of COVID-19. From the dataset, 15 relevant characteristics were chosen for research. In this study, the SVM produced the best classification results with very little significance, with AUC, sensitivity, and specificity of 85%, 68%, and 85%, respectively, when compared to previous work. A study [20] analyzed and applied six state-of-the-art methods like the SVM, MLP, NB, RT, Bayesian Networks (BN), and RF to a dataset from the Brazil's Albert Einstein Hospital which consists of 564 samples, including 559 COVID-19 positive samples. The SMOTE was used for oversampling due to limited size of the dataset. Two PSO-based algorithms, the evolutionary search algorithm, and a manual method were all used for feature selection. The BN model performed the best overall, achieving accuracy, precision, specificity, and sensitivity values of 95.159%, 93.8%, 93.6%, and 96.8%, respectively.

Almansoor and Hewahi [21] collected Kaggle data with patient information from the Brazil's Albert Einstein Hospital, containing 5644 instances and 111 features. Data preprocessing was carried out using a one-sided selection technique to balance the data. The SVM, AdaBoost, random forest and k-nearest neighbour classifiers were used to detect COVID-19. Cabitza et al. [22] compared the performance of their model using the random forest, logistic regression, k-NN, SVM and Naïve Bayes algorithms. Three types of datasets, namely, the CBC, OSR, and a COVID-19-specific dataset were utilized. It was observed that the random forest and SVM performed the best with 88% accuracy for the OSR dataset, while the k-NN and SVM outperformed other algorithms on the COVID-19 specific dataset. The CBC dataset produced good results with the k-NN algorithm.

Akhtar et al. [23] used various machine learning algorithms like the k-NN, SVM, Naïve Bayes, multi-layer perceptron and decision tree to detect COVID-19 using the CBC dataset uploaded on the Kaggle website. The CBC dataset contains the CBC parameters of 5644 patients. Performance-wise, the multi-layer perceptron outclassed other algorithms. Abayomi-Alli et al. [24] introduced an ensemble learning model for COVID-19 detection using blood test samples. They combined custom convolutional neural networks (CNN) with 15 supervised machine learning algorithms. This ensemble model, incorporating DNN and ExtraTrees, achieved a remarkable accuracy of 99.28% and an AUC of 99.4% on the San Raffaele Hospital dataset, outperforming other COVID-19 diagnostic methods. Gong et al. [25] present a methodology for achieving explainable AI-driven rapid COVID-19 diagnosis. They employed ensemble learning algorithms to analyze data collected from 1,737 participants hospitalized at San Raphael Hospital during the period of February to May 2020. The study applied four distinct ensemble learning algorithms, namely random forest, adaptive boosting, gradient boosting decision tree (GBDT), and extreme gradient boosting (XGBoost). Notably, the GBDT model demonstrated superior performance, achieving an accuracy of 86.4% in effectively distinguishing COVID-19 patients from the control group. Roy and Singh [26] introduced a framework employing the weighted average of predictive accuracy from individual transfer learning models, including ResNet50V2, DenseNet201, and InceptionNetV3. The framework demonstrated exceptional performance in detecting COVID-19 from Chest X-ray images, achieving an impressive F1-score of 0.997.

Andueza et al. [27] used ARIMA and SARIMA Machine Learning models to predict the impact of COVID-19 on tobacco sales in Spain (January 2020 to December 2021) in euros, packs, and per capita packs. The study highlights a significant decline in cigarette sales, particularly in provinces popular among tourists and those sharing

borders with France. Sales during border closures were up to 66.74% lower than the initial forecasts, emphasizing the notable impact of COVID-19 restrictions on provincial tobacco sales in Spain. This suggests a disruption in the typical patterns of tourism and cross-border purchases between Spain and France, as well as Spain and Gibraltar. Cowley et al. [28] suggested a novel approach that integrates the outcomes of supervised random forest classification with unsupervised clustering to forecast patient risk. The model demonstrated superior performance, achieving an accuracy of 92%.

A. Motivation and Justification

COVID-19, a communicable disease that has spread throughout the globe, has such common symptoms that it has facilitated the publication of numerous open source clinical datasets. The literature review makes it plain that clinical parameters help in early screening of the disease. A person infected by the coronavirus shows variations in blood components. Given that the C-Reactive Protein (CRP), DC:Neutrophils, D-Dimer, and the lymphocytes vary rapidly from their normal values, they help identify infected individuals with early screening and have thus motivated the creation of an open source clinical dataset. The proposed clinical dataset contains COVID-19 patients' blood test results, and it is anticipated that it will help researchers develop a tool for an initial screening of the disease.

It is clear from the literature survey that the datasets used for the research have missing values and are imbalanced. The two issues are addressed by machine learning techniques such as the Multivariate Imputation by Chained Equation (MICE), KNNImputer, and SMOTE. The two factors above have motivated the creation of an open source clinical dataset with balanced data and with no missing values. Secondly, it is seen that feature selection is vital to improved performance. The two types of feature selection algorithm that is widely used are filter and wrapper based methods. The filter based method selects features with a score greater than the threshold. The wrapper method selects a subset of features for classification, following which the subset with the best accuracy is selected as the best feature set. Reducing the number of features by selecting optimal ones helps improve the performance of the model. This research is justified in that it carries out numerous experiments, using the proposed clinical dataset, to perform a comparison in the accuracy of the classifier with and without feature selection. Machine learning algorithms such as the ANOVA-F, chi-square, mutual information, Pearson coefficient, SFS and RFE are used for feature selection.

The literature review revealed that machine learning classifiers such as the random forest, XGBoost, logistic regression, extra trees, SVM, Naïve Bayes, and multilayer perceptron help predict the disease most accurately. This research uses the Naïve Bayes, SVM, k-NN and logistic regression to predict the disease using the features selected from the feature selection algorithms mentioned above. To summarize, machine learning algorithms may be used in prediction by training the dataset and incorporating the given input data with the trained data for classification. Most healthcare applications, therefore, use machine learning approaches for prediction.

Classification techniques such as the logistic regression, SVM, k-NN and Naïve Bayes are used to classify the selected features. The findings indicate that the performance of the classifier is enhanced by only using selected features that are picked following the application of feature selection. The prediction model, built with the selected features and classification algorithms in machine learning, produces good accuracy. This research work undertakes a comparative analysis of several feature selection and classification algorithms to discover the most effective feature set and classifier respectively, for detecting COVID-19 using the proposed clinical dataset.

B. Outline of the Work

The overall working of the research process is shown in Fig. 1. Firstly, the dataset is preprocessed to handle missing values, eliminate redundant values and convert categorical values into numerical values. Then the dataset is checked for outliers, the identified outliers are removed, and the dataset is balanced using SMOTE algorithm. Secondly, from the preprocessed dataset, significant features are selected using feature selection algorithm. Thirdly, the data in the selected features are subject to several classification techniques to identify the persons affected by COVID-19. Finally, based on the performance metrics of various classification techniques, best feature selection and classification algorithm is selected.

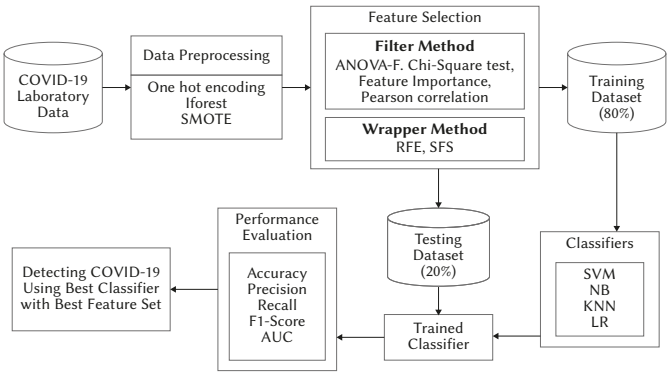

Fig. 1. Outline of the Work.

C. Organization of the Paper

The rest of this research paper is organized as follows. Section 3 discusses the methodology of the model. Section 4 describes the procedure for COVID-19 prediction. Section 5 presents the findings of the experiments and discusses them, while Section 6 concludes the paper and offers directions for future research.

III. Methodology

A. Proposed Dataset Construction

There are, currently, only a few clinically available COVID-19 datasets which, for the most part, however, cannot be used by researchers directly, given that most of the features presented therein have missing values. Data need to be preprocessed. When missing data are handled during preprocessing, a particular feature is either dropped or filled by using statistical formulae. Such procedures, then, fail to produce accurate results. In most clinical datasets, information on positive and negative COVID-19 patients is not balanced. Therefore, a new clinical dataset was constructed using information on 2000 patients with COVID-19 symptoms, all of whom had taken a blood test between August 2020 and August 2021, at a private hospital in Thoothukudi, Tamil Nadu, India. Patient data privacy is maintained by excluding the patient's name and any other personal information. Instead, a fictitious patient number is linked to the collected blood sample data. Based on the values of the 27 features in the blood test, the final result was noted as COVID-19 positive or negative. Table I provides a description of the clinical dataset. Following the acceptance of the paper, the entire dataset will be made accessible publicly via the link: https://github.com/merviname/COVID-19.

B. Data Preprocessing

The data acquired are preprocessed by checking for missing values and duplicate entries. The dataset contains no missing value. The duplicate entries in the dataset are removed. The dataset contains

