A 2D Clustering Based Hotspot Identification Approach for Spatio-Temporal Crime Prediction

Muhammad Faisal Buland Iqbal[1], Aman Ullah[2], Ahmed Alhomoud[3], Tariq Hussain[4]*, Razaz Waheeb Attar[5], Jianquan Ouyang[1]*, Mrim M. Alnfiai[6], Wesam Atef Hatamleh[7]

[1] Key Laboratory of Intelligent Computing & Information Processing, Ministry of Education, Xiangtan University, Xiangtan 411105 (China)
[2] Faculty of Science and Technology, Virtual University of Pakistan (Pakistan)
[3] Department of Computer Science, College of Science, Northern Border University, Arar 91431 (Saudi Arabia)
[4] School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou, 310018 (China)
[5] Management Department, College of Business Administration, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671 (Saudi Arabia)
[6] Department of Information Technology, College of Computers and Information Technology, Taif University, Taif P.O. Box 11099, Taif, 21944 (Saudi Arabia)
[7] Department of Plant Production, College of Food and Agriculture, King Saud University, Riyadh (Saudi Arabia)

* Corresponding author: uom.tariq@gmail.com (T. Hussain), oyjq@xtu.edu.cn (J. Ouyang).

Received 20 May 2023 | Accepted 13 August 2024 | Early Access 4 October 2024

Abstract

This research introduces a method for predicting where crimes will occur based on clustering activity in the area. Hotspots, or locations with a disproportionately high number of crimes, are located by a combination of spatial and temporal grouping methods employed by this strategy. Crime forecasting models use these hotspots to predict where crimes will occur. The approach's efficacy is tested using actual crime data, and it successfully predicts future crimes in high-crime zones. Law enforcement agencies can use the proposed method to protect the public better, and it shows promise as a tool for crime prediction. Academic research into the topic of foreseeing criminal behavior is a newer development. Researchers in this discipline have discovered that criminal behavior has universal patterns. These patterns may help law enforcement agencies plan for criminal activities. Predictive policing, hotspot analysis, and geographical profiling are examples of when crime forecasting has been useful. Several aspects of the census, such as the average yearly income and literacy rate, are related to the prevalence of crime in a certain area. Indicators of potentially criminal behavior, these characteristics may be seen as markers. This investigation aims to discover if any clues can be gleaned from past criminal behavior that may be utilized to forecast future criminal behavior. Using machine learning and 2-D Hotspot analysis, we propose a method for the spatiotemporal prediction of criminal activity within the scope of this study. Clustering is a method used in 2-dimensional hotspot analysis. Methods of modern categorization, both with and without hotspot analysis, are used to evaluate the suggested model's efficacy. It is found that the model that incorporates hotspot analysis performs better than the one that does not.

Keywords

2-D Hotspot Analysis, Crime Prediction, Forensics, Predictive Policing, Spatial, Temporal Clustering.

DOI: 10.9781/ijimai.2024.10.001

I. Introduction

THE clustering-based hotspot identification is a method for predicting crime by identifying areas, or "hotspots," where crime is likely to occur [1]-[4]. This approach involves grouping similar crime incidents using clustering algorithms and then identifying areas where these clusters are concentrated. These hotspots can then be used to predict where future crime may occur and to focus crime prevention efforts. This approach is based on the idea that crime tends to be concentrated in specific areas and that identifying these areas makes it possible to predict and prevent future crime [5]-[9].

Please cite this article as:
M. F. B. Iqbal, A. Ullah, A. Alhomoud, T. Hussain, R. W. Attar, J. Ouyang, M, M. Alnfia, W. A. Hatamleh. A 2D Clustering Based Hotspot Identification Approach for Spatio-Temporal Crime Prediction, International Journal of Interactive Multimedia and Artificial Intelligence, (2024), http://dx.doi.org/10.9781/ijimai.2024.10.001

There are several reasons why clustering-based hotspot identification is a valuable approach for crime prediction. One primary motivation is that it identifies areas where crime is likely to occur, which can then be used to focus crime prevention efforts. This is particularly useful for law enforcement agencies, as it allows them to allocate resources more effectively and to target specific areas where crime is known to occur [10]-[14].

Another motivation for using this approach is that it can help identify crime patterns and trends that may not be immediately obvious from raw crime data. By grouping similar crime incidents using clustering algorithms, it is possible to identify patterns and trends that may not be apparent when looking at the data individually. This can provide valuable insights into the causes of crime and help inform crime prevention strategies.

Additionally, clustering-based hotspot identification can be combined with other crime prediction methods to improve the accuracy of crime predictions. This approach can identify areas where crime is likely to occur. Other methods, such as time-series analysis or machine learning, can predict the likelihood of crime occurring in those areas [15]-[19].

Lastly, using clustering-based hotspot identification can help evaluate the effectiveness of crime prevention strategies. Monitoring changes in crime hotspots over time makes it possible to determine whether a particular strategy is having the desired effect of reducing crime in the targeted area. A research problem for a clustering-based hotspot identification approach for crime prediction is investigating the effectiveness of a clustering-based hotspot identification approach in predicting and preventing crime in urban areas. This research problem aims to determine the efficacy of using clustering-based hotspot identification to predict and prevent crime in urban areas. The research could involve collecting crime data from a specific urban area, applying clustering algorithms to identify hotspots, and comparing the accuracy of crime predictions using this approach to other methods. Additionally, the research could evaluate the effectiveness of crime prevention strategies targeted at identified hotspots. This research problem could be refined by focusing on specific aspects of the approach, such as:

- Identifying the most suitable clustering algorithm for crime hotspot identification.
- Comparing the performance of different clustering algorithms in identifying crime hotspots.
- Investigating the impact of different spatial and temporal resolutions on the accuracy of crime predictions.
- Evaluating the effectiveness of crime prevention strategies in reducing crime in identified hotspots.
- Examining the challenges and limitations of implementing clustering-based hotspot identification in practice.

A research objective for the proposed research problem "Investigating the effectiveness of clustering-based hotspot identification approach in predicting and preventing crime in urban areas is to evaluate the efficacy of clustering-based hotspot identification approach in predicting and preventing crime in urban areas by comparing its performance to other crime prediction methods and assessing the effectiveness of crime prevention strategies targeted at identified hotspots. This research objective is focused on determining the effectiveness of using clustering-based hotspot identification to predict and prevent crime in urban areas. It aims to do this by comparing the accuracy of crime predictions made using this approach to other methods and evaluating the effectiveness of crime prevention strategies targeted at identified hotspots. The main objectives of this study are:

- To determine the most suitable clustering algorithm for crime hotspot identification.
- To compare the performance of different clustering algorithms in identifying crime hotspots.
- To investigate the impact of different spatial and temporal resolutions on the accuracy of crime predictions.
- To evaluate the effectiveness of crime prevention strategies in reducing crime in identified hotspots.
- To examine the challenges and limitations of implementing clustering-based hotspot identification in practice.

