An optimized machine learning and big data approach to crime detection.

PALANIVINAYAGAM, A., GOPAL, S.S., BHATTACHARYA, S., ANUMBE, N., IBEKE, E. and BIAMBA, C.

2021

Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 5291528, 10 pages
https://doi.org/10.1155/2021/5291528

Research Article

An Optimized Machine Learning and Big Data Approach to Crime Detection

Ashokkumar Palanivinayagam,[1] **Siva Shankar Gopal**,[1] **Sweta Bhattacharya**,[2] **Noble Anumbe**,[3] **Ebuka Ibeke**,[4] **and Cresantus Biamba**[5]

[1]*Sri Ramachandra Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Tamil Nadu, India*
[2]*School of Information Technology and Engineering, VIT, Tamil Nadu, India*
[3]*Department of Mechanical Engineering, University of South Carolina, Columbia, SC, USA*
[4]*School of Creative & Cultural Business, Robert Gordon University, Aberdeen, UK*
[5]*Faculty of Education and Business Studies, University of Gavle, Sweden*

Correspondence should be addressed to Cresantus Biamba; cresantus.biamba@hig.se

Received 23 June 2021; Accepted 10 October 2021; Published 13 November 2021

Academic Editor: Vishal Sharma

Crime detection is one of the most important research applications in machine learning. Identifying and reducing crime rates is crucial to developing a healthy society. Big Data techniques are applied to collect and analyse data: determine the required features and prime attributes that cause the emergence of crime hotspots. The traditional crime detection and machine learning-based algorithms lack the ability to generate key prime attributes from the crime dataset, hence most often fail to predict crime patterns successfully. This paper is aimed at extracting the prime attributes such as time zones, crime probability, and crime hotspots and performing vulnerability analysis to increase the accuracy of the subject machine learning algorithm. We implemented our proposed methodology using two standard datasets. Results show that the proposed feature generation method increased the performance of machine learning models. The highest accuracy of 97.5% was obtained when the proposed methodology was applied to the Naïve Bayes algorithm while analysing the San Francisco dataset.

1. Introduction

In the last few decades, there has been an exceptional growth in urban population which has led to the demand for a secured, hospitable, and sustainable society. With the ever-expanding growth of city, engulfing suburbs and rural spaces, the management of urbanization remains a major challenge for administrative authorities. Cities are getting overpopulated, compelling governments to undertake smart city initiatives that would help achieve better management of infrastructure and overcome the major challenges of security, sustainability, and development. Although smart city initiatives have gained immense momentum with promises to enhance quality of life, it does have its own challenging aspects as well. One of the major challenges in smart city life is public safety. Various studies have been conducted to help understand crime patterns and its relationship to the social economic development of particular regions, the human characteristics, their level of education, and family bonding [1].

Crime investigating organizations have identified various types of crimes. The four main categories include killing, molestation, looting, and intensive attacks. Killing or murder refers to the willful assassination of a person by another. Molestation means the sexual abuse of a woman, man, or child against their wish. This crime is as heinous as rape, having significant consequences. Looting refers to the act of stealing goods from a human domain, using excessive physical force or violence. Finally, intensive attacks refer to illegal confrontation by one person against another to achieve something or to

simply harm the individual [2]. Crime detection is a necessity in urban life, and machine learning is a popular crime detection and prevention technique. Several organizations across the globe have been experimenting with these techniques.

It has been observed that crimes are often predictable, and it just requires the processing of high volumes of data that would reveal interesting patterns suitable for law enforcement. In many of the instances, crimes conducted often remain unreported due to external pressures from all verticals of the society. Intelligent systems can promptly detect crimes and help eradicate such manipulative activities by bypassing individuals and automatically informing relevant authorities. As an example, the research by Borges et al. [1] discussed the case study of San-Francisco, USA, and Natal in Brazil where criminal activities were prevalent. The various attributes of urbanization in these two cities were analysed, and then, machine learning models were implemented to detect criminal activity hotspots. As per [2], they created a regression model to predict crime rates in various Indian states. Supervised and unsupervised learning techniques were also deployed to achieve enhanced accuracy in crime prediction. In [3], fuzzy C-means algorithm was used for the clustering of crime data for various cognizable crimes, namely, kidnapping, murder, theft, robbery, and crimes against women. Similarly, K nearest neighbour methods have been deployed for the observation of crime rates which have helped to understand crime types and time/place of occurrence.

Considering the various studies conducted, it is observed that most of the existing works emphasize the use of crime history and population density for the crime prediction. The present work presents four attribute generation methods for the detection of crimes. The dataset holds various crime locations in an area where K-means clustering is applied, yielding crime hotspots. Then, a crime ratio matrix is constructed leading to the prediction of crime probability when subjected to a machine learning model. As part of the proposed methodology, crime monitoring is performed with the help of the following methods:

(i) Crime transition probability computes the connection of one crime to another

(ii) Vulnerability of an area indicates how safe an area is

Many existing works use artificial intelligence and machine learning to extract crime patterns and to detect and prevent crime incidents. Most of the existing works have few limitations which include incompetence of finding links between different crime incidents and vulnerability analysis. In this paper, we propose four unique stages of crime detection which uses the combination of locations, vulnerability, correlation, and temporal patterns.