TABLE I. Covid-19 Clinical Dataset Description

S.No	Field	Data Type	Description	Normal Range values
1	AGE	Numerical	Patient Age	>0
2	GENDER	Categorical	Patient Gender	M-Male F-Female
3	HB	Numerical	It is the measurement of Hemoglobin in blood. COVID-19 patients have a decrease in HB which indicates the low oxygen carrier in blood.	Male-13.5– 17.0 g/dl Female–12.0–15.5 g/dl
4	TC	Numerical	It stores the count of white blood cells. Patients affected by COVID-19 have a significant increase of white blood cells.	4000-11000 cells
5	DC:NEUTROPHILS	Numerical	It stores the count of neutrophils in the blood. Patients affected by COVID-19 have increase in neutrophil count.	40-65 %
6	LYMPHOCYTES	Numerical	It stores the lymphocytes count. The lymphocytes count decrease in COVID-19 patients.	30-50 %
7	EOSINOPHILS	Numerical	It stores the Eosinophil count in the blood to measure the allergic disease, infections etc.COVID-19 patients has a decrease in eosinophil count.	100-400 cells/mL
8	MONOCYTES	Numerical	It is a kind of white blood cell which fight against disease and infections. COVID-19 patients has a decrease in Monocytes count.	200-800 /mL
9	BASOPHILS	Numerical	It is used to measure the allergic reaction. COVID-19 patients has a lower basophils count.	0-300 /mL
10	ESR (60 MIN)	Numerical	Erythrocyte Sedimentation Rate-It is used to measure and identify the inflammation. There is an increase of ESR in COVID-19 patients.	Male : 0-17 mm/hour Female : 1-25 mm/hour Children : 0-10 mm/hour
11	PC	Numerical	It stores the average platelets count in the blood. COVID-19 patients has a lower platelet count.	150 - 410 thousands/cmm
12	PCV	Numerical	Packed Cell Volume - It stores the Red Blood Cells proportion in blood. Patients affected by COVID-19 have a decrease in PCV.	Male : 40-52% Female : 35- 47%
13	MCV	Numerical	Mean Corpuscular Volume – It stores the size of Red Blood Cell. Patient affected by COVID-19 has low MCV value.	80-95 fL
14	MCH	Numerical	Mean Corpuscular Hemoglobin – It stores the Hemoglobin amount in Red Blood Cell. Patient affected by COVID-19 has low MCH value.	27.5 - 33.2 pg/cell
15	MCHC	Numerical	Mean Corpuscular Hemoglobin Concentration – It stores the average amount of hemoglobin in the group of Red Blood Cells. Patient affected by COVID-19 has low MCHC value.	32 – 36 b/dL
16	RBC	Numerical	It stores the number of Red Blood cells in the blood. In severe COVID-19 affected patients shows low RBC.	Male: 4.7 – 6.1 million cells/microliter Female: 4.2– 5.4 million cells/microliter
17		Numerical	Red Cell Distribution Width – It stores the variances in the size and volume of Red Blood Cells. Severe COVID-19 affected patient shows high RDW-CV count.	Male : 11.8 – 14.5 % Female : 12.2 – 16.1 %
18	RBS	Numerical	Random Blood Sugar Test. It stores the level of Blood Sugar of a non-fasting person. COVID-19 patients will have increase in the blood sugar level.	< 140 mg/dL
19	UREA	Numerical	It stores the amount of Urea in the Blood sample. Urea level in blood is high for COVID-19 patients.	6 – 24 mg/dL
20	CREATININE	Numerical	It stores the measure of creatinine in the blood sample. Creatinine in blood sample has an increase in COVID-19 patients.	Male :0.74 – 1.35 mg/dL Female : 0.59 – 1.04 mg/dL
21	CRP	Numerical	It stores the measure of C-reactive protein in the blood. There is significant increase in the CRP value for COVID-19 patients.	0-5 mg/L
23	D-DIMER	Numerical	It stores the measure of protein fragments of blood clots floating in the blood. COVID-19 patients have a higher D-Dimer value.	< 500 mg/mL
24	LDH	Numerical	It stores the amount of Lactate Dehydrogenase in the blood. There is significant increase in the LDH amount for COVID-19 patients.	125 – 343 U/L
25	DIRECT BILLIRUBIN	Numerical	It stores the measure of conjugated bilirubin. There is an increase in the direct Bilirubin value for COVID-19 patients.	0 – 0.3 mg/dL
26	BILLIRUBIN T	Numerical	It stores the sum of Direct and Indirect Bilirubin. There is an increase in the Indirect Bilirubin value for COVID-19 patients.	0.1–1.2 mg/dL
27	INDIRECT BILLIRUBIN	Numerical	It stores the measure of unconjugated bilirubin There is an increase in the Indirect Bilirubin value for COVID-19 patients.	0.2–0.8 mg/dL
28	SGOT	Numerical	Serum-Glutamic-Oxaloacetic-Transaminase – It stores the measure of enzyme found in liver, heart and other tissues. There is an increase in the SGOT value for COVID-19 patients.	8–45 units/litre

categorical and numeric values. Machine learning algorithms works well with numerical data. Hence, to convert categorical values into numerical values, one-hot encoding is used. The outliers in the dataset are removed using iForest algorithm and then the dataset is balanced using SMOTE algorithm.

1. One-Hot Encoding

Features with string values refer to categorical data, while most machine learning algorithms work with numerical values. Hence, these categorical values have to be mapped with numerical values. This conversion helps the algorithm for better prediction [29]. In this study, the categorical feature, 'gender' is converted into a numerical feature by creating two columns, Gender_1 and Gender_2 by using one-hot encoding method.

2. iForest

Anomalies in a dataset differ from normal records both in terms of quantity and quality. Removing these outliers can significantly enhance the performance of a classification model. In the context of this study, the Isolation Forest (iForest) [30] technique was employed to identify and eliminate outliers from the proposed COVID-19 clinical dataset. iForest identifies outliers by calculating the average path lengths for instances within its tree structures, with outliers being instances having notably shorter average path lengths.

iForest demonstrates efficient performance when used with a relatively small subsample size and an appropriate number of trees. The 'contamination' parameter serves the purpose of specifying the proportion of outliers present in the dataset. For this particular study, the chosen parameter configuration led to the detection of 24 outliers in the dataset mentioned above. After the removal of these outliers, the subsequent step involved addressing the issue of dataset imbalance. Imbalanced data can significantly impact the performance of a classification model, especially during training. Imbalanced data often causes the classification model to exhibit bias toward the majority class, leading to an increased occurrence of both false positives and false negatives. This, in turn, diminishes the overall performance of the classification model. Therefore, to enhance the performance, the proposed classification model balanced the COVID-19 data.

The clinical dataset proposed for this study displayed a significant imbalance, consisting of 997 COVID-positive cases and 999 being COVID-negative cases. This stark imbalance tilted the dataset heavily toward negative cases. To address this imbalance, the proposed model employed the Synthetic Minority Over-sampling Technique (SMOTE) to randomly generate minority class instances, effectively oversampling the minority class and rebalancing the dataset. Then, the entire dataset was randomly split into an 80% training set and a 20% test set.

C. Feature Selection

The columns/attributes in the dataset are termed features and only essential ones are needed to train an optimal model Feature selection, which is the process of choosing essential features, is critical to building a machine learning model because it reduces data redundancy and thus maximizes the model's performance. The objectives of feature selection techniques are (i) to reduce the model's complexity by removing irrelevant features, (ii) to help the machine learning algorithm train a model faster, and (iii) to avoid overfitting by reducing the dimensions [31]. Based on its interaction with the classifier, the feature selection algorithm is divided into three types they are: (i)filter method, (ii)wrapper method and (iii)embedded method. This study makes use of filter and wrapper methods.

1. Filter Methods

Statistical techniques are used in the filter method to assess the dependence between the input variable and the target variable. Statistical measures such as Fisher score, mutual information, chi-square test, correlation coefficient and variance threshold identify important features [32]. The techniques calculate the scores based on variance, correlation, consistency and distance, depending on the data's intrinsic properties. Thereafter, the features are ranked from best to worst, based on the said scores [33]. Fig. 2. shows the operation of the filter method. This paper employs the following four filter-based feature selection methods: (i) ANOVA-F (ii) chi-square (iii) mutual information (iv) Pearson correlation.

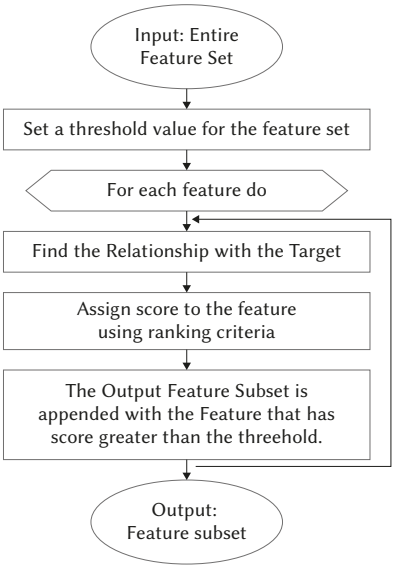

Fig. 2. Filter Method [41].

a) ANOVA-F:

ANOVA (Analysis of Variance) analyses each feature individually to examine the feature - target relationship. Features with a tenuous relationship with the target are eliminated. The F-Test is a statistical function that computes the ratio of the variance values. The variance denotes the dispersal measure of the data points from the mean. The ANOVA-F is used for the numerical input variable with the classification target variable [34]. Based on the test results, the best features with a high F-statistic score are selected. Features with a low F-statistic score, which are independent of the target variable, are removed from the dataset.

b) Chi-Square:

The chi-square test identifies attributes that are highly dependent on the target variable. It measures dependencies by examining the deviation of the expected count from the observed count. The chi-square value is small when the observed count is close to the expected count, indicating that the input feature is independent of the response. The higher chi-square value shows that the dependencies between the feature and response is high [35]. The chi-square feature selection algorithm selects features by calculating the chi-score and the p-value. The most significant features have a high chi-square score and a low p-value.

c) Mutual Information:

Mutual Information calculates the entropy for each feature with reference to the target feature [36]. Mutual information is calculated for each independent feature, following which the features are ranked, based on the calculated information gain for each feature. A threshold is set for selecting features with information gain above the threshold value. Mutual information thus helps find the most useful features that differentiate the target class.

d) Pearson Correlation:

Pearson correlation constructs a correlation matrix that measures the linear association between two features. The values in the matrix range from -1 to 1. Values closer to -1 and 1 indicate strong negative correlation and positive correlation, respectively. Values closer to 0 indicate weak correlation, while features with a value of 0 have no correlation [37]. A threshold is set to select the best features, and those with a higher score than the threshold is selected while others are removed from the dataset.

2. Wrapper Methods

The wrapper-based feature selection technique selects the best feature subset by producing a number of candidate feature subsets whose accuracy is evaluated using a classification algorithm. The best feature set is defined as the feature subset with the highest accuracy [38]. Fig. 3 shows the working of the wrapper-based feature selection method. The wrapper method such as Recursive Feature Selection (RFE) and Sequential Forward Selection(SFS) are used in this research.

a) Recursive Feature Elimination (RFE):

The Recursive Feature Elimination (RFE) algorithm first determines the most significant features and subsequently removes the least important ones, one at a time, in each iteration. The features are eliminated repeatedly until an optimal threshold is obtained from the classification algorithm. The final feature set obtained is the best [39]. Each feature is ranked using the rfe_ranking and features with '1' in the rfe_ranking column are selected for classification.

b) Sequential Forward Selection (SFS):

The Sequential Forward Selection (SFS) algorithm initially has an empty set of features, with features added on to the feature set at each iteration. The best feature set is obtained when the iteration yields a reduced misclassification rate [40]. The average score for each feature subset is calculated. Initially, the average score starts with a single feature and, at each iteration, another feature is added to the subset. The feature subset with the highest average score is selected as the best, and the features within it are selected for classification.

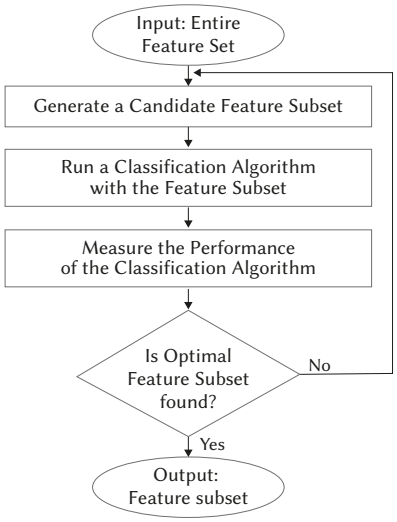

Fig. 3. Wrapper Method [41].

D. Classifiers

A classifier is a machine learning algorithm that can be used to identify the class of given input data. It takes the input data and outputs discrete class labels that define a set of possible classes. Classifiers, once trained with machine learning classification algorithms, can be used to make predictions on new data points and identify the class to which a training set belongs [42]. Several machine learning classification algorithms are used in this paper to predict COVID positive cases based on patients' blood test results.