II. Related Work

As crime prediction tools improve, geographical trends emerge. Recurring concepts with better crime forecasting, it's evident that crime tracks geographical trends. These tendencies may help authorities predict and prevent crime. Predicting criminal behaviour has helped police prevent crimes. Crime prediction is used in predictive policing, Hotspot analysis, and geographical profiling. Crime rates correspond with census statistics like wealth and literacy. These parameters predict crime: crime mapping, geo-profiling, forecasting, and hotspot analysis. Crime is connected to dates, weather, latitude, and census population numbers (such as median household income or literacy rate). Crime may result from any of these. This study [8] used criminal symptoms. This study proposes a 2-D Hotspot analysis approach for predicting criminal behaviour. Two-dimensional hotspot experiments showed the link between past and present criminality. This technique uses cluster analysis to predict crime locations and times. The suggested technique is compared to 2-D Hotspot-based machine learning classification algorithms. 2-D hotspot analysis uses clustering. The author analyzed the recommended model's performance using cutting-edge classification algorithms with and without hotspot analysis. The model testing is successful with or without hotspot analysis.

In this study [10], the author used a Support Vector Machine-based spatial clustering approach to predict where violent crimes will occur. The author used Principal Component Analysis on six separate violent crime datasets covering 2014 through 2019 to predict where crimes will occur the next day in Lagos, Nigeria. With an accuracy of 82.12%, the data collected is reliable enough to be used in forecasting violent crime. This study [5] offered a summary of methods for analyzing and projecting criminal activity based on available data. The ability to anticipate where and when crimes will occur will significantly aid law enforcement in lowering crime rates. By analyzing historical crime data and predicting where and when future crimes will occur, law enforcement agencies can be better prepared to deal with them [6].

Police in high-crime regions can significantly benefit from predictive hotspot mapping. Popular approaches like kernel density estimation (KDE) exist even though time dimensions in crime are frequently ignored. This study [20] presented a spatio-temporal paradigm for identifying, evaluating, and predicting hotspots, building on prior work in related disciplines. The author of this study [21] presented a data-mining strategy for predicting and classifying criminal activity in San Francisco. Both the K-NN classifier and the Naive Bayes classifier are compared and contrasted in this technique. The K-NN classifier utilized not one but two separate processes: uniform and inverse.

Given the constraints on law enforcement resources, it seems sensible to prioritize those locations with the highest crime rates. As part of this study [22], the author renounced the more common "higher than" average criteria in favor of a more practical approach. The outcomes show that classification performs better than clustering on this dataset. Criminality is a worldwide issue that calls for collective

action from all members of society. The techniques of Random Forest (RF) and Extremely Randomized Trees (XRT) are known to be used to reduce bias and variability (ERT). In this study [11], the author described a hybrid method that combines the best aspects of RF and ERT, such as bootstrap aggregation and random feature selection. This novel hybrid approach was then evaluated in terms of RF and ERT. The findings showed that this hybrid approach was superior to its rivals in terms of both prediction accuracy and computing complexity. Planning for criminal acts is challenging since they can happen anywhere and at any moment. The research proposes a crime prediction model by analyzing and comparing three widely used prediction classification algorithms—Naive Bayes, Random Forest, and Gradient Boosting Decision Tree. The analysis and prediction of this study [2] can benefit security agencies in making optimal use of existing resources, foreseeing criminal activity at specific periods, and overall doing good for society.

The recurrent neural network model has been extensively tested and shown to be effective in discovering patterns in time series. This study [23] recommended the spatiotemporal neural network (STNN), which integrates geographical and temporal data for precise crime hotspot predictions. The primary goal of this study [7] was to conduct a literature review to gather, synthesize, and evaluate existing methods for identifying and predicting crime hotspots throughout time and space (SLR). The SLR aims to provide a fundamental platform for future research of crime hotspot detection and prediction applications while identifying different issues related to their accuracy.

Because of the increase in crime data collection and the introduction of data analytics, new approaches have been created to glean information from criminal records to understand criminal behaviour better and, perhaps, prevent future crimes. Clustering and association rule mining algorithms make up the bulk of these approaches, while prediction models of criminal behaviour make up a smaller subset. In this study [13], the author discussed crime prediction models and antisocial behaviour prediction models based on LSOA (Lower Layer Super Output Area) codes (a geographical classification system used by the UK police). Classification models based on supervised machine learning have been developed and deployed for predictive modeling purposes. This research [24] used big data and AI to forecast violent events in the future. The study aimed to categorize and assess a crime prediction technique that uses a five-part classification scheme. This research [25] described great accuracy and efficiency in forecasting hot spots despite using a different data collection method and kind of crime than the original algorithm intended. The data suggests several approaches are employed to forecast major and minor crimes.

It is now a challenging task to anticipate where criminal activity will surge. Algorithms that can process large crime datasets are necessary for accurate predictions about future crime hotspots. Spatial and temporal data mining is beneficial when dealing with crime statistics. This study [15] aimed to create a model for foreseeing instances of serial crime using a spatial clustering technique based on sparse matrix analysis. Crime data spanning four years (2010-2014) in the main cities of India (Delhi, Mumbai, Kolkata, and Chennai) is clustered using three different methods before the sparse matrix analysis-based spatial clustering methodology is applied. In the first phase of this study [16], machine learning algorithms were used to extract knowledge from the large datasets and find the hidden relationships among data, which was then used for reporting and discovering crime patterns, which is a significant source of information for a crime analyst to analyze these crime networks through several interactive visualizations for crime prediction, and is thus very helpful in preventing crime. The ultimate goal of this study [4] was to create data-driven algorithms capable of predicting future criminal behaviour. First, the author looks at Halifax, Nova Scotia, from a geographical perspective. The author proposed a density-based spatial grouping technique to locate locations with high crime rates and applied a reverse geocoding method to retrieve Open Street Map spatial data [2].

Predicting where crimes are likely to occur is crucial for protecting the public. In addition, this study [12] provided a generic framework for adapting the spatial data classification job to datasets in other geographical domains analogous to the crime datasets used for analysis. In this study [9], the suggested approach forecasts areas likely to be affected by criminal activity and divides them into several "hotspot" types. In this study [26], [27], spatial data analysis using a geographic information system (GIS) may help law enforcement agencies recognize trends in criminal activity, identify potential trouble locations, and plan for future patrols. This study [28] introduces a spatial-temporal crime prediction method that uses a structured crime categorization algorithm to predict criminal behaviour in a given neighborhood.

Hotspot identification and other exploratory approaches can help map out areas with a high crime rate so that resources can be directed there. Even though several methods have been developed to pinpoint these troublesome locations, there has been a shortage of research evaluating their efficacy, especially regarding their ability to pinpoint complex-shaped crime hotspots. Therefore, this study [29] investigated several techniques for locating these hotspots, with an eye toward density and organization. The results show that the assessment framework and indicators presented are helpful for quantitatively comparing approaches by defining the size, concentration, and form features of the observed hotspots. In the police department, GIS maps hotspots, compiles crime data (CompStat), and conducts spatial profiling. The main objective of this study was [14] to gather more information on criminal behaviour, especially in high-crime areas.

The world's governments are working together to use innovative approaches to solve these issues. The study's results [30] were used to calculate crime rates and expedite handling of criminal cases. Informational infrastructure and analysis are being adopted rapidly by intelligent cities to inform public safety decisions better. In this study [31], the author analyzed the framework using data from two cities, Natal (Brazil) and Boston (US), which includes twelve different types of crime. The author used the standard police prediction approach, also used in Natal, as this starting point. According to the findings, the prediction strategy outperforms the status quo by a 1.6-3.1. In this research [6], the author constructed a model that can foretell the time and place of criminal activity as well as the nature of that activity. Rising crime rates in emerging countries are linked to unstable psychological and economic environments.