The unique contributions of the proposed method are highlighted below:

(i) Ability to analyse the relationships between time zones, namely, morning, evening, and night for each type of crime

(ii) Prediction of crime probability for the following day considering the present-day crime history

(iii) Generation of crime hotspots in the form of geolocations indicating occurrence of a greater number of crimes

(iv) Performing vulnerability analysis to identify locations more prone to criminal activities in the future

This paper contains five sections: Section 2 discusses previous studies. Section 3 describes the four-feature generation process used in the proposed work. The results of our work are discussed in Section 4. Finally, Section 5 contains the conclusion and future work.

2. Related Works

Various studies have been conducted that are relevant to crime detection, analysis of the various factors that contribute significantly towards crime occurrence and its impact on the socio-economic status of various regions. Machine learning approaches have been a predominant and popular area of research interest in the crime detection domain. This section summarizes some of the interesting studies conducted in crime detection, analysis, and prediction. The overview will help highlight research gaps or limitations in this field.

A research proposed by [4] implemented deep learning approaches on CCTV camera images to detect crimes, eliminating traditional (manual) monitoring systems that rely on human supervision. In the traditional system, the CCTV cameras are installed at various positions in the public and private surroundings which capture videos and images with the prime objective of monitoring and preventing incidences of crime. However, the detection of crimes does not happen automatically as it requires human supervision and constant monitoring of CCTV screens. This physical monitoring system is often prone to errors due to the chance of missing important incidents since effectively monitoring multiple screens at the same time is often difficult. To overcome the challenge, [4] developed a pretrained deep learning model VGGNet19 that detected criminal events in real time and generated an alert for the human supervisor to ensure immediate action is taken. The results were evaluated against GoogleNet and InceptionV3, with the VGGNet19 model yielding higher training accuracy. However, the model detects criminal intentions but does not provide any insight on crime hotspots nor does it highlight the probabilities of crime occurrence.

One more research [5] proposes a visual surveillance system that would detect hostile intent and behaviour inside the elevators. The surveillance camera that captures images of the small, confined elevator space based on the illumination of the opening and closing of the elevator doors was used for this study. The implementation involved a three-layered approach for the detection of violent events. The low-level feature-segmented foreground blobs from the background and their motion velocities were captured using an optical flow method. In the second or midlevel feature, the velocity and directions were computed to analyse motions of the

images captured. Sequences of image frames having more than one person in the elevator were analysed, and whenever an average velocity magnitude exceeded a threshold value, a violent event occurrence was assumed to have been detected.

The methodology proposed by [6] is aimed at predicting crime without human intervention using computer vision and machine learning approaches. The paper implements rectified linear unit (ReLU) and convolutional neural networks (CNN) for the detection of weapons such as knives or guns from a particular image. This helped to validate the occurrence of a crime and identify the location of occurrence as well. The accuracy of the results seemed quite promising, which achieved almost 92% accuracy for a testing dataset.

Another interesting research [7] discusses the excessive surge in document forging incidents using powerful photo editing software used as a tool for creating fake documents. Such fake documents are scanned and forgotten in minutes with the help of automated editing tools used exclusively for the said purpose. The study involves the use of a GUI which is designed to detect if an image is manipulated or not. The GUI helps to load and preprocess the image, enhancing its global contrast. The image is then partitioned into three segments using the K-means clustering approach. The segment containing most of the information is further analysed, extracting its features. These are compared with the scanned images in the database to identify the occurrence of tampering. Support vector machine (SVM) and ANN were implemented, but SVM yielded better accuracy and thus was considered the most suitable.

A ML model [8] proposed a fraud detection system using a hybrid machine learning approach emphasizing on electronic transactions. It has been observed that most economic frauds involve business transactions relating to credit cards. The paper uses feature engineering approach on the dataset and then SVM and random forest implementation as a hybrid technique to detect fraudulent transactions.

One more research work by [9] developed a machine learning-based approach for the detection of spam images. In the present day and age, email is one of the vital modes of communication almost among all stakeholders in the society. Email not only acts as digital letters but also enable the attachment of documents, pictures, videos, and music to be sent to recipients. There are certain miscreants who send unsolicited emails to users to weaken the internet traffic. The spammers also sent such emails to users attracting them to buy products which are prohibited. The study involved using chi-square test for feature engineering and sequential minimal optimization (SMO) algorithm. Post feature selection method, multilayer perceptron (MLP) algorithm is used for the detection of spams. Both SMO and MLP yielded an F-score of 98.5% and 98.4%, respectively.

The work done by [10] developed a machine learning model that would help to predict potential crimes in a geographic location, analysing the existing crime and repeating incident occurrence datasets. The paper used the Chicago Police Department CLEAR dataset and selected 9 features from the dataset for further analysis. Finally, Naïve Bayes- and decision tree-based approaches were used to predict potential crimes. This was intended to help create contingency plans and keep the society safe, promoting hospitable and secured living. The results highlighted the superiority of the decision tree-based approach considering 7, 8, and 9 features for the matrices: correctly classified instances (CCI), accuracy (AC), ROC, precision, and recall, respectively.