Supervised machine learning models such as the SVM, Naive Bayes, KNN and logistic regression are applied for learning the preprocessed data after selecting the best features. The dataset is divided into training and testing data in 80:20 ratio. The classifier algorithm is used to train the model using training set. After that 20 percent of the set is used as testing data. The implementation process of the models used in this study follow below.

1. Support Vector Machines

The support vector machine is a statistics-based supervised machine learning algorithm that is used for classification and regression [43], which enables it to predict COVID-19 with its features The SVM creates a decision boundary known as the hyperplane which differentiates between COVID-19 positive and negative classes. The selected features from the clinical dataset are trained using the SVM algorithm. The training process results in a set of support vectors and a decision boundary. A predictive analysis is carried out using w*xi-c = +1 and w*xi-c = -1 (where 'w' is the vector which is normal to the hyperplane and 'c' the offset) by dividing the points on the hyperplane [44]. The SVM classifies the given new input vectors by calculating the distance from the decision boundary. The distance (d) from the point (a0,b0) to the line Mx+Ny+ O is calculated using Eq. 1.

$$d = \frac{|Ma_0+Nb_0+O|}{sqrt(M^2+N^2)} \quad (1)$$

Similarly, the distance between the hyperplane $w^T(\Phi(x)) + c$ and the given vector $\Phi(a_0)$ is given by Eq 2.

$$d_H(\Phi(a_0)) = \frac{|w^T(\Phi(a_0))+b|}{||w||_2} \quad (2)$$

where 'w' is the vector that is normal to the hyperplane, 'b' the offset and ||w||2 the length of *w* in the Euclidean norm. ||w||2 is given by ||w||2 = sqrt(w12 + w22 + w32 +....wn2).

For better accuracy using the SVM, the algorithm maximizes the distance and gives space to the hyperplane. Hence, to maximize the minimum distance, Eq. 3 is used.

$$w^* = \text{argwmax}[\text{minn}d_H(\Phi(a_n))] \quad (3)$$

If a point is substituted in the hyperplane equation (w*x + c > 0) and is greater than zero, then the given data is COVID-19 positive. If a point is substituted in the hyperplane equation (w*x + c < 0) and is less than zero then the given data is COVID-19 negative.

Pseudocode for the Support Vector Machine

Input: D = [X, Y]; X (dataset with m features), Y (class labels)

Output: Test case class

1. Initialize the model with random values for the weights, w.
2. Split the dataset, D, into two parts: a training set, T, and a testing set, T1.
3. For each training data, xi, from dataset T:
   3.1 Compute the margin, yi(w) = w^Txi.
   3.2 If the margin is greater than 1, add xi to the support vector set, S.
4. Find the optimal weights, w*, by solving the quadratic optimization problem, subject to yi(w) >= 1 for all i in S.
5. Return the support vector set, S, and the optimal weights, w*.
6. Assign the test data as positive if the yi(w) >= 1, otherwise classify it as negative.

2. Naïve Bayes:

The Naïve Bayes algorithm is a statistical supervised machine learning algorithm that predicts class membership using probability. The algorithm works well for small datasets but even better for large ones, offering high accuracy and speed [45].

Pseudocode for the Naïve Bayes [46]
Input: D = [X, Y]; X (dataset with n features), Y (class labels)
F = (f1, f2, f3, ..., fn) // features in the testing dataset
Output: Test case class
1. Split the dataset, D, into two parts: a training set, T, and a testing set, T1.
2. Read the training dataset, T.
3. Repeat to Compute the probability of fi using the Gauss density equation in each class until the probability of all predictor variables (f1, f2, f3, .., fn) has been calculated
4. Compute the likelihood for each class.
5. Select the greatest likelihood.

The Naïve Bayes algorithm is based on Bayes' theorem, written as in Eq. 4

$$P(C|D) = \frac{P(D|C) * P(C)}{P(D)} \quad (4)$$

where P(C) is the prior probability denoting the probability of occurrence C and P(D) the marginal probability denoting the probability of occurrence D. The probability values are independent and do not refer to each other. P(C|D) is known as the posterior probability which represents the probability of occurrence of C, given that D has occurred. This algorithm does not depend on other parameters and uses Eq. 5 to predict COVID-19.

$$P(C|D1 \ldots Dn) \alpha \prod_{i=1}^{n} P(D_i|C) P(C) \quad (5)$$

Eq. 6 below is used to calculate the highest probability.

$$H = \text{argmax}C \prod_{i=1}^{n} P(D_i|C) P(C) \quad (6)$$

3. K-Nearest Neighbors (KNN):

The k-NN algorithm is the most fundamental supervised machine learning algorithm used for classification. It classifies the given blood samples by using the majority of the classes in the clinical dataset's k-nearest neighbors. To find the nearest neighbors for a given data point, the algorithm typically employs the Euclidean distance metric. The distance metric formula is given in Eq. 7:

$$d(xi, xj) = \sqrt{\sum_{k=1}^{n} w_k (f_k(x_i) - f_k(x_j))^2} \quad (7)$$

where x = (f1, f2, f3, ..., fn), n is the number of attributes, fk is the kth attribute with its weight denoted by w_k, and d (xi, xj) is the distance between xi and xj [46].

Pseudocode for the k-Nearest Neighbor Algorithm [46]
Input: D=[X,Y]; X,Y(class labels)
Output: Test case class
1. Initialize the model with random values for the weights, w.
2. Split the dataset, D, into two parts: a training set, T, and a testing set, T1.
3. Read the training dataset, T.
4. For i=1 to n, do
   Compute distance d(Ti,T1).
   End for
5. Compute set1 containing indices for the k smallest distances, d(Ti,T1).
6. Return a majority label for {Yi where i € I}.

4. Logistic Regression:

Logistic regression predicts using a logistic function. The logistic function is a sigmoid function that takes a real-value number as input and maps it between 0 and 1 [47]. The 15 features selected using the feature selection algorithm are given as input and the class membership probability is calculated using Eq. 8

$$y = \frac{e^{(b_0 + b_1 * x)}}{(1 + e^{(b_0 + b_1 * x)})} \quad (8)$$

where the predicted output is denoted by y, b0 is the intercept term, and b1 is the coefficient of input x[48]. The binary classification made is based on the value of y. Here the binary class value is either 0 or 1. A class value of 0 is COVID-19 negative and a class value of 1 is COVID-19 positive as shown in Eq. 9 and Eq. 10.

$$0 \text{ if } y < 0.5 \quad (9)$$

$$1 \text{ if } y >= 0.5 \quad (10)$$

Pseudocode for Logistic Regression
Input: D = [X, Y]; X (dataset with n features), Y (class labels)
Output: Test case class
1. Initialize the model with random values for the weights, w.
2. Split the dataset, D, into two parts: a training set, T, and a testing set, T1.
3. Read the training dataset, T.
4. For each data in the training set, T,
   Calculate the probability using the formula in Eq. 8
5. If y<0.5, assign the class label 0, otherwise assign the class label 1.

IV. COVID-19 Prediction Procedure

An early screening procedure for predicting COVID-19 is given in the form of pseudocode. The blood test values are given as input and the output class gives the information whether the person is affected by COVID-19 or not. The working of COVID-19 prediction is given as a pseudocode below:

Pseudocode for COVID-19 Prediction
Input: COVID-19 Clinical Dataset D with Y classes

Begin
D1: Convert categorical values to numerical values using one-hot encoding // data preprocessing
X: Select relevant features from D1 // feature selection
T: 80% samples from X // training set
T1: Remaining 20% of samples from X // testing set
N: Number of samples in T
F: Feature labels from f1 to fn
C: Classification algorithm
For each feature in f1 ... fn,
    Construct a new label vector for the Y classes.
Apply T to C // the classifier is trained using training dataset.
Testing data (T1) is given as input to the trained classifier.
Calculate the confusion matrix for T1.
Evaluate the performance of the classifier using the confusion matrix.
Choose the classifier and feature selection algorithm with the best accuracy.
End

Output: Prediction of COVID-19 using the best feature selection and classifier algorithm

V. Experimental Findings and Discussion

This section included several experiments to determine the best feature selection and classification algorithm for COVID-19 prediction. The first experiment was carried out to identify the best feature selection algorithm which selects significant features from the dataset. The second experiment was carried out to find a suitable classifier for the selected features. The clinical dataset taken for the experimental set-up is described in the following section.

A. Dataset Description

The features of the proposed dataset are age, gender, D-Dimer, C-Reactive Protein (CRP), Lactate DeHydrogenase (LDH), total number of white blood cells (TC), platelet count (PC), packed cell volume (PCV), monocytes, eosinophils, basophils, mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), erythrocyte sedimentation rate (ESR (60 min)), lymphocytes, serum glutamic-oxaloacetic transaminase (SGOT), random blood sugar (RBS), billirubin T, direct billirubin, indirect billirubin, DC:Neutrophils, emoglobin (HB), red blood cells (RBC), red cell distribution width RDW-CV), urea and creatinine The last feature 'class' is used to identify whether the person is affected by COVID-19 or not, with '1' and '0' indicating COVID-19 positive and negative, respectively. Based on the 'class' feature, the dataset is divided into the types shown in Table II.

TABLE II. Statistical Information of the Dataset

2000 Blood Samples			
Positive		Negative	
1000		1000	
Female	Male	Female	Male
408	592	480	520

B. Ground Truth and Predicted Output Using the Existing Classifiers

The details of ten patients were given as input to the existing SVM, Naïve Bayes, k-NN, and logistic regression classifiers to predict whether the person is affected by COVID-19 or not. The results are tabulated in Table III, wherein the actual values correspond to the physician's diagnosis outcome, and the predicted values correspond to the target values predicted.

The "actual" values presented in the table signify the outcomes provided by a physician, aligning with the model's predictions. This verification in a real-world context serves to showcase the model's practicality and its relevance within a clinical setting.

1. Performance Metrics

The performance metrics used for selecting the best feature selection method and classification algorithm are discussed below.