III. Material and Methods

A research methodology for investigating the effectiveness of a clustering-based hotspot identification approach in predicting and preventing crime in urban areas could involve the following steps:

Data collection: Collect crime data from a specific urban area for a specified time. This data should include information such as the type of crime, location, and time of occurrence.

Data pre-processing: Clean and prepare the data for analysis by removing any missing or irrelevant information.

Clustering: Apply clustering algorithms to the data to group similar crime incidents and identify hotspots where clusters are concentrated.

Crime prediction: Use the identified hotspots to predict where future crime may occur by applying time-series analysis or machine learning techniques.

Comparison: Compare the accuracy of crime predictions made using the clustering-based hotspot identification approach to other methods.

Crime prevention strategy evaluation: Implement crime prevention strategies in identified hotspots and evaluate their effectiveness in reducing crime.

Challenges and limitations: Examine the challenges and limitations of implementing the clustering-based hotspot identification approach.

Fig. 1 shows the flowchart of the current study:

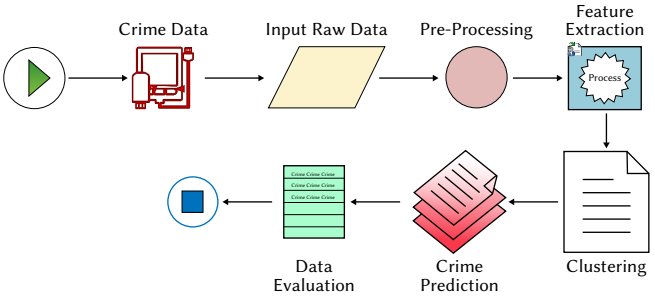


Fig. 1. Flowchart of Study.

A. Data Collection

For the proposed study, we performed our experiments on a dataset available on Kaggle called the "Crime in Chicago" dataset. This dataset includes crime incidents that occurred in Chicago from 2001 to the present and includes information such as the type of crime, location, and time of occurrence. This dataset has 22 total variables and 529531 records. Except for homicides, where data exists for each victim, this dataset contains information about crimes reported to have taken place in Chicago between 2001 and the current day, minus the most recent seven days. Information is gathered from the CLEAR (Citizen Law Enforcement Analysis and Reporting) database maintained by the Chicago Police Department. Addresses are only displayed at the block level, and individual places are not disclosed to preserve the privacy of crime victims. Some of this information comes from tips given to the Police Department without verification. There is always the risk of mechanical or human error, and further examination may result in a revision of the preliminary criminal categorization. As a result, you shouldn't use the data for long-term comparisons, and the Chicago Police Department makes no promises about the data's correctness, completeness, timeliness, or proper sequencing. Table I shows the features of the current dataset:

TABLE I. The Features of the Current Dataset

Feature	Description	Type
ID	Unique ID of crime	Input Variable
Type of Crime	Type of Crime, i.e., Robbery, Rape etc.	Input Variable
Criminal Possession	Whether criminal was justified or not	Input Variable
Domestic	Whether it's domestic or international	Input Variable
Location	Description of Location of crime	Input Variable
Outcome	Clustered Output	Output Variable

B. Data Visualization

Data visualization is an essential step in analyzing the Chicago crime dataset. It allows researchers to quickly identify patterns and trends in the data, providing valuable insights into the causes of crime and informing crime prevention strategies.

One way to visualize the Chicago crime dataset is by creating maps showing the distribution of crime incidents across the city. This can be done by plotting each crime incident's location on a Chicago map and using color coding or symbols to indicate the type of crime. This allows researchers to quickly identify areas where crime is concentrated and compare the distribution of different types of crime.

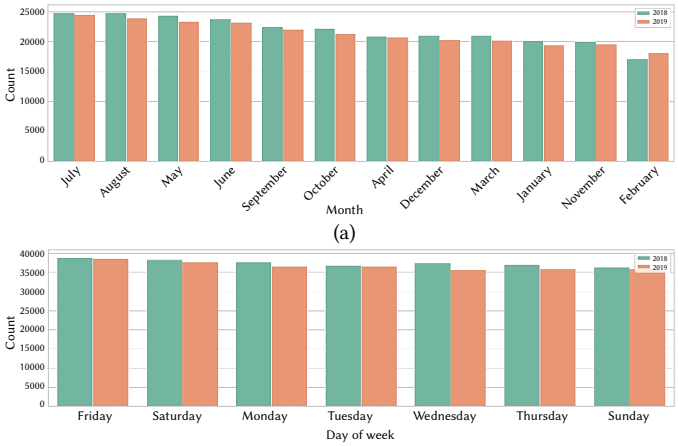


Fig. 2. Crime Rate in Each year on different days.

Another way to visualize the Chicago crime dataset is by creating charts and graphs. For example, line charts or bar charts could show the number of crime incidents that occurred over time or by month. This can help researchers identify crime trends and understand when and where crime is most likely.

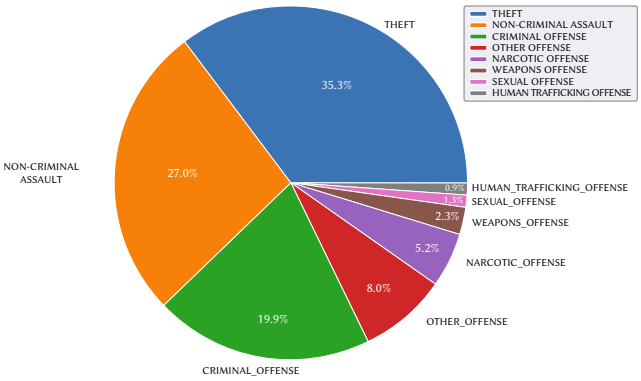


Fig. 3. Distribution of Type of Crime.

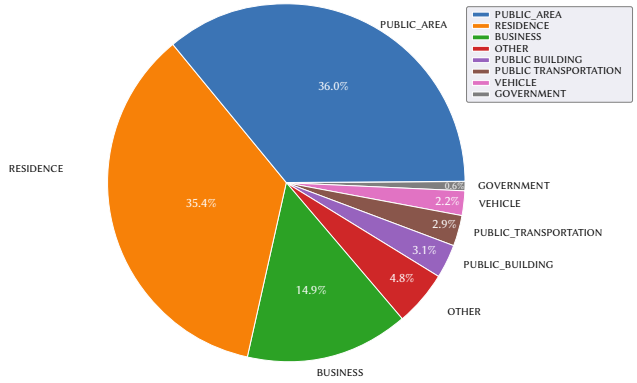


Fig. 4. Locations of Crime.

Additionally, data visualization can be used to show the relationship between different factors and crime rates. For example, heat maps can show the relationship between crime rate and other factors such as income, population density, education level, etc.

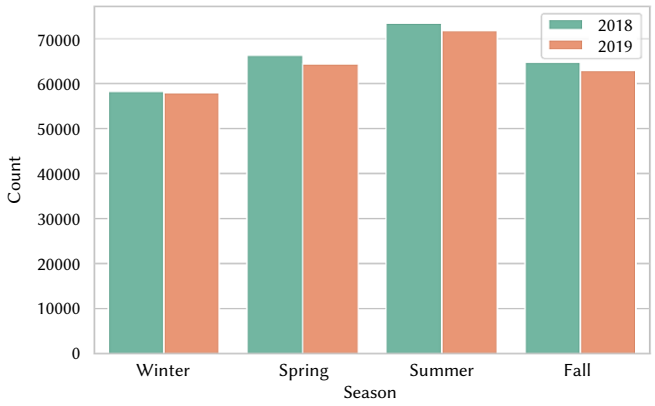


Fig. 5. Season Wise Crimes.