The study in [11] focused on comparing two images by identifying the query image from the source image, which would help in the recognition of a particular person or object in the image. The frames that matched were generated as an output after implementation of the scale invariant feature transformation (SIFT) method. SIFT was used to extract features that were invariant to image scaling, rotation, presence of noise, or all changes in the image lighting. Once the feature points in an image were identified, they were compared with the feature points in the frame implementing homographic estimation. The Euclidian distance formula was used for the comparison.

The work by [12] targeted the occurrences of road transport crimes and identified methods to reduce them. Road transport is often used by criminals for escaping after conducting heinous crimes. Moreover, a lot of crimes remain unregistered and unresolved due to lack of evidence on the roads. To eliminate such occurrences, a machine learning algorithm was deployed in the study using text and facial recognition techniques. The system extracts characters from the vehicle number plates using a text recognition mechanism. On the other hand, the facial recognition algorithm helps in the identification of the face of the suspects. The extracted feature is mapped to the relevant features of the images saved in the database, and in case of mismatch, an alert is generated. In the same way, the facial images are compared with criminal face images available in the database, and in case of anomaly, an alert is generated. KNN and SVM in association with face detection classifier were used to achieve the proposed objective [13].

In [14], news is analysed using machine learning algorithms and provides a report on the classified crime news. The traditional system involves reading the complete news and manually analysing the same which is prone to errors. Moreover, the approach is quite time consuming. To overcome this challenge, a machine learning-based classification approach is implemented involving the use of three classifiers. The result segregates crime-related data and noncrime-related data. The website or newspaper contents are fed into the system, a crawling program is implemented written in Python, and the data is finally stored in a temporary database. The result generated display crime and noncrime data presented in a tabular format to the user. Table 1 shows the summary of related works performed.

Another research work [15] concentrates on crime hotspot detection. They have used data from 2 million crime data between 2006 and 2018 to train GAN model. Their research work proposes a new city plan based on the crime distribution. The simulated new city plan seems to have much lower crime rate than the original city.

The crime data is imbalanced most of the time. [16] uses data argumentation and loss function to develop samples

Table 1: Summary of related works.

References	Dataset	Methods used	Evaluation metrics	Limitations
Navalgund and K. (2018) [4]	YouTube and Google	VGGNet -19	Accuracy, recall, F1-score and support	Detection of crime hotspots and probability of occurrences not included.
Younghyun Lee et al. (2011) [5]	Real-time elevator data collected using surveillance camera of 320 ∗ 240 pixels	Violent frame detector, motion vector extraction, and foreground segmentation	Detection rate, no. of people in the elevator, false-positive rates (FPR)	Includes only detection but not prediction or probabilities of occurrence results
Nakib et al. (2018) [6]	Real-time data	Softmax regression model, CNN	Accuracy	The size of the dataset was relatively small. The model was not evaluated against the other classical models.
Ranjan et al. (2018) [7]	Image collected from various internet sources and then morphed to test the methods	SVM and ANN	Accuracy, sensitivity and specificity	Comparison of the results with other traditional approaches were not included. Availability of larger dataset also is a challenge
Vynokurova et al. (2020) [8]	Real-time dataset	SVM and random forest-based hybrid approach	Accuracy	Comparison of the results with other traditional approaches were not included. Availability of larger dataset also is a challenge

and improve the minority class. They have used neural network to enhance the crime detection problem.

3. Preparing the Model

In this section, we present the working of the proposed model and the four attribute generation methods such as fraction of day, crime growth factor, distance from crime hotspot, and vulnerability analysis. The overall flow of the proposed method is shown in Figure 1.

3.1. Fraction of the Day. Crimes are more likely to occur at certain times of the day, for example, more crimes occur between 6 p.m. and 12 a.m. (next day) than between 6 a.m. and 12 p.m. Hence, to increase the prediction success rate, it will be better to consider a fraction of the day instead of the day as a whole [17, 18].

Consider 100 crimes that happened on day X. Since most of the crimes are more likely to occur at night, in the proposed model, we consider the impact of different fractions of the day instead of the whole day. In this case, we divide a single day into four fractions such as

(i) Fraction 1: between 00:00 AM and 06:00 AM

(ii) Fraction 2: between 06:01 AM and 12:00 PM

(iii) Fraction 3: between 12:01 PM and 06:00 PM

(iv) Fraction 4: between 06:01 PM and 11:59 PM

For each crime, the number of crime events is noted and stored in crime counter (CC) as per Figure 2.

$C_1, C_2, , \cdots C_n$ are different crimes and F_1, F_2, F_3, and F_4 are the four fractions, respectively. $N_{ci,Fj}$ represents the number of crimes i that occurred at fraction j. The time fractions can be made dynamic; however, dividing a day into four fractions makes the segregation of crimes simpler and more meaningful.