TABLE III. Ground Truth and Predicted Output Using Existing Classifiers

Input											
Features		**Patients**									
		1	2	3	4	5	6	7	8	9	10
AGE		67	35	55	78	74	60	40	28	39	62
GENDER		F	M	F	M	M	F	F	M	M	M
HB		10.8	15.9	12.8	15.2	11.9	18.6	18.3	13.6	16.3	14.9
TC		7400	4500	9400	17200	18300	5000	6200	12300	20400	7400
DC: NEUTROPHILS		58	61	70	85	25	85	60	89	23	70
LYMPHOCYTES		23	31	21	15	20	12	38	17	12	27
EOSINOPHILS		2	5	1	0	1	0	0	2	1	1
MONOCYTES		5	6	2	2	1	1	4	2	2	1
BASOPHILS		1	1	1	0	1	1	1	0	0	1
ESR(60 MIN)		25	10	74	57	12	13	15	45	83	43
PC		2.8	1.9	4.3	2.6	1.6	4.4	1.3	2.2	4.5	2.5
PCV		33	48	31	39	36	27	47	37	38	38
MCV		83	89.8	91	88	89	74	86	89	87	85
MCH		30	27.9	31	32	32	24	30	31	31	31
MCHC		35	31.1	35	36	35	32	35	34	35	36
RBC		3.9	5.3	3.3	4.5	4	3.5	5.4	4	4.3	4.4
RDW-CV		15.2	15.5	15.3	15.1	11.8	12.4	12.5	11.5	15.1	18.6
RBS		169	189	196	230	220	169	143	350	467	220
UREA		15	35	37	41	58	22	20	24	60	30
CREATININE		0.8	1.5	1.3	1.2	1.2	1	0.8	1.2	1.3	1.1
CRP		79.8	10	40	60	55	43	3	36	26	15
D-DIMER		550	240	200	150	140	110	115	110	100	600
LDH		112	186	294	289	190	170	190	278	147	282
DIRECT BILLIRUBIN		0.4	0.3	0.3	0.4	0.4	0.6	0.3	0.4	0.3	0.3
BILLIRUBIN T		0.8	0.9	0.5	0.7	0.8	1.5	0.5	0.8	0.5	0.7
INDIRECT BILLIRUBIN		0.4	0.6	0.2	0.3	0.4	0.9	0.2	0.4	0.2	0.4
SGOT		29	25	27	20	35	54	30	39	38	37
Classifier	Output										
	Actual	1	0	1	1	1	1	0	1	1	1
SVM	Predicted	1	0	1	0	1	1	1	1	1	1
NB	Predicted	1	1	1	0	1	1	1	1	1	1
K-NN	Predicted	1	1	1	0	1	1	1	0	1	1
LR	Predicted	1	0	1	0	1	1	1	0	1	1

The performance of the classifier can be illustrated using confusing matrix. Most of the metrics are measured using confusion matrix. The accuracy, recall, precision, F1-score, and AUC were used to evaluate the results. Table IV displays the confusion matrix for COVID-19 prediction.

TABLE IV. Confusion Matrix for Covid-19 Prediction

	Has COVID-19 Disease	Does not have COVID-19 Disease
Has COVID-19 Disease	True positive	False Positive
Does not have COVID-19 Disease	False Negative	True Negative

True Positive (TP) :

If the actual class is COVID-19 positive and the model also predicts the class value as COVID-19 positive, then it is termed as True Positive.

True Negative (TN):

If the actual class is COVID-19 negative and predicted class value is also COVID-19 negative, then it is termed as True Negative.

False Positive (FP):

If the actual class is COVID-19 positive but the predicted class result is COVID-19 negative, then it is termed as False Positive

False Negative (FN):

If the actual class is COVID-19 negative but the predicted class result is COVID-19 positive, then it is termed as False Negative.

Accuracy:

Accuracy is computed by adding the number of correctly predicted positive and negative predictions and then dividing it by all types of predictions (TP, TN, FP, FN) [49] as shown in Eq. 11.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{11}$$

Precision:

Precision is the fraction of number of correctly predicted positive instances and the total number of correct or incorrect predicted positive instances (TP, FP) [50]. Precision is also termed the Positive Predictive Rate (PPR) as shown in Eq. 12.

$$\text{Precision} = \frac{TP}{(TP+FP)} \tag{12}$$

Recall:

Recall is the fraction of correctly predicted positive (TP) instances and the sum of correctly predicted positive and incorrectly predicted negative instances (TP, FN) [50] as shown in Eq.13. It is otherwise called the True Positive Rate (TPR).

$$\text{Recall} = \frac{TP}{(TP+FN)} \tag{13}$$

F1-score:

F1-score is calculated as the weighted average of precision and recall as shown in Eq. 14. Since it takes into account false positive and false negative predictions, these metric measures accuracy for uneven datasets better [50].

$$\text{F1-Score} = 2 \times \frac{(Precision \times Recall)}{(Precision+Recall)} \tag{14}$$

AUC:

AUC stands for Area under the curve. It is the measure of how well it distinguishes between each class. It is also known as Receiver Operator Characteristics (ROC) curve summary. It is used as a metric in binary classification problem.

C. Validation Methods

This section deals with the two types of cross-fold and split dataset validation methods used in the research.

K-fold validation:

The K-fold validation method trains and evaluates the model "k" times for different samples [51]. Performance metrics are used to evaluate each fold, and the fold with the highest accuracy is selected the best.

Data split validation:

Based on the number of samples, the dataset is divided into a train set split and a test set split. The split ratio normally commences with 80:20, 75:25, and 70:30, and goes on likewise. Metrics are used in every split to measure the performance of the model and the split with the best accuracy.

D. Comparison of State-of the Art Benchmarks in COVID-19 Prediction

In this section, a table is presented to outline the different machine learning algorithms employed in preprocessing, feature selection, and classification techniques. Table V compiles the state-of-the-art benchmarks in predicting COVID-19.

TABLE V. Comparison of State-of the Art Benchmarks in COVID-19 Prediction

Ref. No.	Dataset Used	Pre Processing	Feature Extraction / Selection	Classification/ Techniques Used	Accuracy (%)
		Noise removal / Handling Outliers			
[15]	IRCCS Ospedale San Raffaele	N/A	Feature Importance	Random Forest	82
[17]	Albert Einstein Hospital dataset	Highly correlated attributes are eliminated to reduce noise in the data.	Pearson Cor-relation and feature importance	Logistic regression, random forest, k nearest neighbours and Xgboost	92
[21]	Albert Einstein Hospital in São Paulo, Brazil	N/A	Correlation Matrix and The Chi-Squared Test	Ensemble of Support Vector Machines, Adaptive Boosting, Random Forest and K-Nearest Neighbors	69.9
[23]	Albert Einstein Hospital (Kaggle)	N/A	N/A	K Nearest Neighbor, Radial Basis Function, Naive Bayes, kStar, PART, Random Forest, Decision Tree, OneR, Support Vector Machine and Multi-Layer Perceptron	88
	Proposed Clinical Dataset	iForest to Handle Outliers	Mutual Information	SVM	**94.8**

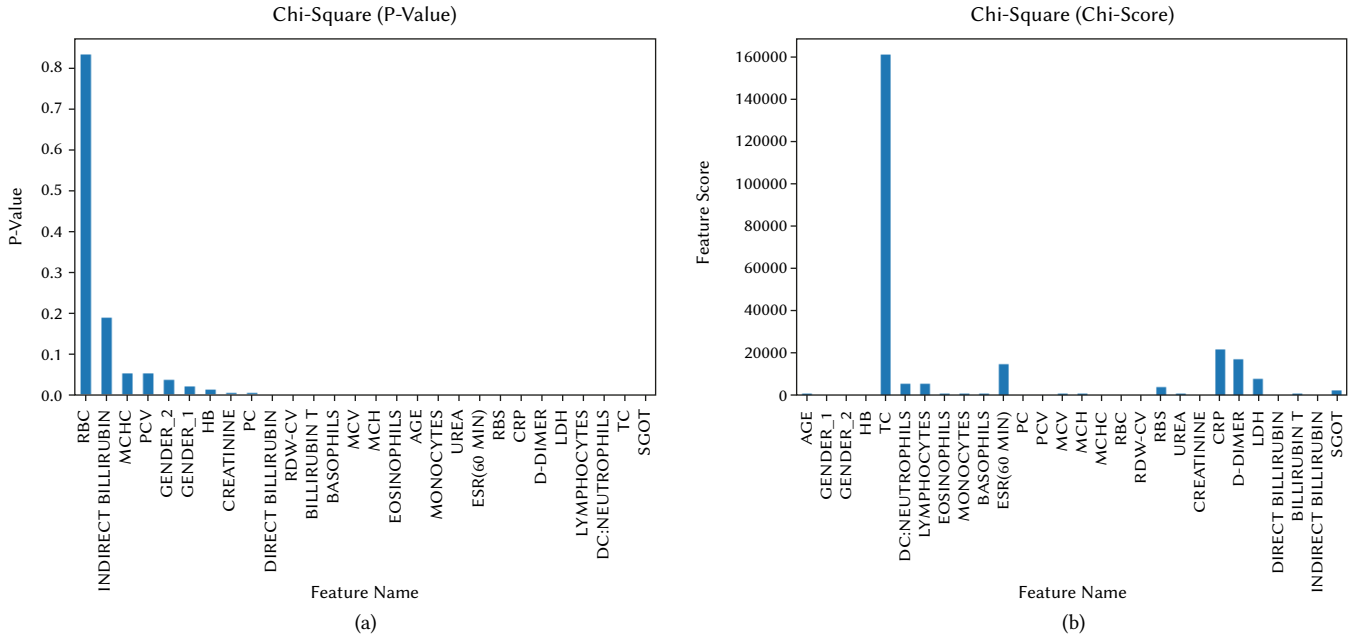

Fig. 4. (a) P-Value of the features using Chi-Square. (b) Chi-Score of the features using Chi-Square.

Based on the information provided in the table above, it is clear that the model using Albert Einstein Hospital (Brazil) dataset and the proposed clinical dataset that checks and handle outliers demonstrates improved accuracy than other models. Therefore, addressing outliers has effectively enhanced the model's robustness.

E. Finding the Best Features Using Feature Selection Methods

In the feature selection stage, which is indispensable to designing the model, the most appropriate features that maximize the model's performance are chosen. This section shows the results of the feature selection algorithm in the process of selecting the features from the clinical dataset. Feature selection methods used for the study include the ANOVA-F, chi-square, mutual information and Pearson correlation filter-based methods, as well as the RFE and SFS wrapper-based methods.

1. Chi-Square Test:

The outputs of the chi-squared test are the p-value and the chi score. A large p-value shows target-independent input features that are not selected for training. Target-dependent features with a high chi score, on the other hand, are selected for training. Fig. 4(a) and 4(b) show the graph representing the chi-square, based on the p-values and chi-score, respectively. The threshold for the chi-score is set to 100 and for the p-value to 0.05. Features selected for training include the MCV, MCH, Eosinophils, age, monocytes, urea, ESR (60 min), RBS, CRP, D-Dimer, LDH, lymphocytes, DC: Neutrophils, TC, and SGOT. These features have a p-value and a chi-score that are less than and greater than the threshold, respectively.

2. ANOVA-F:

In the ANOVA-F, the impact of the feature with the target variable is determined by the feature's variance. A low score implies that the feature has no impact on the target feature. Fig. 5. graphically depicts the feature score of all the features in the dataset using ANOVA-F feature selection. The top 15 features selected for classification are the TC, monocytes, RBS, Direct Billirubin, DC:Neutrophils, lymphocytes, basophils, ESR (60 min), MCH, D-Dimer, CRP, LDH, Eosinophils, Billirubin T and SGOT.