Overall, data visualization is an essential tool for understanding the Chicago crime dataset and can identify patterns and trends in crime that may not be immediately obvious from the raw data. Using different visualization techniques to uncover different insights and provide different data perspectives is also essential. Fig. 2 (a, b), Fig. 3, Fig. 4, Fig. 5 Fig. 6 and Fig. 7 shows the visualization of the dataset.

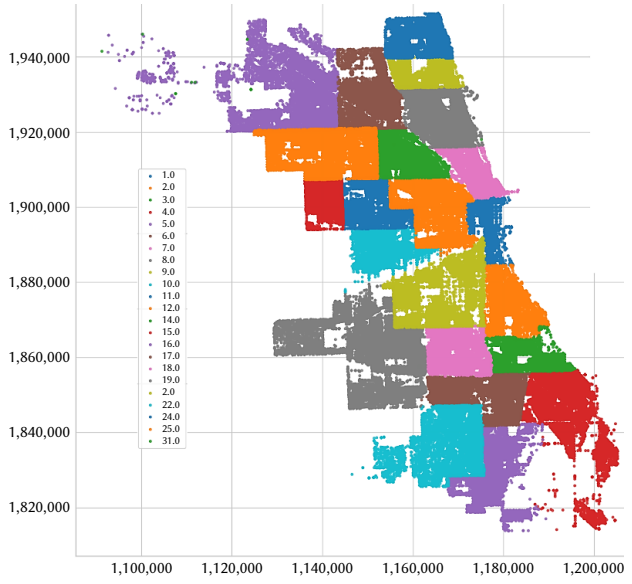


Fig. 6. Crimes by District.

C. Data Pre-Processing

Data pre-processing is an essential step in analyzing the Chicago crime dataset. It involves cleaning and preparing the data for analysis by removing any missing or irrelevant information. This step is crucial to ensure the data is accurate and suitable for research.

The following are some common data pre-processing steps that can be applied to the Chicago crime dataset:

- Data cleaning: Before proceeding with the analysis, we performed steps to handle missing values, outliers, and duplicate entries. We achieved this by identifying and removing any rows or columns with missing values or by imputing missing values.

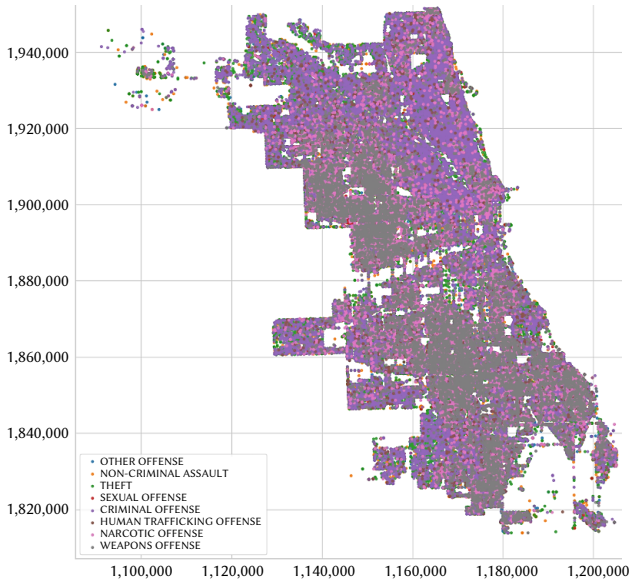


Fig. 7. Crimes by Types of Crime.

- Data transformation: This step involves transforming the data into a suitable format for analysis. For example, the location data may need to be transformed into longitude and latitude coordinates.
- Data normalization: This step involves transforming variables to a standard scale, useful when comparing variables measured in different units. We opted for normalization over other techniques like Box-Cox, standardization, Yeo-Johnson, or robust standardization because it is effective when the features in the dataset have different ranges and units; normalization does not make any assumptions about the distribution of the data, machine learning algorithms, especially distance-based methods like K-means, perform better when the features are normalized and last but not the least normalization is sensitive to outliers.
- Data reduction: This step involves reducing the number of variables in the dataset by removing irrelevant or redundant variables, which can help reduce the data's complexity and improve the analysis's accuracy.
- Data integration: this step involves combining different data sources into a single dataset, which can provide a more complete picture of the data and be useful for cross-referencing information.

D. Clustering Algorithm

Clustering algorithms are vital to the Clustering Based Hotspot Identification Approach for Crime Prediction. They are used to group similar crime incidents and to identify hotspots where crime is concentrated.

Some of the most commonly used clustering algorithms for crime hotspot identification include:

K-means: This widely used clustering algorithm groups data points into k clusters based on similarity. It is a simple and efficient algorithm but can have difficulty handling non-globular clusters and high-dimensional data.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): This algorithm groups data points into clusters based on the density of data points in the area. It can identify clusters of arbitrary shape and handle data with varying densities.

Hierarchical Clustering: This algorithm creates a hierarchy of clusters by successively merging or splitting smaller clusters. It can be used to identify clusters at different levels of granularity and helps explore the data structure.

1. K-Means

K-means is a popular clustering algorithm that groups data points into k clusters based on similarity. The basic idea behind the algorithm is to partition a dataset into k clusters, where each cluster is defined by the mean of the data points that belong to it.

The K-means algorithm consists of the following steps:

- Initialization: Select k initial centroids, which are the means of the clusters. These centroids can be selected randomly or using a specific method such as k-means++.
- Assignment: Assign each data point to the cluster whose centroid is closest to it. This is done by calculating the distance between each data point and each centroid and assigning the data point to the cluster whose centroid is closest.
- Update: Recalculate the centroid for each cluster by taking the mean of all the data points that belong to that cluster.
- Repeat steps 2 and 3 until the centroids no longer change or the maximum number of iterations is reached.

The central equation used in the K-means algorithm is the distance metric. The most common distance metric used in K-means is the Euclidean distance, the square root of the sum of the squared differences between the coordinates of two points.

The equation (1) for the Euclidean distance between two points, x and y, is:

$$d(x,y) = sqrt((x1 - y1)^{\wedge 2} + \cdots + (xn - yn)^{\wedge 2}) \quad (1)$$

Where x and y are two n-dimensional vectors.

In addition to the distance metric, K-means uses the mean of the data points in each cluster as the new centroid. The equation for the mean of a set of data points is eq. (2):

$$New\ centroid = \left(\frac{1}{n}\right) * (x1 + x2 + \cdots + xn) \quad (2)$$

Where x1, x2... xn are the data points in the cluster, and n is the number of data points in the cluster. Overall, the K-means algorithm is simple, easy to implement, and computationally efficient, but it can have difficulty handling non-globular clusters and high-dimensional data.

2. DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that groups data points into clusters based on the density of data points in the area. Unlike k-means and other centroid-based clustering algorithms, DBSCAN does not require the number of clusters to be specified in advance.

The DBSCAN algorithm defines clusters as dense data points separated by lower-point density areas. The algorithm starts by selecting a random point from the dataset and then searches for all points within a certain distance (EPS) from that point. If the number of points in this neighborhood (MinPts) is greater than or equal to a specified threshold, a new cluster is created, and all points that are reachable from the starting point within the EPS distance are added to the cluster. The process is then repeated for each point in the new cluster until all density-connected points have been added.

