International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Text Analysis of Smart Cities: A Big Data-based Model

Ahmet E. Topcu*[1], Yehia Ibrahim Alzoubi[2], Hüsrev Abdulcelil Karacabey[3]

Submitted: 09/05/2023 **Accepted** : 14/07/2023 **Accepted**: 06/08/2023

Abstract*:* Traditional criminal protection approaches rely heavily on the knowledge of legal authorities' agents. However, it is difficult for them to extrapolate their knowledge to utilize answers in real scenarios. Furthermore, the traditional approaches frequently lose their capacity to avoid problems since they take longer to respond to them. If the crime had occurred, the victims would have already suffered. As a result, it is in the best interests of legal authorities to avoid a crime from happening. We can argue that in order to prevent crime, legal authorities effectively must use data-driven techniques and real-time data analysis. This paper suggests a big data analytics model for analyzing data from various departments. In this study, we proposed using automatic number plate recognition, call detail records, and advanced passenger information data. The proposed model may help to predict potential criminal activities before they occur. Accordingly, this model may allow authorities to stop and prepare for possible criminal activities.

Keywords: *Advanced passenger information, automatic number plate recognition, call detail record, criminal activity, data security*

1. Introduction

Crime refers to the purposeful performance of an act that is commonly judged as socially damaging or hazardous and clearly defined, forbidden, and penalized under criminal law [1]. It is a public omission that every state and legal authority's agency strives to avoid. It is in the best interests of enforcement agencies to avert a crime from occurring rather than seeking to investigate the crime, which accomplishes little to nothing to mend or rehabilitate the victims. In today's environment, a nation's security requires the capabilities of big data to support agencies in their efforts to uncover criminals by merging data from many areas. As a result, it might be claimed that developing a mechanism to evaluate and use massive data is an obvious requirement for modern legal authorities.

Big data are bigger and more complicated data sets than typical data sets. Big data refers to extraordinarily massive data collections that may be computationally examined to identify patterns, correlations, and so on. Therefore, big data may be utilized to resolve security issues that were previously unsolvable [2]. Present legal authorities, particularly those with federal or state-wide responsibilities, obtain terabytes of data daily from various sources (other state agencies, smart city networks, and so on). Yet, they seldom effectively use data to avoid illegal conduct. Big data attributes include volume (amount of data created and maintained), velocity (the rapid rate at which data is acquired and processed), veracity (data's validity and relevance), and variety (several forms of data accessible), referred to as the four Vs [3].

This study uses Call Detail Records (CDR) data, Advanced Passenger Information/Passenger Name Records (API/PNR) data, and Automatic Number Plate Recognition (ANPR) data, all of which are instances of big data. To avoid crime, we made a significant assumption in our study about the person who would conduct the crime, namely that someone with a criminal past is more likely to commit a crime. However, before alerting legal authorities about a suspected crime committed by that specific person, we must first examine real-time data (i.e., CDR, ANPR, API/PNR).

This paper contributes to developing and offering a flexible model for detecting suspected illicit actions to avoid criminal activity. We present several examples of how to use various data kinds to prevent crime. Because the model is adaptable, it can be modified to meet the changing demands of legal authorities in a modular fashion. It may also combine and engage with numerous systems to accurately measure the hazard to a person's or an object's behaviors. The remainder of this study is structured as follows. The related literature is discussed in Section 2. The problem formulation and the supposed model requirements are covered in Section 3. The proposed model architecture and components are presented in Section 4. Section 5 discusses the research's future work and concludes this study.

2. Literature Review

Big data analytics and computer system assistance for legal authorities have long been fascinating research areas [4]. Data-driven methods are well-exemplified by police big data analytics platforms and research that map possible crime sites and utilize machine learning techniques to prevent crime [5, 6]. This section explores the most prevalent works in this context. The prior research on criminal detection is compiled in Table 1. The literature can be broadly divided into three themes: employing

[1] *College of Engineering and Technology,*
American University of the Middle East, Kuwait
ORCID ID : 0000-0003-1929-5358
[2] *College of Business Administration*
American University of the Middle East, Kuwait
ORCID ID : 0000-0003-4329-4072
[3] *Department of Computer Engineering,*
Ankara Yildirim Beyazit University Ankara, Turkey
ORCID ID : 0000-0002-4272-0407
** Corresponding Author Email: ahmet.topcu@aum.edu.kw*

machine learning, classifying data using data mining, and using big data as detection approaches. Since the detection method works with a large amount of data, machine learning has been the main focus of all recent literature.