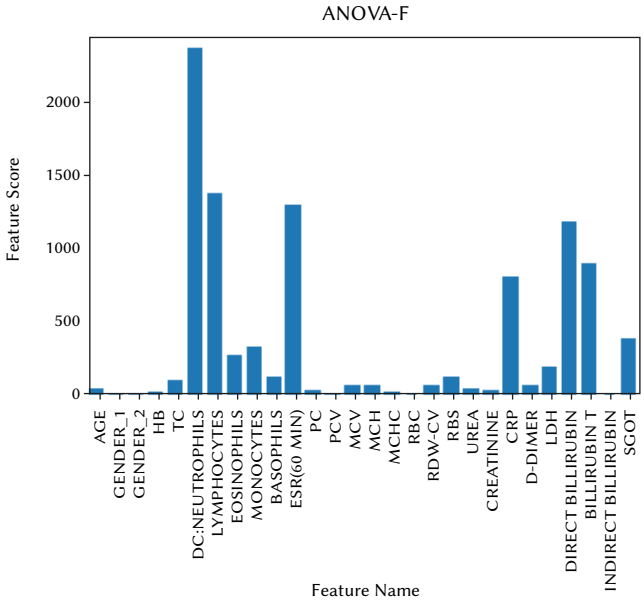

Fig. 5. Feature and its score using ANOVA-F.

3. Mutual Information:

Mutual information calculates the information gain for each feature. Fig. 6. shows the feature score calculated using mutual information for each feature. The top 15 features with high information gain are selected for classification. The selected features are DC: Neutrophils, Lymphocytes, CRP, Billirubin T, ESR (60 min), Direct Billirubin, D-Dimer, LDH, MCV, MCH, RBS, RBC, UREA, Eosinophils, and PC.

4. Pearson Correlation:

Fig. 7. Depicts the correlation matrix of various clinical dataset features. Highly correlated features are selected for classification. The selected features include DC:Neutrophils, Eosinophils, ESR (60 min), monocytes, basophils, MCV, MCH, RBS, Direct Billirubin, CRP, RDW-CV, Billirubin T, LDH, SGOT, and D-Dimer.

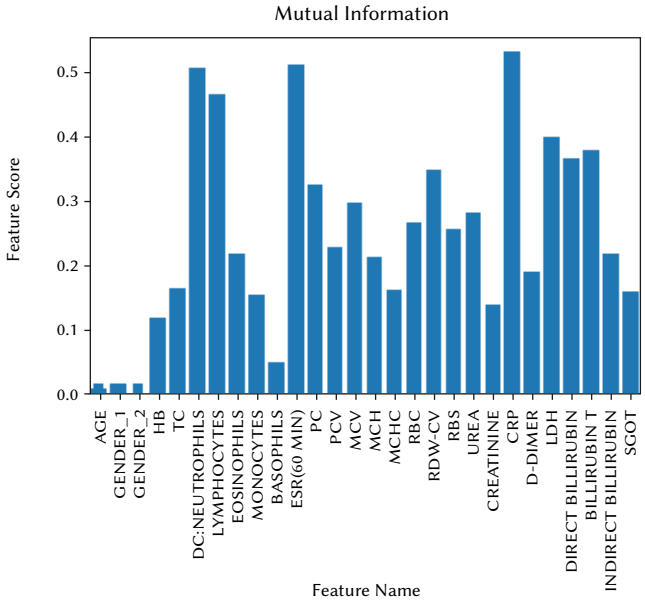

Fig. 6. Feature and its score using Mutual Information.

5. Recursive Feature Elimination (RFE):

Table VI shows a list of features selected using Recursive Feature Elimination (RFE). The selected features are those with a 'true' value in the RFE_Support column, while the RFE_Ranking column provides information on the rank of each feature. Features with '1' in the RFE_Ranking and 'True' in the RFE_Support are selected as the best which include Lymphocytes, DC: Neutrophils, Eosinophils, Monocytes, Basophils, ESR(60 min), PC,PCV,MCV, RBS, RDW-CV, urea, creatinine, CRP, D-Dimer and LDH.

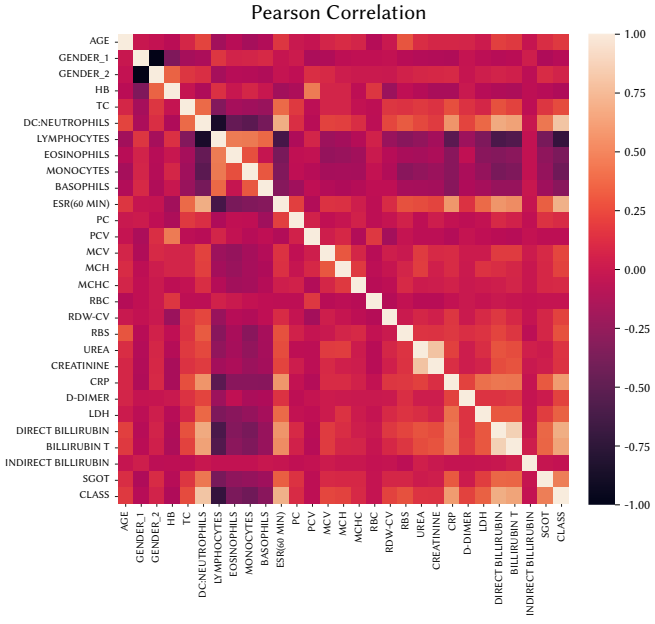

Fig. 7. Correlation Matrix of Clinical Dataset features.

6. Sequential Forward Selection (SFS):

Table VII shows a list of features and the average score of the feature subset selected using the Sequential Forward Selection (SFS) algorithm. The Feature_Names column displays a list of features selected in each iteration and the Avg_Score column gives the average

TABLE VI. FEATURES SELECTED USING RFE ALGORITHM

S. No.	Feature Name	RFE_Support	RFE_Ranking
1	AGE	False	14
2	GENDER_1	False	13
3	GENDER_2	False	12
4	HB	False	9
5	TC	False	8
6	DC: NEUTROPHILS	True	1
7	LYMPHOCYTES	True	1
8	EOSINOPHILS	False	2
9	MONOCYTES	True	1
10	BASOPHILS	True	1
11	ESR (60 MIN)	True	1
12	PC	True	1
13	PCV	True	1
14	MCV	True	1
15	MCH	False	10
16	MCHC	False	11
17	RBC	False	6
18	RDW-CV	True	1
19	RBS	True	1
20	UREA	True	1
21	CREATININE	True	1
22	CRP	True	1
23	D-DIMER	True	1
24	LDH	True	1
25	DIRECT BILLIRUBIN	False	3
26	BILLIRUBIN T	False	4
27	INDIRECT BILLIRUBIN	False	5
28	SGOT	False	7

of the feature score of the selected features in each iteration. Features selected by the SFS algorithm are Gender_1, Gender_2, HB, TC, DC Neutrophils, Lymphocytes, Eosinophils, Monocytes, Basophils, ESR (60 min), PC, PCV, MCV, CRP, and LDH.

F. Features Selected By Various Feature Selection Method:

The experiments above have selected certain features by different feature selection method. Some features are selected by more than one feature selection method. Table VIII presents an analysis of the votes gained by each feature.

It is found from the above table, features like CRP, DC neutrophils, lymphocytes, eosinophils, basophils, ESR (60 min), MCV, RBS, D-dimer, LDH, direct bilirubin, and billirubin T have a high vote.

G. Comparison of Classification Techniques Performance Based on Feature Selection Methods

Table VIII shows the comparison of classification technique performance based on various feature selection techniques. The dataset contains 28 features, of which 15 were selected using the ANOVA-F, chi-square test, mutual information, Pearson correlation, RFE and SFS feature selection methods. These 15 features are used for classifying patients' with COVID-19. The efficiency of the classifier is determined by various performance metrics. The confusion matrix helps to view the efficiency of the classifier pictorially. Fig.8, 9, 10, and 11 show, respectively, the confusion matrix obtained for the test data after training the model using the SVM, Naïve Bayes, k-NN and logistic regression classifiers. The classifiers work with the features selected using mutual information.

TABLE VIII. Features Selection Method And Selected Features

Name of the Feature	List of Selected Features with Feature Selection (15)					
	Filter Method				Wrapper Method	
	ANOVA-F	Chi-Square	Mutual Information	Pearson Correlation	RFE	SFS
AGE						
GENDER_1						✓
GENDER_2						✓
HB						✓
TC	✓	✓				✓
DC: NEUTROPHILS	✓	✓	✓	✓	✓	✓
LYMPHOCYTES	✓	✓	✓		✓	✓
EOSINOPHILS	✓	✓	✓	✓		✓
MONOCYTES	✓	✓		✓	✓	✓
BASOPHILS	✓			✓	✓	✓
ESR(60 MIN)	✓	✓	✓	✓	✓	✓
PC			✓		✓	✓
PCV					✓	✓
MCV		✓	✓	✓	✓	✓
MCH	✓	✓	✓	✓		
MCHC						
RBC			✓			
RDW-CV				✓	✓	
RBS	✓	✓	✓	✓	✓	
UREA		✓	✓		✓	
CREATININE					✓	
CRP	✓	✓	✓	✓	✓	✓
D-DIMER	✓	✓	✓	✓	✓	
LDH	✓	✓	✓	✓	✓	✓
DIRECT BILLIRUBIN	✓		✓	✓		
BILLIRUBIN T	✓		✓	✓		
INDIRECT BILLIRUBIN						
SGOT	✓	✓		✓		

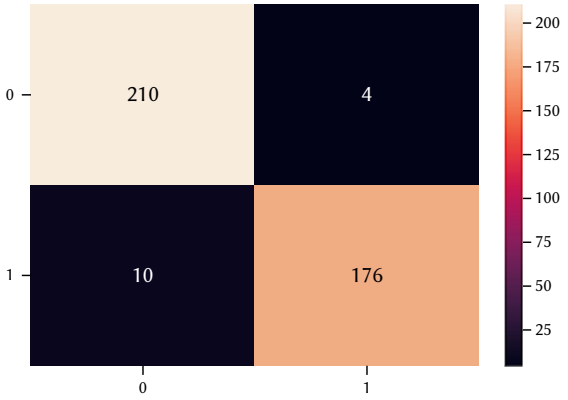

Fig. 8. Confusion Matrix – SVM.

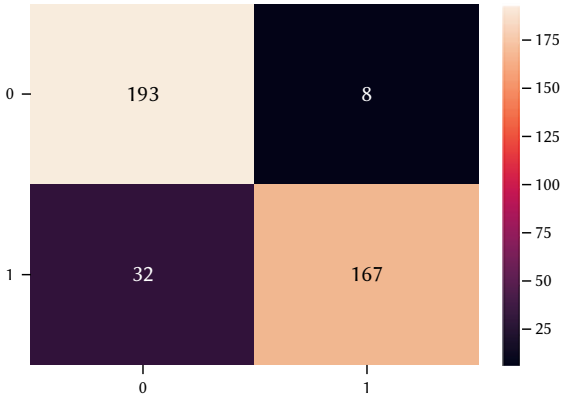

Fig. 9. Confusion Matrix –Naïve Bayes.