DBSCAN has two main parameters:

- EPS (epsilon): The maximum distance between two points to be considered as part of the same cluster
- MinPts: The minimum number of data points in a cluster. DBSCAN algorithm is efficient and effective in identifying clusters of arbitrary shape, and it can handle data with varying densities. However, it can be sensitive to the choice of parameters and may not be suitable for high-dimensional data or data with varying densities.

3. Hierarchal Clustering

Hierarchical Clustering is a clustering method that creates a hierarchy of clusters by successively merging or splitting smaller clusters. It can be used to identify clusters at different levels of granularity and helps explore the data structure. There are two main types of hierarchical clustering: Agglomerative and Divisive.

Agglomerative Hierarchical Clustering: This method starts by treating each data point as a separate cluster and then repeatedly merges the closest pair of clusters until all points are in the same cluster.

The algorithm follows these steps:

- Start by treating each data point as a single-point cluster.
- Compute the pairwise distances between all points and clusters
- Merge the closest pair of clusters into a new cluster.
- Repeat steps 2 and 3 until all points belong to a single cluster.
- The resulting tree, called a dendrogram, shows the merging process and the data structure.

The central equation used in Agglomerative Hierarchical Clustering is the linkage criterion. The linkage criterion is a measure of the distance between two clusters. The most common linkage criteria are:

Single linkage: The distance between two clusters is defined as the minimum distance between any two points, one from each cluster.

$$d(A,B) = \min\big(d(i,j)\big)\ for\ all\ i\ in\ A, j\ in\ B$$

Complete linkage: The distance between two clusters

E. Crime Prevention Strategy Evaluation

The most prevalent type of unsupervised learning is called "clustering." Clustering is a method for organizing data in which related observations are grouped and, unlike ones, kept as far apart as possible without labels. An algorithm's efficacy cannot simply be measured by tallying the number of errors or its precision and recall, as in supervised learning.

The average distance between cluster nodes measures cluster similarity or dissimilarity. To evaluate a clustering algorithm's effectiveness, it must be able to group similar observations and separate those that differ. The Silhouette coefficient and Dunn's Index are two commonly used metrics for evaluating clustering techniques.

1. Silhouette Coefficient

The Silhouette Coefficient is a metric employed to gauge the quality of the clusters created by a clustering algorithm. It ranges from -1 to 1, where a high value indicates that the object is well-matched to its cluster and poorly matched to neighboring clusters. It measures how well each data point is clustered by comparing the distance between the data point and its cluster members to the distance between the data point and other clusters.

The Silhouette Coefficient is higher when the data points within a cluster are close to each other. This is because the average distance between a data point and its cluster members will be lower. The Silhouette Coefficient is higher when the data points in different clusters are far apart. This is because the smallest average distance between a data point and other clusters will be higher. It considers both the compactness and separation of the clusters. The Silhouette Coefficient's range makes it easy to interpret the quality of clusters, which is crucial for understanding crime hotspots in an actionable way. Unlike other metrics, the Silhouette Coefficient does not assume any particular geometry of the clusters, making it versatile for different types of data distributions. It provides insight into the distance between the resulting clusters. More distant clusters lead to better clustering, vital for accurately identifying distinct crime hotspots.

2. Indices by Dunn

A clustering technique can also be judged using Dunn's Index (DI), which is a similar metric. Minimal inter-cluster distance divided by the most significant cluster size equals Dunn's Index. Note that a higher DI value results from greater separation between clusters (more considerable inter-cluster distances) and smaller cluster sizes (more compact clusters). A larger DI indicates better clustering. It presumes that compact, well-separated clusters indicate superior clustering.

3. Inter and Intra Cluster Distances

Intra-cluster distance is the distance between data points within the same cluster, as shown in Fig 8. It measures how well the clustering algorithm has grouped similar data points. A lower intra-cluster distance indicates better clustering results, as the data points within a cluster are close to each other. Inter-cluster distance refers to the distance between different clusters. It measures how well the clustering algorithm has separated dissimilar data points. A higher inter-cluster distance indicates better clustering results, as the clusters are well-separated and distinct. There are several types of inter-cluster distance, including Minimum distance, which is the minimum distance between any two data points in different clusters. Maximum distance: The maximum distance between any two data points in different clusters. Average distance: The average distance between all pairs of data points in different clusters. Centroid distance: The distance between the centroids of two different clusters.

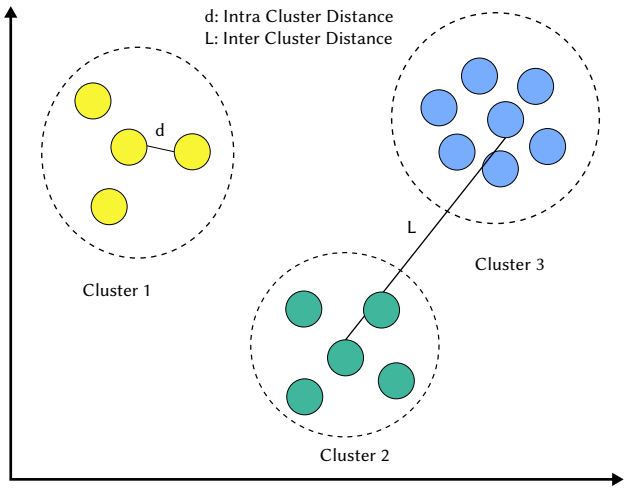


Fig. 8. Inter and intra-cluster.

IV. Results and Discussion

After the data has been appropriately preprocessed, we present clustering techniques, a crucial aspect of the research. Several methods, including hierarchical clustering, DBSCAN, and K-means, will be used to locate hotspots of concentrated criminal activity and group-related crime episodes during the experiments. For instance, the K-means method uses an iterative centroid assignment and recalculation process to divide the data into clusters based on similarity. Alternatively, DBSCAN uses density-based clustering to find any shape of the cluster. Various validation methods will be used to ascertain the ideal number of clusters, such as the Davies Bouldin Score, CALINSKI H Score, Elbow Method, and Silhouette Score. These methods assist in determining the applicability and validity of various cluster sizes. To provide a thorough understanding of the outcomes, it is crucial to improve the justification for the use of these validation techniques as well as their explanation. Concurrently, the study will employ dendrograms and other connection strategies in hierarchical clustering, expanding the range of clustering methodologies utilized. The clustering results are validated using silhouette scores; a table comparing the number of clusters obtained from various clustering techniques with their corresponding validation procedures and time complexity was created to support the findings.

A. K-Means Clustering

K-means is a popular clustering algorithm that groups data points into k clusters based on similarity. The basic idea behind the algorithm is to partition a dataset into k clusters, where each cluster is defined by the mean of the data points that belong to it. Fig. 9 shows the results of the KMEANS.

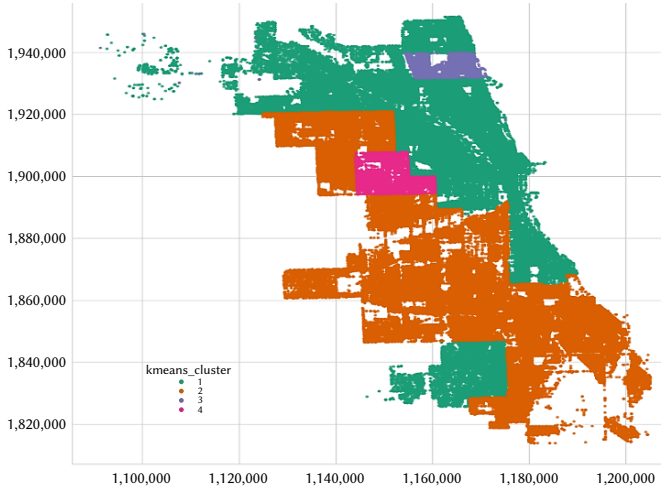


Fig. 9. KMeans Clustering of Crimes by District.

B. DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that groups data points into clusters based on the density of data points in the area. Unlike k-means and other centroid-based clustering algorithms, DBSCAN does not require the number of clusters to be specified in advance. Fig. 10 shows the results of DBSCAN.

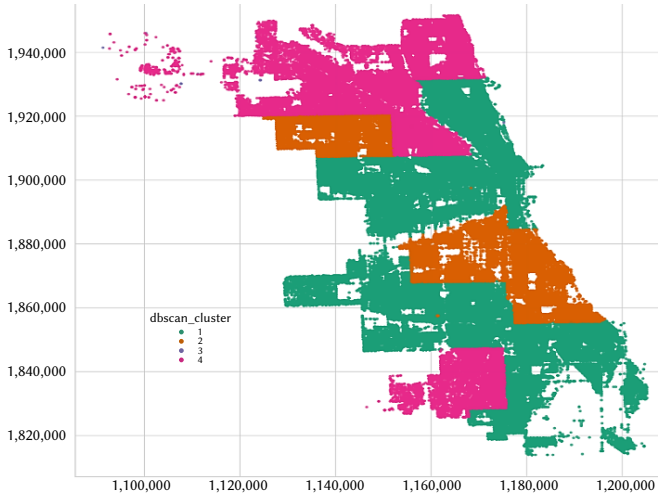


Fig. 10. DBScan Clustering of Crimes by District.

C. Hierarchal Clustering

Hierarchical Clustering is a clustering method that creates a hierarchy of clusters by successively merging or splitting smaller clusters. It can be used to identify clusters at different levels of granularity and helps explore the data structure. There are two main types of hierarchical clustering: Agglomerative and Divisive. Fig. 11 shows the Hierarchical clustering.

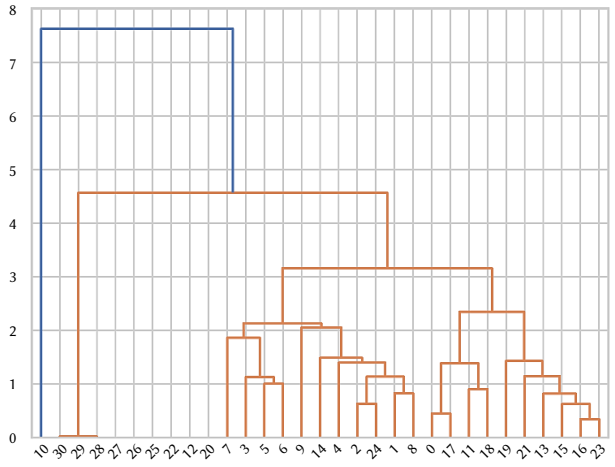


Fig. 11. Shows the Hierarchal Clustering.

D. Comparison of Other Clustering Techniques on the Crime Dataset

The dataset for the analysis is downloaded from Kaggle, and the data is suitable for EDA and machine learning. Hierarchal clustering techniques like agglomerative and divisive clustering were deployed and different visualization techniques were used, with the difference being in the same agglomerative clustering. After these clustering techniques, DBSCAN, OPTICS, and DENCLUE were applied to get the optimal clustering value, validated the clustering results with the Elbow Method with the Distortion technique, and found the optimal value of the same cluster processes for validation and used CALINSKI HARABASZ score method and also used Silhouette score. This study also uses the dendrogram for validation and analysis. The next step is to find the center points and error in the centroid, the number of clusters for each centroid, and the error of the centroid. Ultimately, all these methods are deployed based on the features of the crime data points to find the possible crime and predict how to deal with potential criminals.

Table II compares clusters and validation methods, and each method's time complexity helps summarize the whole study on one page. It was noticed that the Hierarchical clustering is good at some points. Still, density-based clustering also helps with batter clustering, and validation techniques help us find the optimal value of K to find the optimal number of clusters in the data for batter classification. After the compression of the clustering techniques, we can see that divisive is the most time-consuming clustering technique. In the figure, the shortest clustering technique is dendrogram and different methods. The progressive is DBSCAN, and the clustering is optimal. Time is also the validation method for DBSCAN, and we observe that the outliers are also removed in DBSCAN. Overall, methods are suitable, and most use a very short time for clustering, but it also depends on the dataset features.

The Table II provides a comprehensive overview of the different clustering techniques applied in an analytical study. It includes information on the number of clusters generated, the validation methods used, the number of optimally validated clusters, and the time complexity—expressed in seconds—for every clustering strategy. The first section of the table presents Agglomerative Clustering and Dendrogram-based approaches that use several linkage techniques such as Single, Complete, Average, Weighted, Centroid, Median, and Ward Linkage. These techniques employ a variety of validation

TABLE II. Comparison of Other Clustering Techniques on the Crime Dataset

Sr. No	Clustering Technique	No. of Clusters	Validation Cluster	Valid optimal No. of Clusters	Time Complexity in Seconds
1	Agglomerative Clustering	1	Elbow Method, Silhouette Score, CALINSKI H Score Elbow, Davies Bouldin Score	4,2,4	34.7
2	Dendrogram with Single Linkage	3	Elbow Method, Silhouette Score, CALINSKI H SCORE Elbow	4,2,4	34.7
3	Dendrogram with Complete Linkage	3	Elbow Method, Silhouette Score, CALINSKI H Score Elbow	4,2,4	33.8
4	Dendrogram with Average Linkage	3	Elbow Method, Silhouette Score, CALINSKI H Score Elbow	4,2,4	33.2
5	Dendrogram with Weighted Linkage	4	Elbow Method, Silhouette Score, CALINSKI H Score Elbow	4,2,4	33.1
6	Dendrogram with centroid Linkage	2	Elbow Method, Silhouette Score, CALINSKI H Score Elbow	4,2,4	47.7
7	Dendrogram with median Linkage	3	Elbow Method, Silhouette Score, CALINSKI H Score Elbow	4,2,4	34.1
8	Dendrogram with ward Linkage	2	Silhouette Score	3	34.8
9	Agglomerative Clustering with the scatter plot	2	Silhouette Score	2	0.3
10	Agglomerative Clustering with the scatter plot	3	Silhouette Score	2	0.2
11	Agglomerative Clustering with the scatter plot	4	Silhouette Score	4	0.2
12	Agglomerative Clustering with the scatter plot	5	Silhouette Score	2	0.3
13	Agglomerative Clustering with the scatter plot	6	Silhouette Score	2	0.3
14	DBSCAN with EPS 0.5	4	Silhouette Score	3	0.8
15	DBSCAN with EPS 2.0	1	Silhouette Score	3	0.8
16	OPTICS Clustering	4	Silhouette Score	3	0.8
17	DEBCLUE Clustering	4	Silhouette Score	2	0.4
18	Divisive Clustering	3	NA	NA	973.2

methodologies, including the Elbow Method, CALINSKI H Score, Davies Bouldin Score, and Silhouette Score. According to the results, the optimal number of clusters is between two and four, and the temporal complexity of these techniques ranges from 33 to 47 seconds. The second section, which focuses on Agglomerative Clustering using a scatter plot, employs the Silhouette Score as the validation metric. It demonstrates that the optimal number of clusters is between two and three, with temporal complexity as low as 0.2 to 0.3 seconds. In the third portion of the table, alternative clustering techniques such as DBSCAN, OPTICS, and DENCLUE are presented. The Silhouette Score is used as a validation tool for each method's evaluation. Between three and four clusters is the optimal number generated by these algorithms, and between 0.4 and 0.8 seconds is the least amount of temporal complexity.