3.2. Crime Growth Vector. The most important aspect of crime forecasting system is detecting the probability of crime each day [19, 20]. The probability of crime i can be found by calculating the percentage of the number of crime i events in the total number of all crimes. The crime vector CV stores the probability of all crimes. Equation (1) shows the structure of the CV. Each value in the vector is calculated by Equation (2)

$$\text{Crime Vector}(\text{CV}) = (P_{c1}, P_{c2}, \cdots, P_c), \tag{1}$$

$$P_{C_i} = \frac{\text{Number of crimes } i}{\text{Total number of crimes}}. \tag{2}$$

Transition probability matrix (TPM) is one of the methods which can help to forecast the probabilities of future days. TPM needs a vector (to denote the initial probability) and a matrix (to represent the Markov chain). In this context, we use the crime vector as the initial probability matrix. The crime growth factor can be used as Markov chains. A crime growth factor between two crimes A and B is how much likely a crime B is to happen on day $d + 1$ when crime A has happened on day d. Equations (3) and (4) can be used to calculate the likelihood of two crimes happening on day d and day $d - 1$. The values are normalized so that

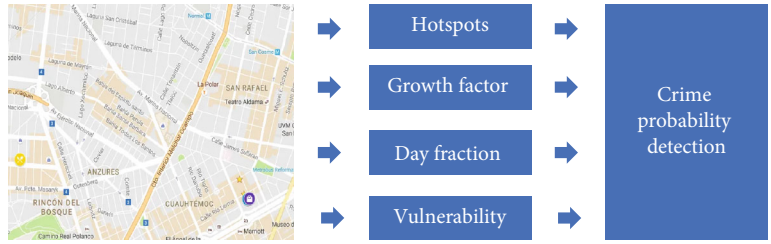

Figure 1: The working of the proposed model.

$$\mathrm{CC} = \begin{bmatrix} N_{C_1F_1} & N_{C_1F_2} & \cdots & N_{C_1F_m} \\ N_{C_2F_1} & N_{C_2F_2} & \cdots & N_{C_2F_m} \\ \cdots & \cdots & \ddots & \cdots \\ N_{C_nF_1} & N_{C_nF_2} & \cdots & N_{C_nF_m} \end{bmatrix}$$

Figure 2: The crime counter.

the factors are turned into probability values. The crime growth factor is calculated for each day fraction separately and finally merged into a single matrix as per Equation (5).

$$GF_{i,j}^{d} = \frac{g_{ij}^{d}}{\sum_{i=1}^{n} g_{ik}^{d}}, \tag{3}$$

$$g_{i,j}^{d} = \frac{\text{Number of crime } j \text{ on day } d}{\text{Number of crimes } i \text{ on day } d * 1}, \tag{4}$$

$$\text{Final Value}\left(FV_{i,j}\right) = \sqrt[n]{\prod_{k=1}^{n} g_{i,j}^{k}}, \tag{5}$$

Next Day Crime Probability Vector

$$=: \left[P_{C_1}, P_{C_2}, \cdots, P_{C_n}\right] \begin{bmatrix} FV_{1,1} & FV_{1,2} & \cdots & FV_{1,n} \\ FV_{2,1} & FV_{2,2} & \cdots & FV_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ FV_{n,1} & FV_{n,2} & \cdots & FV_{n,n} \end{bmatrix}. \tag{6}$$

Using this TPM, the next day probability can be easily calculated by multiplying the CV and the final value matrix. The calculation is mentioned in Equation (6).

3.3. Determining Hotspots. Hotspot identification is an important factor to consider for crime detection. A hotspot represents highly frequent crime locations; hence, accurate prediction of the crime hotspots increases the accuracy of the crime detection process. Hotspot represents a spatial relationship between the occurrences of crime.

The calculation of hotspots is as follows: first, the coordinates of all crime reporting are grouped based on the type of crime. For example, the coordinates of "VEHICLE–STOLEN" are grouped into a separate list; second, the X and Y locations are clustered using K-means clustering. Finally, the distance from the nearest cluster is found. The working of hotspot identification is presented in Algorithm 1. The algorithm converges when there are no more additional changes in the clusters. Figure 3 illustrates the working of hotpot identification.

3.4. Vulnerability Analysis. In this subsection, we present the vulnerability analysis, which can detect the possible areas where there are more chances for a crime to occur. Suppose we consider an area X, a crime Y has happened, that means the area is open to attacks or there are fewer or insufficient security measures. Hence, the area surrounding X is more likely to become vulnerable to Y. We use kNN to analyse the vulnerability. Let us say, there is a vehicle theft at a place X, that means X has less security for monitoring the crime Y ; hence, the same area or the surrounding areas are too likely to become a vulnerable point. Link-based algorithms such as [21] will be helpful in creating a graph; the latter kNN algorithm can easily predict the crime spots. Figure 4 shows a visualization of crime in San Francisco; the visualization shows which areas are vulnerable and lack security monitoring.

We have considered 5 as the value for k and the kNN used in this model produces 86.61% accuracy.

The proposed crime detection model works as follows: Firstly, a day is fragmented into four sections because it enhances the identification of temporal patterns of crimes. Few crimes such as robbery and chain snatching mostly occur at night, whereas other crimes such as hit and run and kidnapping occur during the day. Segregation of the day into various time quantum can help the prediction process. Secondly, the relationship between various crimes is established, i.e., how different crimes are linked to each other. The proposed method uses the crime correlation and growth rate to increase the prediction of the crime events. Thirdly, the hotspots of crime are identified. A hotspot represents a small geographical location where many crime incidents have been reported. Finally, vulnerability identification allows the proposed method to recommend an area where crime events are likely to occur in the future. By using both temporal and spatial inputs, the proposed model develops an increased ability to correctly predict crime events.