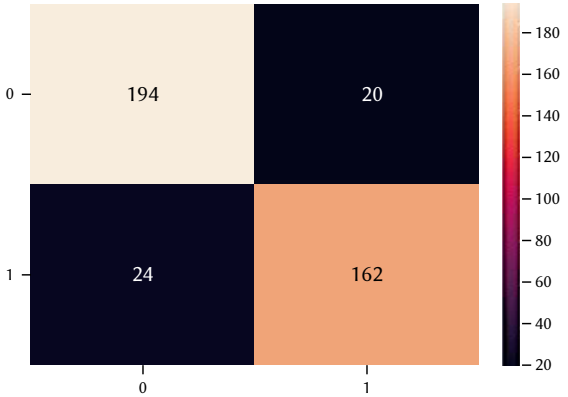

Fig. 10. Confusion Matrix – k-NN.

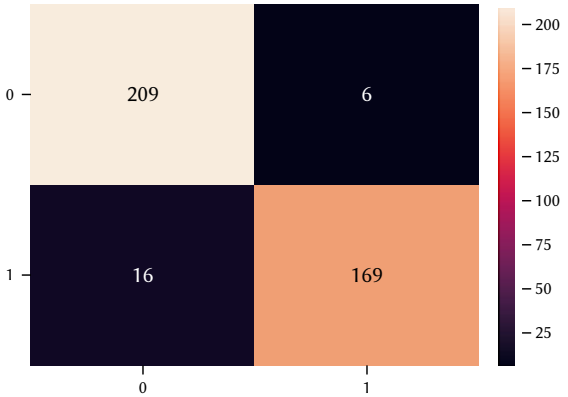

Fig. 11. Confusion Matrix – LG.

Note : 1 - COVID-19 Positive ; 0 – COVID-19 Negative

TABLE VII. Features Selected Using SFS Algorithm

S. No	Feature_Names	Avg_Score
1	CRP	0.71
2	CRP, LDH	0.75
3	TC, CRP, LDH	0.76
4	GENDER_1, TC, CRP, LDH	0.73
5	GENDER_1, GENDER_2, TC, CRP, LDH	0.76
6	GENDER_1, GENDER_2, HB, TC, CRP, LDH	0.78
7	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, CRP, LDH	0.74
8	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, CRP, LDH	0.75
9	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, CRP, LDH	0.79
10	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, CRP, LDH	0.78
11	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, CRP, LDH	0.81
12	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), CRP, LDH	0.83
13	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, CRP, LDH	0.84
14	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, CRP, LDH	0.86
15	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS, MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, CRP, LDH	**0.88**
16	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, CRP, LDH	0.84
17	GENDER_1, GENDER,_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC,CRP, LDH	0.81
18	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, CRP, LDH	0.82
19	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, CRP, LDH	0.76
20	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, CRP, LDH	0.74
21	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA, CRP, LDH	0.82
22	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA , CFREATININE, CRP, LDH	0.78
23	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA , CFREATININE, CRP, D-DIMER, LDH	0.79
24	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA , CFREATININE, CRP, D-DIMER, LDH, DIRECT BILLIRUBIN	0.81
25	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA , CFREATININE, CRP, D-DIMER, LDH, DIRECT BILLIRUBIN, BILLIRUBIN T	0.79
26	GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA , CFREATININE, CRP, D-DIMER, LDH, DIRECT BILLIRUBIN, BILLIRUBIN T, INDIRECT BILLIRUBIN, SGOT	0.77
27	AGE,GENDER_1, GENDER_2, HB, TC, DC:NEUTROPHILS, LYMPHOCYTES, EOSINOPHILS,MONOCYTES, BASOPHILS, ESR(60 MIN), PC, PCV, MCV, MCH, MCHC, RBC, RDW-CV, RBS, UREA , CFREATININE, CRP, D-DIMER, LDH, DIRECT BILLIRUBIN, BILLIRUBIN T, INDIRECT BILLIRUBIN, SGOT	0.77

TABLE IX. Performance of Different Classifiers With and Without Different Feature Selection Methods

Feature Selection Algorithm	No. of Selected attribute		Classifiers	Performance Metrics				
				Accuracy (%)	Precision	Recall	F1-score	AUC
Without Feature selection	28		SVM	92.7	0.93	0.93	0.93	0.93
			Naïve Bayes	89.7	0.90	0.90	0.90	0.90
			KNN	69.7	0.75	0.70	0.68	0.70
			LR	91.7	0.92	0.92	0.92	0.93
Filter	ANOVA-F	15	SVM	93.7	0.94	0.94	0.94	0.94
			Naïve Bayes	92.7	0.92	0.92	0.92	0.93
			KNN	90	0.89	0.90	0.89	0.90
			LR	93.7	0.94	0.94	0.94	0.94
	Chi-Square	15	SVM	93.7	0.94	0.94	0.94	0.94
			Naïve Bayes	90.6	0.90	0.90	0.90	0.91
			KNN	89	0.89	0.89	0.89	0.89
			LR	91.25	0.92	0.91	0.91	0.92
	Mutual Information	**15**	**SVM**	**94.8**	**0.95**	**0.95**	**0.95**	**0.95**
			Naïve Bayes	90	0.91	0.90	0.90	0.90
			KNN	89	0.89	0.89	0.89	0.89
			LR	91.25	0.91	0.92	0.91	0.92
	Pearson Correlation	15	SVM	92.7	0.93	0.93	0.93	0.93
			Naïve Bayes	92.7	0.93	0.93	0.93	0.93
			KNN	88.5	0.87	0.89	0.89	0.89
			LR	92.7	0.93	0.93	0.93	0.93
Wrapper	RFE	15	SVM	91.6	0.92	0.92	0.92	0.92
			Naïve Bayes	88.5	0.89	0.89	0.89	0.89
			KNN	88.5	0.89	0.89	0.89	0.89
			LR	91.6	0.92	0.92	0.92	0.92
	SFS	15	SVM	92.7	0.93	0.93	0.93	0.93
			Naïve Bayes	92.7	0.92	0.93	0.92	0.93
			KNN	88.5	0.87	0.89	0.88	0.89
			LR	92.7	0.93	0.93	0.93	0.93

It is inferred from Table IX that the features selected by mutual information perform the best with the SVM classifier compared to other methods, producing 94.8% accuracy

H. Performance Evaluation of Feature Selection Techniques Using K-Fold Validation

According to the results, the SVM classifier paired with feature selection technique works well. Fold validation and training-testing data split validation helps to improve the efficiency of the model by providing us the best fold and split. Table IX shows the comparison of various feature selection techniques performance using the SVM classifier in COVID-19 prediction. To find the effective fold for all filter and wrapper based feature selection methods, cross-fold validation is used. This experiment divides the dataset into 5 fold ranging from 1 to 5 and the above mentioned performance metrics are used for evaluation. Table X shows the comparison of various feature selection technique performance with the SVM classifier.

TABLE X. Comparison of Feature Selection Methods Performance With Svm Classifier Based on Fold Validation

Metrics	Feature Selection Algorithm	Comparison of Feature Selection Methods Performance Based on Fold Validation				
		Fold 1	Fold 2	Fold 3	Fold 4	**Fold 5**
Accuracy (%)	ANOVA-F	61.6	65.04	76.5	88.62	89.3
	Chi-Square	72.4	71.04	81.9	90.6	91.7
	Mutual Information	94.1	93.9	94.08	93.9	**94.5**
	Pearson Correlation	92.1	91.9	92.08	91.9	92.4
	RFE	89.1	88.8	89.1	88.8	89.8
	SFS	91.5	91.4	91.5	91.5	91.9
Precision	ANOVA-F	0.62	0.65	0.77	0.89	0.90
	Chi-Square	0.73	0.71	0.82	0.91	0.92
	Mutual Information	0.93	0.92	0.94	0.94	0.95
	Pearson Correlation	0.91	0.92	0.91	0.91	0.92
	RFE	0.89	0.88	0.89	0.88	0.89
	SFS	0.91	0.90	0.91	0.91	0.91
Recall	ANOVA-F	0.62	0.65	0.77	0.89	0.89
	Chi-Square	0.73	0.71	0.82	0.91	0.92
	Mutual Information	0.93	0.93	0.94	0.94	0.95
	Pearson Correlation	0.91	0.91	0.91	0.91	0.92
	RFE	0.89	0.88	0.89	0.88	0.89
	SFS	0.91	0.90	0.91	0.91	0.91
F1-Score	ANOVA-F	0.62	0.65	0.77	0.89	0.89
	Chi-Square	0.73	0.71	0.82	0.91	0.91
	Mutual Information	0.93	0.92	0.94	0.94	0.95
	Pearson Correlation	0.91	0.91	0.91	0.91	0.92
	RFE	0.89	0.88	0.89	0.88	0.89
	SFS	0.91	0.90	0.91	0.91	0.91

It is observed, from the results of Table X that 5th fold gives the best results. Moreover, the performance metrics show that the mutual information technique outperforms all the others.

I. Performance Evaluation of Feature Selection Techniques Using Data Splitting Validation

Many researchers do not focus on fold or split validation. The importance of data splitting is highlighted in this research, with experiments carried out to determine the suitable split for testing and training. The above mentioned metrics are used to evaluate the performance of feature selection methods with SVM classifier in order to determine the best fold and data splitting range for predicting

TABLE XI. Comparison of the Performance of Feature Selection Methods Based on Data Splitting Validation

Metrics	Feature Selection Algorithms	Comparison of the performance of feature selection methods based on data splitting validation												
		20-80	25-75	30-70	35-65	40-60	45-55	50-50	55-45	60-40	65-35	70-30	75-25	**80-20**
Accuracy (%)	ANOVA-F	80.6	80.6	80.2	81.2	81.9	80.5	82.1	81.6	83.1	84.9	85.8	89.1	93.7
	Chi-Square	80.1	81.5	82.7	83.6	84.2	85.6	86.1	87.7	88.8	89.9	90.1	91.7	93.7
	Mutual Information	83.2	84.9	85.2	86.8	87.2	88.8	89.4	90.2	91.8	92.6	93.1	93.8	**94.8**
	Pearson Correlation	80.6	81.7	80.3	82.3	83.4	85.1	86.7	87.1	88.4	89.1	90.5	91.2	92.7
	RFE	78.5	79.2	80.6	81.1	82.1	83.4	84.6	93.5	85.1	86.9	88.1	90.1	91.6
	SFS	75.2	76.2	77.8	78.2	80.2	81.5	82.9	84.1	88.2	89.4	90.1	91.8	92.7
Precision	ANOVA-F	0.78	0.80	0.80	0.81	0.82	0.81	0.82	0.82	0.83	0.85	0.86	0.89	0.94
	Chi-Square	0.80	0.81	0.83	0.84	0.84	0.86	0.86	0.88	0.89	0.90	0.91	0.92	0.94
	Mutual Information	0.84	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.93	0.94	**0.95**
	Pearson Correlation	0.81	0.82	0.81	0.82	0.83	0.85	0.87	0.87	0.88	0.89	0.91	0.92	0.93
	RFE	0.78	0.79	0.81	0.81	0.82	0.82	0.85	0.84	0.85	0.87	0.88	0.90	0.92
	SFS	0.75	0.77	0.78	0.78	0.80	0.82	0.83	0.84	0.88	0.89	0.90	0.91	0.93
Recall	ANOVA-F	0.77	0.81	0.81	0.82	0.82	0.81	0.82	0.82	0.83	0.85	0.86	0.89	0.94
	Chi-Square	0.79	0.81	0.83	0.84	0.83	0.86	0.86	0.87	0.88	0.89	0.90	0.92	0.94
	Mutual Information	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.94	**0.95**
	Pearson Correlation	0.81	0.82	0.81	0.82	0.83	0.85	0.87	0.87	0.88	0.89	0.91	0.92	0.93
	RFE	0.78	0.79	0.81	0.81	0.82	0.82	0.85	0.84	0.85	0.87	0.88	0.90	0.92
	SFS	0.75	0.77	0.78	0.78	0.80	0.82	0.83	0.84	0.88	0.89	0.90	0.91	0.93
F1 Score	ANOVA-F	0.77	0.81	0.81	0.82	0.82	0.81	0.82	0.82	0.83	0.85	0.86	0.89	0.94
	Chi-Square	0.79	0.81	0.83	0.81	0.83	0.86	0.84	0.87	0.88	0.89	0.89	0.91	0.94
	Mutual Information	0.85	0.86	0.87	0.88	0.89	0.90	0.91	0.92	0.93	0.94	0.95	0.94	**0.95**
	Pearson Correlation	0.81	0.82	0.81	0.82	0.83	0.85	0.87	0.87	0.88	0.89	0.91	0.92	0.93
	RFE	0.78	0.79	0.81	0.81	0.82	0.82	0.85	0.84	0.85	0.87	0.88	0.90	0.92
	SFS	0.75	0.77	0.78	0.78	0.80	0.82	0.83	0.84	0.88	0.89	0.90	0.91	0.93