Row 11 (Agglomerative Clustering with scatter plot initially set to 4 clusters) stood out for its consistency, maintaining four as the optimal number of clusters per the Silhouette Score and doing so with a meager time complexity of 0.2 seconds. Based on these comparisons and outcomes, row 11(Agglomerative Clustering with scatter plot) appears to offer a good balance between time complexity and consistency in determining the optimal number of clusters, making it the best candidate for identifying crime hotspots effectively.

V. Conclusion

The study's findings are firmly supported by the comprehensive analysis of results produced by using various clustering methods to the "Crime in Chicago" dataset. The study began with a comprehensive data-gathering phase, followed by meticulous data preprocessing, clustering, and validation procedures. K-means, DBSCAN, and Agglomerative clustering are just a few of the clustering approaches that were put to use to help find crime hotspots and foresee where crimes could occur. This study uses clustering activity to predict crime. This concept uses spatial and temporal grouping to find crime hotspots. These hotspots help crime forecasting models predict future crimes. Using actual crime data, we discover that the Agglomerative clustering technique predicts future crimes in high-crime areas better. The proposed strategy can help police predict crime and safeguard the public. Foreseeing criminal behaviour in academia is new. This field found universal criminal behaviour patterns. These tendencies may assist law enforcement in anticipating crimes. Crime forecasting is vital in predictive policing, hotspot analysis, and geographical profiling. Crime rates are linked to census data like average income and literacy. These traits may indicate criminality. This project seeks to determine if past crimes can predict future crimes. We anticipate spatiotemporal criminal activity using machine learning and 2-D Hotspot analysis. 2-D hotspot analysis uses clustering. Modern categorization, with and without hotspot analysis, tests the model. Hotspot analysis improves model performance.

Conflicts of Interest: "The authors declare that they have no conflicts of interest to report regarding the present study."

Funding: This research is supported by the National Natural Science Foundation of China under Grant 62172366 and Key Projects of the Ministry of Science and Technology of the People's Republic of China (2018AAA0102301). This research is also supported by the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R 343) Riyadh, Saudi Arabia. The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA for funding this research work through the project number "NBU-FFR-2024-1092-12". This research was funded by Taif University, Saudi Arabia, Project No. (TU-DSPP-2024-41).

Data Availability Statement: Publicly available data was used to carry out this research.

Acknowledgments

Muhammad Faisal Buland Iqbal and Tariq Hussain contributed equally to this work and are the first co-authors.

References

[1] M.-S. Baek, W. Park, J. Park, K.-H. Jang, and Y.-T. Lee, "Smart policing technique with crime type and risk score prediction based on machine learning for early awareness of risk situation," *IEEE Access,* vol. 9, pp. 131906-131915, 2021.

[2] M. Khan, A. Ali, and Y. Alharbi, "Predicting and preventing crime: a crime prediction model using san francisco crime data by classification techniques," *Complexity,* vol. 2022, p. 4830411, 2022.

[3] X. Wang, M. S. Gerber, and D. E. Brown, "Automatic crime prediction using events extracted from twitter posts," in *International conference on social computing, behavioral-cultural modeling, and prediction*, 2012, pp. 231-238.

[4] F. K. Bappee, "Prediction of Crime Occurrences: A data-driven approach for single domain and cross-domain learning," 2020. http://hdl.handle.net/10222/80092

[5] A. Almaw and K. Kadam, "Survey paper on crime prediction using ensemble approach," *International Journal of Pure and Applied Mathematics,* vol. 118, pp. 133-139, 2018.

[6] M. David, E. S. Mbabazi, J. Nakatumba-Nabende, and G. Marvin, "Crime Forecasting using Interpretable Regression Techniques," in *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)*, 2023, pp. 1405-1411.

[7] U. M. Butt, S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir, and H. H. R. Sherazi, "Spatio-temporal crime hotspot detection and prediction: a systematic literature review," *IEEE access,* vol. 8, pp. 166553-166574, 2020.

[8] G. Hajela, M. Chawla, and A. Rasool, "A clustering based hotspot identification approach for crime prediction," *Procedia Computer Science,* vol. 167, pp. 1462-1470, 2020.

[9] G. Hajela, M. Chawla, and A. Rasool, "Crime hotspot prediction based on dynamic spatial analysis," *ETRI Journal,* vol. 43, pp. 1058-1080, 2021.

[10] E. G. Badreddine, H. Larbi, A. Houda, and R. Mohammed, "Spatio-Temporal Crime Forecasting: Approaches, Datasets, and Comparative Study," in *International Conference on Advanced Intelligent Systems for Sustainable Development*, 2022, pp. 231-251.

[11] S. S. May, O. E. Isafiade, and O. O. Ajayi, "Hybridizing extremely randomized trees with bootstrap aggregation for crime prediction," in *Proceedings of the 2021 4th International Conference on Artificial Intelligence and Pattern Recognition*, 2021, pp. 536-541.

[12] K. Kianmehr and R. Alhajj, "Effectiveness of support vector machine for crime hot-spots prediction," *Applied Artificial Intelligence,* vol. 22, pp. 433-458, 2008.

[13] A. Jan and G. M. Khan, "Real world anomalous scene detection and classification using multilayer deep neural networks," , vol. 8, no. 2, pp. 158–167, 2023.

[14] P. Pawale, S. Bagal, S. Ajabe, and K. Shikalagar, "GEO STATISTICAL APPROCH FOR CRIME HOTSPOT DETECTION AND PREDICTION," *International Research Journal of Engineering and Technology,* Vol: 04 Issue: 05, pp, 2703-2706, 2017.

[15] S. Das and M. R. Choudhury, "A geo-statistical approach for crime hot spot prediction," *International Journal of Criminology and Sociological Theory*, vol. 9, no. 1, pp. 1–11, 2016.

[16] S. B. Mehta and R. D. Doshi, "Automatic clustering crime region prediction model using statistical method in data mining," *International Journal of Engineering Research And,* vol. 9, pp. 697-703, 2020.

[17] M. Boukabous and M. Azizi, "Image and video-based crime prediction using object detection and deep learning," *Bulletin of Electrical Engineering and Informatics,* vol. 12, pp. 1630-1638, 2023.