4. Results and Discussion

The proposed algorithm is to predict the probability of a given crime for a given area. We performed a comparison of our results with other machine learning algorithms such

procedure HOTSPOT Generation
  K←The number of hotspot
  Output ←**S**{}, the crime location in each hotspot
begin:
  Initialize the midpoints $m^{(1)}$={**R**andom **K** points}
  for*i*=1 to *k***do**
    Add respective midpoint to the hotspot. $C_i^{(1)} = m_j^{(1)}$
  iter=1
  while True **do**
    for each point P**do**
      min=0
      Cluster=None
      for*i* = 1 to *k***do**
        dist=$\|p - m_i^{iter}\|^2$
        if min<dist **then**
          min=dist
          cluster=*i*
      $S_{\text{cluster}} = \mathbf{S}_{cluster}^{iter} \cup \{\boldsymbol{P}\}$
      iter=iter+1
      for*i* = 1 to *k***do**
        $m_i^{iter} = 1/|S_i^{iter-1}| \sum x \in S_i^{iter=1x}$
    break when past 2 **S** values are same
  return **S**

Algorithm 1: Identification of hotspots.

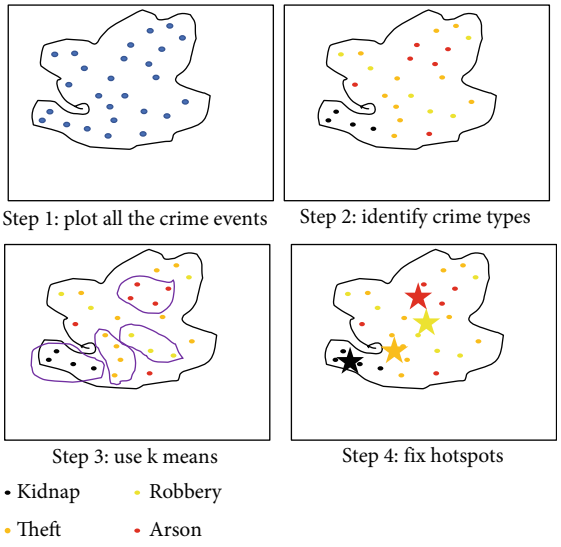

Figure 3: Illustration of hotspot identification.

as Naive Bayes, kNN, random forest, and support vector machines.

4.1. Dataset Description. We have used four attributes present in the Los Angeles dataset and six attributes in San Francisco dataset. The attributes used for the dataset are shown in Table 2. These attributes are used to train the existing machine learning algorithms.

In addition to the attributes present in the dataset, we have added four new attributes as discussed in Section 3 and fed into the proposed method.

4.2. Evaluation Metrics. Our evaluation metrics include accuracy, precision, and recall. The outputs of all classifiers are binary; hence, we can define the terms true positive (TP), true negative (TN), false positive (FP), and false negative (FN) as follows.

(i) TP: when a crime event is predicted as a crime event

(ii) TN: when a noncrime event is predicted as a noncrime event

(iii) FP: when a noncrime event is predicted as a crime event

(iv) FN: when a crime event is predicted as a noncrime event

We used three parameters (i.e., accuracy, precision, and recall) to test and evaluate the performance of the proposed model using existing machine learning algorithms.

Accuracy: accuracy is defined as the quality of correctness, and it is calculated by using the formula given by the equation

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{7}$$

Precision: precision explains how many positives out of the total positives predicted are. Precision is calculated based on

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{8}$$

Figure 4: Crime incidents and the vulnerable areas (data taken from San-Francisco crime dataset: Mohan [22, 23]).

Table 2: Dataset description.

Dataset	Attributes used	Attributes generated
Crime in Los Angeles	(1) Crime code (2) Date occurred (3) Time occurred (4) Location	(1) Predicted probability (2) X (3) Y (4) Day of week (5) Fraction of day
San Francisco Crime Dataset	(1) Category (2) Day of week (3) Date (4) Time (5) X (6) Y	(1) Predicted probability (2) Fraction of day

Table 3: Performance evaluation: San Francisco Dataset.

Classifier	15 days average			7 days average			2 days average		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
NB	95	94	95.91	92.5	93.68	90.8	90.5	89.98	90.81
NB (with proposed features)	97.5	97.97	97	94.5	91.91	96.8	92.5	91.91	92.85
RF	93.5	94.05	93.13	89.5	87.87	90.6	90	87.87	91.57
RF (with proposed features)	97	96.03	97.97	93	88.11	97.8	91.5	88.11	94.68
kNN	95	93	96.87	94	92.85	94.79	90	89.21	91
kNN (with proposed features)	95.5	95.95	95	94	92.92	94.84	91.5	92.92	90.19
SVM	96.5	97	96.03	91.5	91.57	90.62	90	91.57	87.87
SVM (with proposed features)	97	98	96.07	92.5	97.8	87.25	91.5	95.69	87.25

Recall: the recall is a measure to calculate how many actual positives from all predicted positives found by the classifier. The recall is calculated by

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{9}$$

4.3. Day History Analysis. We considered three different values for the construction of Final Value FV_{ij}, i.e., 15, 7, and 2. The performance evaluations for both datasets are shown in Tables 3 and 4, respectively. We found that classifier performed better given more historical data. We also found that Naive Bayes resulted in the best performance when the number of days was 15 (i.e., 15-day average) compared to other classifiers.