TABLE XII. Comparison of Other Dataset Performance With Svm Classifier

S. No.	Ref. No.	Dataset	No. of Features	Instances	Features Selected	Accuracy (%)
1.	[15]	IRCCS Ospedale San Raffaele	15	219 (COVID-19 positive - 177 COVID-19 Negative - 102)	Gender, Age, WBC, platelets, CRP, AST, ALT, GGT, LDH, Neutrophils, Lymphocytes, Monocytes, Eosinophils, Basophils, Swab	82
2.	[21]	Albert Einstein Hospital in São Paulo, Brazil	111	5644 (COVID-19 positive - 558 COVID-19 Negative – 5086)	Monocytes, Age, Red Blood Cells, Serum Glucose Hematocrit, Hemoglobin, Leukocytes, Lymphocytes, Mean Platelet Volume, Creatinine, Calcium, Magnesium, Potassium, Sodium, Urea, Vitamin B12, Phosphor	69.79
3.	[23]	Albert Einstein Hospital (Kaggle)	72	1624 (COVID-19 positive - 786 COVID-19 Negative – 838)	LDH, AST, FG, CA, PCR, GLU, ALT, CO2POC, SO2POC, GLUEMO, WBC, FCOPOC, RDW, HHBPOC, AGE, HCT, FO2POC, BAT, XDP, GGT.	88
4.	-	**Proposed COVID-19 Clinical Dataset**	27	2000 (COVID-19 positive - 1000 COVID-19 Negative – 1000)	DC:Neutrophils, Lymphocytes, CRP, Billirubin T, ESR(60 Min), Direct Billirubin, D-Dimer, LDH, MCV, MCH, RBS, RBC, UREA, Indirect Billirubin, PC	**94.8**

COVID-19 using clinical data. Table XI lists a comparison of the performance of feature selection methods with the SVM classifier to predict COVID-19. To get the best training and testing splitting range, the split is listed in ranges from 20 - 80% to 80% - 20% as depicted in Table XI.

It is evident from Table XI that the 80%-20% training-testing data splitting shows high accuracy. The result shows that mutual information outperforms other feature selection techniques.

J. Comparing the Performance of the Datasets

This section compares the performance of open source clinical datasets with the proposed clinical dataset. It is found from the literature survey that open source datasets are the preferred choice for model-building. The total number of features and instances, as well as features chosen by the feature selection algorithm, are analysed. The features were classified using the SVM classifier and its performance.

It is observed from Table XII that the open source datasets used have imbalanced data, unlike the proposed dataset. This proposed dataset has been used to build a model that selects relevant features and predicts COVID-19 using the SVM classifier with 94.8% accuracy, outclassing other datasets. The hyperparameters such as C (penalty parameter), kernel, gamma, coef0 can be tuned to improve the performance of the model.

VI. Conclusion

This research was carried out to publish a new clinical dataset on GitHub. Further, it focused on selecting the best features using feature selection techniques and finding a suitable classifier to predict COVID-19. To this end, a literature survey was completed to examine the feature selection methods and classification algorithms used for COVID-19 prediction using a clinical dataset. Different experiments were conducted using the clinical dataset in order to determine the suitable feature selection algorithm that selects the most relevant features, along with an appropriate classifier for the prediction. Based on the experiments, the mutual information filter-based feature selection algorithm was identified to be the best of its kind. The SVM classifier, with a high 94.8% accuracy, outperformed the rest. While the model excels at predicting COVID-19, its primary limitation lies in its lack of generalizability, stemming from its reliance on data from a single hospital for model training. Future directions include extending the research by modifying the mutual information algorithm to select the best feature to enhance the performance of the classifier. Likewise, two classifiers can be combined to form an ensemble classifier that can be used to build a high-performance classifier for COVID-19 prediction using the clinical dataset.

References

[1] T. Singhal, "A Review of Coronavirus Disease-2019 (COVID-19)," *Indian Journal of Pediatrics*, vol. 87, no.4, pp.281-286, 2020, doi: 10.1007/s12098-020-03263-6.

[2] S. Mubareka, J. B. Gubbay, W. C. Chan, "Diagnosing COVID-19: the disease and tools for detection," *American Chemical Society Nano*, vol. 14 no.4, pp. 3822-35, 2020, doi:10.1021/acsnano.0c02624.

[3] D. Li, D. Wang, J. Dong, N. Wang, H. Huang, H. Xu, C. Xia, "False-negative results of real-time reverse-transcriptase polymerase chain reaction for severe acute respiratory syndrome coronavirus 2: role of deep-learning-based CT diagnosis and insights from two cases," *Korean journal of radiology*, vol. 21, no. 4, pp. 505-508, 2020, doi: 10.3348/kjr.2020.0146.

[4] A. Ulhaq, J. Born, A. Khan, D. P. S. Gomes, S. Chakraborty, M. Paul, "COVID-19 Control by Computer Vision Approaches: A Survey," *IEEE Access*, vol. 8, pp. 179437-179456, 2020, doi: 10.1109/ACCESS.2020.3027685.

[5] J. Bao, C. Li, K. Zhang, H. Kang, W.Chen, B. Gu, "Comparative analysis of laboratory indexes of severe and non-severe patients infected with COVID-19," *Clinica Chimica Acta*, vol. 509, pp. 180-194, 2020, doi: 10.1016/j.cca.2020.06.009.

[6] B. E. Fan, "Hematologic parameters in patients with COVID-19 infection: a reply," *American journal of hematology* vol. 95, no. 6, 2020, doi: 10.1002/ajh.25774.

[7] Y. Gao, T. Li, M. Han, X. Li, D. Wu, Y. Xu, Y. Zhu, Y. Liu, X. Wang, L. Wang, "Diagnostic utility of clinical laboratory data determinations for patients with the severe COVID-19," *Journal of medical virology*, vol. 92, no. 7, pp. 791-796, 2020, doi: 10.1002/jmv.25770.

[8] T. A. Khartabil, H. Russcher, A. Ven, Y. B. Rijke, "A summary of the diagnostic and prognostic value of hemocytometry markers in COVID-19 patients," *Critical reviews in clinical laboratory sciences*, vol. 57, no. 6, pp. 415-431, 2020, doi:10.1080/10408363. 2020.1774736.

[9] A. J. Rodriguez-Morales, J. A. Cardona-Ospina, E. Gutiérrez-Ocampo, R.Villamizar-Peña, Y. Holguin-Rivera, J. P. Escalera-Antezana, et al., "Clinical, laboratory and imaging features of COVID-19: A systematic review and meta-analysis," *Travel medicine and infectious disease*, vol.34, 2020, doi: 10.1016/j.tmaid.2020.101623.

[10] J. A. Siordia, "Epidemiology and clinical features of COVID-19: A review of current literature," *Journal of Clinical Virology*, vol. 127, 2020, doi: 10.1016/j.jcv.2020.104357.

[11] Y. Liu, Y. Yang, C. Zhang, F.Huang, F. Wang, J. Yuan, et al., "Clinical and biochemical indexes from 2019-nCoV infected patients linked to viral loads and lung injury," *Science China Life Sciences.*, vol. 63, no. 3, pp. 364-74, 2020, doi: 10.1007/s11427-020-1643-8.

[12] L. Muhammad, M. M. Islam, S. S. Usman, S. I. Ayon, "Predictive data mining models for novel coronavirus (COVID-19) infected patients' recovery," *SN Computer Science*, vol. 1, no. 4, pp. 1–7, 2020, doi: 10.1007/s42979-020-00216-w.

[13] J. Wu, P. Zhang, L. Zhang, W. Meng, J. Li, C. Tong, Y. Li, J. Cai, Z. Yang, J. Zhu et al., "Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results," *MedRxiv*, 2020, doi:10.1101/2020.04.02.20051136 .

[14] A. Bastug, H. Bodur, S. Erdogan, D. Gokcinar, S. Kazancioglu, B. D. Kosovali, B. O. Ozbay, G. Gok, I. O. Turan, G .Yilmaz, C. C. Gonen, F. M. Yilmaz. "Clinical and laboratory features of COVID-19: Predictors of severe prognosis," *International Immunopharmacology*, vol. 88, no. 106950, 2020, doi: 10.1016/j.intimp.2020.106950.

[15] D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi, and F. Cabitza, "Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study," *Journal of medical systems*, vol. 44, no. 8, pp. 1–12, 2020, doi:10.1007/s10916-020-01597-4.

[16] M. Kukar, G. Gunčar, T. Vovko *et al.*, "COVID-19 diagnosis by routine blood tests using machine learning," *Scientific Reports*, vol. 11, no. 10738, 2021, doi:10.1038/s41598-021-90265-9.

[17] K. Chadaga, S. Prabhu, K. V. Bhat, S. Umakanth and N. Sampathila,. "Medical diagnosis of COVID-19 using blood tests and machine learning," *Journal of Physics: Conference Series*, vol. 2161(1), 2022, doi:10.1088/1742-6596/2161/1/012017

[18] M. AlJame, I. Ahmad, A. Imtiaz, A. Mohammed, "Ensemble learning model for diagnosing COVID-19 from routine blood tests," *Informatics in Medicine Unlocked*, vol. 1, no. 21, 2020, doi: 10.1016/j.imu.2020.100449

[19] A. F. M. Batista, J.L. Miraglia, T. H. R. Donato, A. D. P. C. Filho, "COVID-19 diagnosis prediction in emergency care patients: A machine learning approach," *medRxiv*, 2020, doi: 10.1101/2020.04.04.20052092 [CrossRef].