[18] M. Kamruzzaman, "New opportunities, challenges, and applications of edge-AI for connected healthcare in smart cities," in *2021 IEEE Globecom Workshops (GC Wkshps)*, 2021, pp. 1-6. doi: 10.1109/

GCWkshps52748.2021.9682055.
[19] A. McMorrow, "Spatiotemporal forecasts of London's crime hotspots," Dublin, National College of Ireland, 2022.
[20] Y. Hu, F. Wang, C. Guin, and H. Zhu, "A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation," *Applied geography,* vol. 99, pp. 89-97, 2018.
[21] N. Abdulrahman and W. Abedalkhader, "KNN classifier and Naive Bayse classifier for crime prediction in San Francisco context," *International Journal of Database Management Systems,* vol. 9, pp. 1-9, 2017.
[22] A. A. Hussein, "Identifying Crime Hotspot: Evaluating the suitability of Supervised and Unsupervised Machine learning," Master's thesis, University of Cincinnati, 2021.
[23] Y. Zhuang, M. Almeida, M. Morabito, and W. Ding, "Crime hot spot forecasting: A recurrent model with spatial and temporal information," in *2017 IEEE International Conference on Big Knowledge (ICBK),* 2017, pp. 143-150.
[24] S. S. Kshatri, D. Bhonsle, R. Verma, A. G. Pillai, and V. Moyal, "Crime Detection Approach Using Big Data Analytics and Machine Learning," *NeuroQuantology,* vol. 20, p. 1480, 2022.
[25] Y. Lee and S. O, "Flag and boost theories for hot spot forecasting: An application of NIJ's Real-Time Crime forecasting algorithm using Colorado Springs crime data," *International journal of police science & management,* vol. 22, pp. 4-15, 2020.
[26] A. F. Mohammed and W. R. Baiee, "Analysis of criminal spatial events in GIS for predicting hotspots," in *IOP conference series: materials science and engineering,* 2020, p. 032071.
[27] M. N. Khan, H. U. Rahman, T. Hussain, B. Yang, and S. M. Qaisar, "Enabling Trust in Automotive IoT: Lightweight Mutual Authentication Scheme for Electronic Connected Devices in Internet of Things," *IEEE Transactions on Consumer Electronics,* 2024.
[28] M. Qasim Gandapur and E. Verdú, "ConvGRU-CNN: Spatiotemporal deep learning for real-world anomaly detection in video surveillance system," vol. 8, nº 4, 2023. https://doi.org/10.9781/ijimai.2023.05.006
[29] Z. He, R. Lai, Z. Wang, H. Liu, and M. Deng, "Comparative study of approaches for detecting crime hotspots with considering concentration and shape characteristics," *International journal of environmental research and public health,* vol. 19, p. 14350, 2022.
[30] K. Jenga, C. Catal, and G. Kar, "Machine learning in crime prediction," *Journal of Ambient Intelligence and Humanized Computing,* vol. 14, pp. 2887-2913, 2023.
[31] A. D. d. Araújo Júnior, "Predspot: predicting crime hotspots with machine learning," Brasil, 2019. https://repositorio.ufrn.br/handle/123456789/29155

Muhammad Faisal Buland Iqbal

Muhammad Faisal Buland Iqbal has completed his BS in Computer Science at AIO University, Islamabad, Pakistan. He received his MS in Software Engineering from Xiangtan University, China 2017. He has been a Key Laboratory of Intelligent Computing & Information Processing, Ministry of Education, Xiangtan University member since 2014. He won the China Scholarship Council award in 2013 and the Hunan Provincial International Student Award in 2015. Currently, he is researching Cryo-EM micrograph Denoising, 3D Construction, and data analysis. He has worked in the industrial sector for four years and has done many international projects. His research interests are Machine Learning, Deep Learning, Data Mining, NLP (Natural Language Processing), Network Data Analysis, Computer Vision, IoT, and Bioinformatics.

Aman Ullah

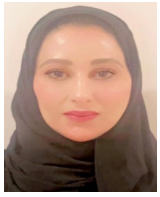
Aman Ullah received his BS in Software Engineering from Virtual University, Lahore, Pakistan, in 2023. He started his academic career as a researcher and freelancer in 2022. Since then, he has been working in academia and an active researcher. His research fields are Application Development, Machine Learning, Artificial intelligence, and IoT.

Ahmed Alhomoud

Dr. Ahmed Alhomoud is Assistant Professor in the Department of Computer Science, Northern Border University, Kingdom of Saudi Arabia. He received his Ph.D. in computer science from the University of Southampton, United Kingdom. His research interests include but are not limited to digital forensics, cyber security, Internet of Things and blockchain.).

Tariq Hussain

Tariq Hussain received his BS and MS Degrees in Information Technology from the University of Malakand, Pakistan (2015) and the Institute of Computer Sciences and Information Technology – the University of Agriculture Peshawar, Pakistan (2019), respectively. He has published many research papers in the area of Computer Networks. He is a doctoral candidate at the School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, China. His research interests are the Internet of Things, Big Data and statistics, 3D Point Cloud QoE Model, and Artificial Intelligence.

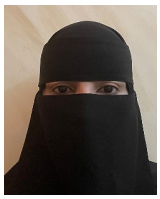


Razaz Waheeb Attar

Dr. Razaz Waheeb Attar, Ph.D., is an Associate Professor in the Department of Management at the School of Business, Princess Nourah Bint Abdul Rahman University, located in Riyadh, Saudi Arabia. She holds a Ph.D. in Business Administration from Dublin City University, an esteemed institution on the Northside of Dublin, Ireland. She has also earned master's degrees in electronic commerce from Dalhousie University in Nova Scotia, Canada, and in Business Research from Kingston University in Kingston upon Thames, London, England. She has further pursued studies at King Abdulaziz University in Jeddah, Saudi Arabia. Her areas of research interest encompass a wide range of topics, including social innovation, higher education, strategic management, contemporary issues in marketing and management, digital strategy, social commerce, digital ethics, sustainability, AI risks, and e-commerce. She scholarly contributions have been recognized through the publication of her research and articles in reputable journals. She has made significant strides in S-commerce design and development; her expertise and passion are evident in her work.

Jianquan Ouyang

Prof. Jianquan Ouyang is a professor, doctoral supervisor, and visiting scholar at the University of Georgia, U.S.A. In 2007, he was selected as the training object of young backbone teachers in general colleges and universities in Hunan Province. He is a senior member of the China Computer Federation (CCF), a liaison of Xiangtan University, a correspondent member of the CCF Youth Working Committee, and a member of CCF Education. He has won many awards. His research interests include Multi-modal Mashup, Multimedia Analysis and Retrieval, Multi-valued Logic, and Trustworthy Computing. His current research focuses on deep learning, MRC image processing, and cybersecurity.



Mrim Alnfiai

Dr. Mrim Alnfiai is an Associate Professor of Information Technology at the Taif University in Saudi Arabia. Her research interests are in assistive technology, Human Computer Interaction, Accessibility, Usable Security, AI and Machine Learning. Mrim publishes several papers at ISI journals and assistive technology, HCI and accessibility conferences including ASSETS, ANT, FNC, CIST, JAIHC, and ICCA. Currently, her research focuses on designing accessible tools for visually impaired people including people with no or low vision. She has conducted several studies and experiences to understand visually impaired abilities and behaviors and design accessible systems that help them interact easily with technology. She has also published several papers related to accessibility and authentication mechanisms for visually impaired users. She has also published papers related to using NFC technology and machine learning to enhance the healthcare system.

Wesam Atef Hatamleh

Dr. Wesam Atef Hatamleh is a distinguished academic author and researcher renowned for his contributions to healthcare engineering, wireless communications, and computational intelligence. His recent publications in 2022 reflect a profound engagement with cutting-edge technologies and methodologies, particularly in machine learning, deep learning, and human-computer interaction. Dr. Hatamleh's work has been prominently featured in high-impact journals such as the Journal of Healthcare Engineering, Wireless Communications and Mobile Computing, and Computational Intelligence and Neuroscience. His research encompasses a wide array of topics, including intracranial tumor detection, multimodal sarcasm detection, speaker emotion recognition, and gender recognition using advanced neural networks like ResNet50 and convolutional neural networks (CNNs).