Crime predictions based on patterns were performed by the classifiers. Additionally, we input four new attributes as mentioned in Section 3 into the classifiers. This allowed

Table 4: Performance evaluation: Los Angeles Dataset.

Classifier	15 days average			7 days average			2 days average		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
NB	91	93.75	88.24	90	91.67	88	89	89.58	87.76
NB (with proposed features)	93	92.93	92.93	92.5	91.92	92.86	90.5	88.35	92.86
RF	90.5	88.24	92.78	89.5	88.24	90.91	88	87.88	87.88
RF (with proposed features)	94.5	92.08	96.88	93.5	90.29	96.88	92.5	90.1	94.79
kNN	88	83.33	93.75	88.5	85.05	92.86	85.5	81.98	91
kNN (with proposed features)	92	92.86	91	91.5	92.78	90	90.5	92.78	88.24
SVM	87	85.71	87.5	88.5	88.42	87.5	89.5	90.53	87.76
SVM (with proposed features)	92.5	97.8	87.25	91	97.8	84.76	89.5	95.7	83.96

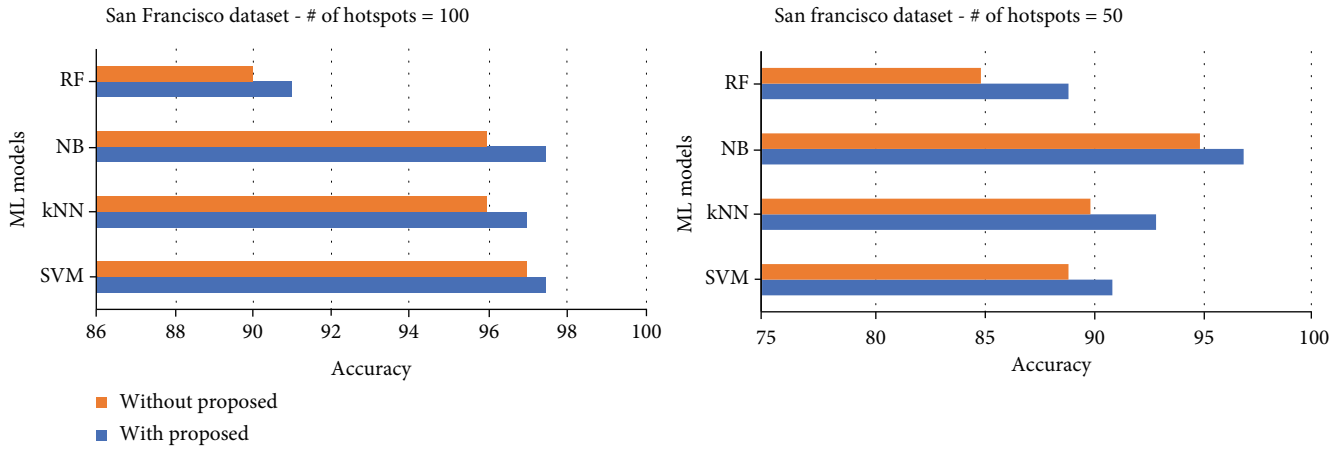

Figure 5: Accuracy of San Francisco Dataset, when # of hotspots is 100 and 50.

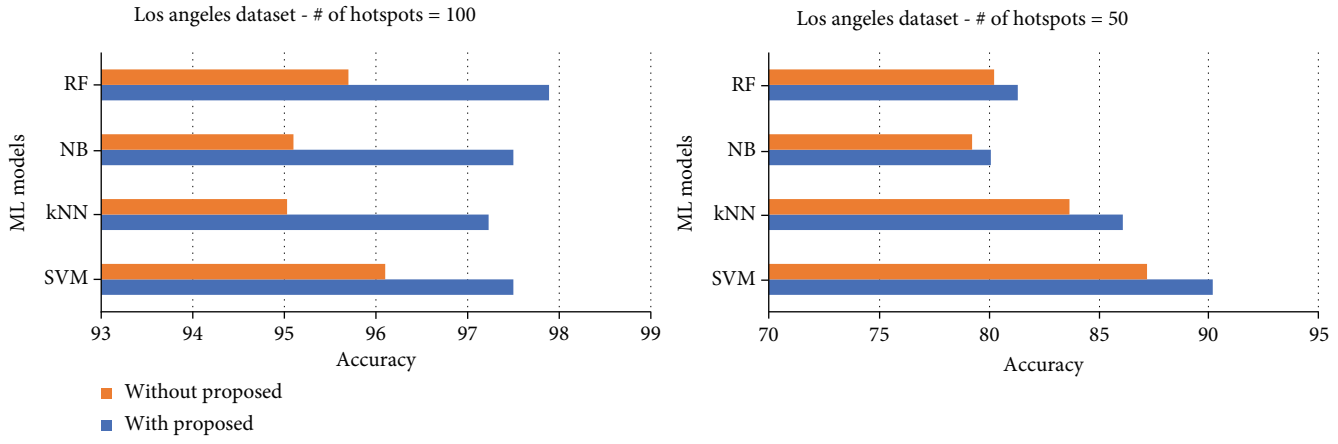

Figure 6: Accuracy of Los Angeles Dataset, when # of hotspots is 100 and 50.

the consideration of crime probabilities, hotspots, and vulnerability analysis. This information helped the classifier to analyse the time series and predict the crime rate better.