[20] V. A. F. Barbosa, J. C. Gomes, M. A. Santana, J. E. A. Albuquerque, R. F. Souza, R. E. Souza, W. P. Santos, "Heg.IA: an intelligent system to support diagnosis of COVID-19 based on blood tests," *Research on Biomedical Engineering*, vol. 38, no. 1, pp. 99–116, 2022, doi: 10.1007/s42600-020-00112-5.

[21] M. Almansoor and N. M. Hewahi, "Exploring The Relation Between Blood Tests And Covid-19 Using Machine Learning," *International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pp. 1-6, 2020, doi: 10.1109/ICDABI51230.2020.9325673.

[22] F. Cabitza, A. Campagner, D. Ferrari, C. Di Resta, D. Ceriotti, E. Sabetta, A. Colombini, E. De Vecchi, G. Banfi, M. Locatelli, A. Carobene, "Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests," *Clinical Chemistry and Laboratory Med*icine, vol. 59, no. 2, pp. 421-431, 2021, doi: 10.1515/cclm-2020-1294.

[23] A. Akhtar, S. Akhtar, B. Bakhtawar, A.A. Kashif, N. Aziz, M. S. Javeid, "COVID-19 Detection from CBC using Machine Learning Techniques," *International Journal of Technology, Innovation and Management*, vol. 1, no. 2, 2021, doi:10.54489/ijtim.v1i2.22.

[24] O. O. Abayomi-Alli, R. Damaševičius, R. Maskeliūnas, S. Misra, "An Ensemble Learning Model for COVID-19 Detection from Blood Test Samples," *Sensors (Basel)*, vol. 22, no. 6, 2022, doi: 10.3390/s22062224.

[25] H. Gong, M. Wang, H. Zhang, M.F. Elahe, M. Jin, "An Explainable AI Approach for the Rapid Diagnosis of COVID-19 Using Ensemble Learning Algorithms," *Front Public Health*, vol. 10, no. 874455, 2022, doi:10.3389/fpubh.2022.874455.

[26] P. K. Roy, A. Singh, "COVID-19 Disease Prediction Using Weighted Ensemble Transfer Learning," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp.13-22, 2023, doi:10.9781/ijimai.2023.02.006.

[27] A. Andueza, M. Á. D. Arco-Osuna, B. Fornés, R. González-Crespo, J. M. Martín-Álvarez, "Using the Statistical Machine Learning Models

ARIMA and SARIMA to Measure the Impact of Covid-19 on Official Provincial Sales of Cigarettes in Spain," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 1, pp. 73-87, 2023, doi:10.9781/ijimai.2023.02.010.

[28] H. P. Cowley, M. S. Robinette, J. K. Matelsky *et al.* "Using machine learning on clinical data to identify unexpected patterns in groups of COVID-19 patients," *Scientific Reports,* vol. 13, no. 2236, 2023, doi:10.1038/s41598-022-26294-9

[29] J.T. Hancock and T.M. Khoshgoftaar, "Survey on categorical data for neural networks," *Jornal of Big Data*, vol. 7, no. 28, pp. 1-41, 2020, doi: 10.1186/s40537-020-00305-w.

[30] W. S. A. Farizi, I. Hidayah, M. N. Rizal, "Isolation Forest Based Anomaly Detection: A Systematic Literature Review," *2021 8th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, Semarang, Indonesia, 2021, pp. 118-122, doi: 10.1109/ICITACEE53184.2021.9617498.

[31] N. Pudjihartono, T. Fadason, A.W. Kempa-Liehr and J.M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," *Frontiers in Bioinformatics,* vol. 2, no. 927312, 2022, doi: 10.3389/fbinf.2022.927312.

[32] K. Dissanayake and M. G. Md Johar, "Comparative Study on Heart Disease Prediction Using Feature Selection Techniques on Classification Algorithms," *Applied Computational Intelligence and Soft Computing*, vol. 2021, no. 1, 2021. doi:10.1155/2021/5581806

[33] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, H. Liu, "Feature Selection: A Data Perspective," Association for Computing Machinery, vol. 50, no. 6, pp. 1-45, 2017, doi: 10.1145/3136625.

[34] V. V. Iyer and A. E. Yilmaz, "Using the ANOVA F-Statistic to Isolate Information-Revealing Near-Field Measurement Configurations for Embedded Systems," *2021 IEEE International Joint EMC/SI/PI and EMC Europe Symposium, Raleigh, NC, USA*, pp. 1024-1029, 2021. doi: 10.1109/EMC/SI/PI/EMCEurope52599.2021.9559360.

[35] A. O. Odetunm , O. A. Adejumo, A. T. Anake, "A study of Hepatitis B virus infection using chi-square statistic," *Journal of Physics Conference Series.* vol. 1734, no. 01, 2021, doi:10.1088/1742-6596/1734/1/012010.

[36] N. Carrara and J. Ernst, "On the estimation of mutual information," *Proceedings of The 39th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*,vol. 33, no.1, 2020, doi:10.3390/proceedings2019033031.

[37] I. M. Nasir, M. A. Khan, M. Yasmin, J. H. Shah, M. Gabryel, R. Scherer, R. Damaševičius, "Pearson Correlation-Based Feature Selection for Document Classification Using Balanced Training", Sensors , vol. 20, no. 23 pp. 1-18, 2020, doi : 10.3390/s20236793.

[38] F. Saberi-Movahed, M. Mohammadifard, A. Mehrpooya, M. Rezaei-Ravari, K. Berahmand, M. Rostami, S. Karami, et al., "Decoding Clinical Biomarker Space of COVID-19: Exploring Matrix Factorization-based Feature Selection Methods," *medRxiv [Preprint]*, 2021, doi: 10.1101/2021.07.07.21259699.

[39] C. A. Ramezan, "Transferability of Recursive Feature Elimination(RFE)-Derived Feature Sets for Support Vector Machine Land Cover Classification," *Remote Sensing*, vol. 14, no. 24, 2022, doi: 10.3390/rs14246218.

[40] C. Zhang, Y. Yi, L. Wang, X. Zhang, S. Chen, Z. Su, S. Zhang, Y. Xue, "Estimation of the Bio-Parameters of Winter Wheat by Combining Feature Selection with Machine Learning Using Multi-Temporal Unmanned Aerial Vehicle Multispectral Images," *Remote Sensing,* vol. 16, no. 3, pp. 1-22, 2024, doi:10.3390/rs16030469

[41] A. Suruliandi, K. Ranjini, S. P. Raja, "Balancing Assisted Reproductive Technology Dataset for Improving the Efficiency of Incremental Classifiers and Feature Selection Techniques," *Journal of Circuits, Systems, and Computers, World Scientific*, vol. 30, no. 06, 2130007,2021, doi:10.1142/S0218126621300075

[42] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Computer Science*, vol. 2, no. 160, 2021, doi:10.1007/s42979-021-00592-x.

[43] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189-215, 2020, doi: 10.1016/j.neucom.2019.10.118.

[44] W. G. Gadallah, N. M. Omar and H. M. Ibrahim, "Machine Learning-based Distributed Denial of Service Attacks Detection Technique using New Features in Software-defined Networks," *International Journal of Computer Network and Information Security,* vol. 3, pp. 15-27, 2021, doi:10.5815/ijcnis.2021.03.02.

[45] C. N. Villavicencio, J. J. E. Macrohon, X. A. Inbaraj, J. H. Jeng, J. G. Hsieh, "COVID-19 Prediction Applying Supervised Machine Learning Algorithms with Comparative Analysis Using WEKA," *Algorithms*, vol.14, no. 7, 2021, doi:10.3390/a14070201.

[46] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet]*, vol. *9*, no. 1 pp.381-386, 2020.

[47] M. Rohini, K. R. Naveena, G. Jothipriya, S. Kameshwaran, M. Jagadeeswari, "A Comparative Approach to Predict Corona Virus Using Machine Learning," *Proceedings of the International Conference on Artificial Intelligence and Smart Systems*, Coimbatore, India, 2021, pp. 331-337, doi: 10.1109/ICAIS50930.2021.9395827.

[48] T. Rymarczyk, E. Kozłowski, G. Kłosowski, K. Niderla, "Logistic Regression for Machine Learning in Process Tomography," *Sensors*, vol. 19, no. 15, 2019, doi:10.3390/s19153400.

[49] S. A. Hicks, I. Strümke, V. Thambawita, M. Hammou, M. A. Riegler, P. Halvorsen, S. Parasa, "On evaluation metrics for medical applications of artificial intelligence," *Scientific Reports,* vol. 12, no. 1. 2022, doi: 10.1038/s41598-022-09954-8.

[50] R. A. Rayan, A. Suruliandi, S.P. Raja, H. B. F. David, "A survey on an analysis of big data open source datasets, techniques and tools for the prediction of corona virus disease," *Journal of Circuits, Systems and Computers,* vol. 32, no. 12, 2023, doi:10.1142/S0218126623300039.

[51] J. White, S. D. Power, "k-Fold Cross-Validation Can Significantly Over-Estimate True Classification Accuracy in Common EEG-Based Passive BCI Experimental Designs: An Empirical Investigation," *Sensors (Basel)*, vol. 23, no. 13, 2023, doi: 10.3390/s23136077.

A. Suruliandi

A. Suruliandi is currently the Professor and Head at Manonmaniam Sundaranar University, Tamil Nadu, India. He previously worked as a lecturer at Kamaraj College, Tamil Nadu. With a Bachelor's degree in Electronics and Communication Engineering, a gold medal in Master's degree, and a Ph.D., he has published over 80 research papers and actively contributes to the academic community through lectures, presentations, and peer reviewing. His research focuses on nurturing young minds and guiding research scholars in exploring innovative ideas.

R. Ame Rayan

R. Ame Rayan is currently pursuing Ph.D. at Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India. Prior to pursuing her Ph.D., she worked as an Assistant Professor in the Computer Science Department at Holy Cross Home Science College in Thoothukudi, Tamil Nadu, India. She has completed her undergraduate studies at St. Mary's College in Thoothukudi, Tamil Nadu, India. She obtained her Master's degree in Computer Applications (MCA) from Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India.

S. P. Raja

S. P. Raja is born in Sathankulam, Tuticorin District, Tamilnadu, India. He completed his schooling in Sacred Heart Higher Secondary School, Sathankulam, Tuticorin, Tamilnadu, India. He completed his B. Tech in Information Technology in the year 2007 from Dr. Sivanthi Aditanar College of Engineering, Tiruchendur. He completed his M.E. in Computer Science and Engineering in the year 2010 from Manonmaniam Sundaranar University, Tirunelveli. He completed his Ph.D. in the year 2016 in the area of Image processing from Manonmaniam Sundaranar University, Tirunelveli. Currently he is working as an Associate Professor in the School of Computer Science and Engineering in Vellore Institute of Technology, Vellore, Tamilnadu, India.