4.4. Analysis of Hotspots. Similar crimes are likely to happen frequently at the same place, which includes highly dense areas or low secured places and so on. This information can be captured using a hotspot cluster. Thus, the distance from a cluster is an important factor to consider for crime prediction. If the distance is very low, then it is more likely for a crime to happen.

A hotspot represents a spatial relationship with high frequent crimes [24]. The accurate prediction of crime hotspots helps the police department to take timely action to avoid crime at specific locations. Determining the number of clusters is an important criterion [25, 26]. We have assumed a

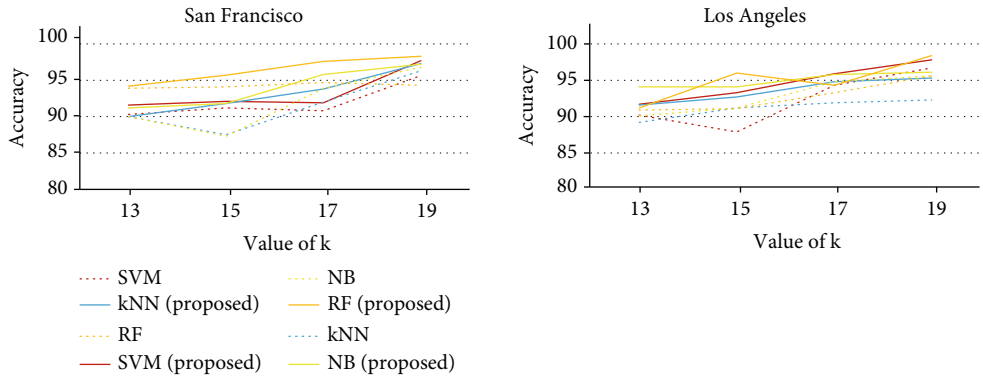

Figure 7: Accuracy measure when n (neighbour) value is 13, 15, 17, and 19.

cluster head per 1 KM^2 or 2 KM^2 and fixed the number of cluster heads as 50 and 100. Figures 5 and 6 show the accuracy comparison of the classifiers with and without the proposed method for the databases San Francisco and Los Angeles, respectively.

4.5. Analysis of Vulnerability. A place is vulnerable for crime when any neighbouring area witness a crime event [27, 28]. We tested the performance of our proposed method using the number of neighbours 13, 15, 17, and 19 [19]. The graph in Figure 7 shows the accuracy of the different classifiers when the value of k changes.

5. Conclusion

Despite many preventive measures, crime rates increase day by day in several regions. This paper concentrates on feature generation methods such as time zone classification, crime probability calculation, analysis of crime hotspots, and vulnerability analysis. The recommended features are fed into four machine learning models which comprises random forest, K nearest neighbour, support vector machines, and Naïve Bayes. The results show that Naïve Bayes produced successful results in predicting the crime incidents.

Symbols

K: How many unique crime events
S: crime locations
m: crime hotspot location
C: clusters
P: temporary points.

Data Availability

Data are available in San Francisco open data https://github.com/ashok0501/ResearchPaperCodes.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] J. Borges, D. Ziehr, M. Beigl et al., "Feature engineering for crime hotspot detection," in *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pp. 1–8, San Francisco, CA, USA, 2017.

[2] S. Yadav, M. Timbadia, A. Yadav, R. Vishwakarma, and N. Yadav, "Crime pattern detection, analysis & prediction," in *2017 International conference of Electronics, Communication and Aerospace Technology (ICECA)*, vol. 1, pp. 225–230, Coimbatore, India, 2017.

[3] B. Sivanagaleela and S. Rajesh, "Crime analysis and prediction using fuzzy c-means algorithm," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 595–599, Tirunelveli, India, 2019.

[4] U. V. Navalgund and K. Priyadharshini, "Crime intention detection system using deep learning," in *2018 International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, pp. 1–6, Kottayam, India, 2018.

[5] Y. Lee, T. Song, H. Kim, D. K. Hant, and H. Ko, "Hostile intent and behaviour detection in elevators," in *4th International Conference on Imaging for Crime Detection and Prevention 2011 (ICDP 2011)*, pp. 1–6, London, 2011.

[6] M. Nakib, R. T. Khan, M. S. Hasan, and J. Uddin, "Crime scene prediction by detecting threatening objects using convolutional neural network," in *2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2)*, pp. 1–4, Rajshahi, Bangladesh, 2018.

[7] S. Ranjan, P. Garhwal, A. Bhan, M. Arora, and A. Mehra, "Framework for image forgery detection and classification using machine learning," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 1872–1877, Tirunelveli, India, 2018.

[8] O. Vynokurova, D. Peleshko, O. Bondarenko, V. Ilyasov, V. Serzhantov, and M. Peleshko, "Hybrid machine learning system for solving fraud detection tasks," in *2020 IEEE Third International Conference on Data Stream Mining Processing (DSMP)*, pp. 1–5, Lviv, Ukraine, 2020.

[9] E. E. Eryilmaz, D. O. Ahin, and E. Kl, "Machine learning based spam e-mail detection system for Turkish," in *2020 5th International Conference on Computer Science and Engineering (UBMK)*, pp. 7–12, Diyarbakir, Turkey, 2020.

[10] B. S. Aldossari, F. M. Alqahtani, N. S. Alshahrani et al., "A comparative study of decision tree and naive bayes machine learning model for crime category prediction in Chicago," in *Proceedings of 2020 the 6th International Conference on Computing and Data Engineering, ser. ICCDE 2020*, p. 3438, New York, NY, USA, 2020.

[11] A. Chowdhary and B. Rudra, "Video surveillance for the crime detection using features," in *Advances in Intelligent Systems and Computing Advanced Machine Learning Technologies and Applications*, pp. 61–71, Cairo, Egypt, 2020.

[12] R. Jain, A. Nayyar, and S. Bachhety, "Factex: a practical approach to crime detection," in *Data Management, Analytics and Innovation Advances in Intelligent Systems and Computing*, p. 503516, Pune, India, 2019.

[13] S. Afzal, M. Asim, A. R. Javed, M. O. Beg, and T. Baker, "URLdeepDetect: a deep learning approach for detecting malicious URLs using semantic vector models," *Journal of Network and Systems Management*, vol. 29, no. 3, 2021.

[14] P. Ashokkumar, N. Arunkumar, and S. Don, "Intelligent optimal route recommendation among heterogeneous objects with keywords," *Computers & Electrical Engineering*, vol. 68, pp. 526–535, 2018.

[15] J. He and H. Zheng, "Prediction of crime rate in urban neighborhoods based on machine learning," *Engineering Applications of Artificial Intelligence*, vol. 106, p. 104460, 2021.

[16] S. A. Chun, V. A. Paturu, S. Yuan, R. Pathak, V. Atluri, and N. R. Adam, "Crime prediction model using deep neural networks," in *Proceedings of the 20th Annual International Conference on Digital Government Research*, pp. 512–514, Dubai United Arab Emirates, 2019.

[17] A. Palanivinayagam and S. Nagarajan, "An optimized iterative clustering framework for recognizing speech," *International Journal of Speech Technology*, vol. 23, no. 4, pp. 767–777, 2020.

[18] A. R. Javed and Z. Jalil, "Byte-level object identification for forensic investigation of digital images," in *2020 International Conference on Cyber Warfare and Security (ICCWS)*, Islamabad, Pakistan, 2020.

[19] N. Deepa, Q. Pham, D. C. Nguyen et al., "A survey on blockchain for big data: approaches, opportunities, and future directions," 2020, http://arxiv.org/abs/2009.00858.

[20] G. T. Reddy, M. P. K. Reddy, K. Lakshmanna et al., "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020.

[21] P. Ashokkumar and S. Don, "Link-based clustering algorithm for clustering web documents," *Journal of Testing and Evaluation*, vol. 47, no. 6, p. 20180497, 2019.

[22] N. Mohan, "Crime analysis with San Francisco open data interactive data visualization with Python and Plotly," June 2019. https://medium.com/@navaneeth.mohan94/crimeanalysis-with-san-francisco-open-data-interactive-data-visualizationwith-python-and-plotly-7b7db7e65d72.

[23] ResearchPaperCodes, "Python code," March 2021, https://github.com/ashok0501/ResearchPaperCodes.

[24] P. Ashok Kumar, G. Shiva Shankar, P. K. R. Maddikunta, T. R. Gadekallu, A. Al-Ahmari, and M. H. Abidi, "Location based business recommendation using spatial demand," *Sustainability*, vol. 12, no. 10, p. 4124, 2020.

[25] A. R. Javed, M. O. Beg, M. Asim, T. Baker, and A. H. Al-Bayatti, "AlphaLogger: detecting motion-based side-channel attack using smartphone keystrokes," *Journal of Ambient Intelligence and Humanized Computing*, vol. 1, 2020.

[26] M. Mittal, C. Iwendi, S. Khan, and A. Rehman Javed, "Analysis of security and energy efficiency for shortest route discovery in low-energy adaptive clustering hierarchy protocol using Levenberg-Marquardt neural network and gated recurrent unit for intrusion detection system," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 6, 2021.

[27] C. Iwendi, Z. Jalil, A. R. Javed et al., "KeySplitWatermark: zero watermarking algorithm for software protection against cyber-attacks," *IEEE Access*, vol. 8, pp. 72650–72660, 2020.

[28] A. Palanivinayagam and D. Sasikumar, "Drug recommendation with minimal side effects based on direct and temporal symptoms," *Neural Computing and Applications*, vol. 32, no. 15, pp. 10971–10978, 2020.