Table 1. Recent studies on machine learning in crime detection

Study	*Summary*	*Proposed solution*
[2]	Analysis of criminal big data using a variety of cutting-edge visualization tools and big data analytics.	A deep learning technique and Prophet model outperform traditional neural network models
[7]	A deep learning model identifies suspicious behavior and sends an alert message to the proper authority by computing the features using video sequence, and then classifying behavior based on the acquired characteristics.	A deep learning-based model to identify suspicious behavior
[8]	A deep learning model based on characterizing persons in video sequence over time, namely if utilizing a monitoring strategy based on the system is considered with 92% efficacy.	A deep learning-based detection technique for criminal activity in shopping centers
[9]	A framework to detect fraudulent communication that combines deep learning with Spark techniques as well as other machine learning algorithms for fraud detection such as decision tree, SVM, random forest, and KNN.	A deep learning-based framework for wireless communication fraud detection
[10]	A supervised machine learning model was developed for distinguishing between lawful and suspect transactions in terms of money laundering.	A machine learning model for detecting money laundering
[11]	A classification model based on machine learning was developed to categorize suspicious Bengali.	A machine learning model to detect suspicious text
[12]	A naive Bayes, a unique crime detection technique, for the analysis and prediction of crimes. The model correctly predicted only 66% of crimes.	A naive Bayes for the analysis and prediction of crimes
[13]	A model for predicting crime by looking at records of crimes that have already been conducted. The adaptive boosting and random forest algorithm were two methods that were utilized to improve the prediction model's accuracy.	A model for predicting crime based on K-nearest neighbor and decision tree classifications
[14]	A machine learning approach to classify data in the dark web compared to other techniques.	Analyze and classify dark web network data
[15]	A framework for detecting real-time anomalies through the use of big data technology. They designed streaming sliding window local outlier factor corset clustering techniques that were subsequently included in the framework.	A framework based on big data technology that focuses on real-time anomaly detection
[16]	A technique used data clustering methods to extract pertinent traffic patterns from the ANPR data, in order to discover and identify unique patterns and abnormal behavior of multi-vehicle activities.	A vehicle contingent analysis technique based on the ANPR data
[17]	A centroids-based activity categorization method after using data-mining tools to analyze activity patterns of vehicles from ANPR data.	An approach for categorizing activity patterns of vehicles using ANPR data
[18]	A framework that helps predict travelers' behaviors based on analyzing three data sources: PNR, the share of wallet, and Web trends.	A framework based on PNR, the share of wallet, and Web trends data to predict travelers' behaviors
[19]	The car license plate is extracted from the digital image using MATLAB software, and a violation SMS is sent to the offender with the help of an Arduino and GSM module (SIM900).	A technique based on ANPR to detect automobiles that disregard traffic signals
[20]	A graph analysis software to examine CDR data belonging to the suspect that was obtained from the service providers in a variety of formats.	A technique to predict suspected individuals from their mobile CDR
[1]	A proposal that used big data techniques to collect and prepare CDR data.	A proposal for a system that can collect CDR data to be retrieved for criminal detection
This study	A model that predicts criminal behaviors based on the analysis of three big data resources: CDR, API/PNR, and ANPR.	A model for criminal behavior prediction using CDR, ANPR, and API/PNR sources

Feng et al. [2] analyzed criminal big data from three US cities using a variety of cutting-edge visualization tools and big data analytics. They revealed that both the Prophet model and the deep learning technique outperform traditional neural network models [2]. Amrutha et al. [7] suggested a deep learning model to identify suspicious behavior and send an alert message to the proper authority. They tracked and analyzed real-time CCTV data to assess the proposed model. This model includes a portion to compute the features using video sequence, and another portion to predict the class as normal or suspicious based on the acquired features [7]. Qin et al. [8] suggested a model that made use of publicly available video surveillance data collecting. The model is based on characterizing persons in video sequence over time, namely if utilizing a monitoring strategy based on the system is considered for addressing occlusions in crowded situations with 92% efficacy. The suggested model was evaluated using actual and private datasets [8]. Sanober et al. [9] provided a framework that integrates a deep learning approach with Spark as well as other algorithms to detect frauds such as SVM, KNN, random forest, and decision trees. Comparative analysis was conducted utilizing different parameters. Both the training and testing datasets achieved greater than 96% accuracy. The dataset utilized covers transactions that happened in two days, including 492 fraudulent transactions out of 284,807 totaling 0.172% of all transactions [9].

Jullum et al. [10] proposed an approach based on machine learning for selecting which bank transactions must be physically scrutinized for probable money laundering. Regular legal transaction data, those data labeled as questionable by the bank's internal warning system, and probable money laundering data were used to train the proposed model. The finding indicated that the conventional practice of not employing regular (underexplored) transactions in model training might result in sub-optimal outcomes [10]. Sharif et al. [11] designed a classification approach based on machine learning to categorize Bengali text as suspicious or non-suspicious based on its original contents. A collection of machine learning classifiers with diverse characteristics was employed, along with 7000 Bengali text documents, 5600 of which were used for training and 1400 for testing. The model's performance was evaluated by comparing it to the human benchmark and existing machine-learning approaches. The model has the greatest accuracy of 84.57% [11].

Wibowo and Oesman [12] proposed a naive Bayes-based criminal detection technique that was used for crime forecasting. The model's accuracy in forecasting crimes was only 66%, and it neglected to take computational speed, resilience, and scalability into account [12]. Hossain et al. [13], suggested a model for predicting crime by looking at a dataset including records of crimes that have already been conducted and their patterns. The suggested model primarily uses decision tree and K-nearest neighbor techniques. The adaptive boosting and random forest algorithm were two methods that were utilized to improve the prediction model's accuracy. The offenses were separated into regular and rare categories in order to improve model findings. The most frequent crimes fell into the regular class, whereas the least common crimes fell into the rare class [13]. Data on criminal activity in San Francisco, California, over a 12-year period was used to feed the proposed model. Using the random forest algorithm and under and over-sampling techniques, the model accuracy increased to 99.16% [13]. Rawat et al. [14] used several approaches to analyze and classify dark web network data. Worms, DDos, backdoors, Spam, DoS attacks, and malicious material were all detected by the proposed study. The suggested study used a term frequency-inverse document frequency and light gradient boosted machine algorithm approach. Based on experiment results, the light-gradient boosted machine approach surpasses the other algorithms with an accuracy of 98.97% [14].

Habeeb et al. [15] suggested a big data-based framework for real-time anomaly detection. They created streaming sliding window local outlier factor corset clustering techniques, which were subsequently integrated into the system. Several current techniques, including spectral clustering, isolated forest, K-means, and agglomerative clustering, were used to assess the proposed framework. The assessment results demonstrated the usefulness of the suggested framework, with a considerably better accuracy rate of 96.51% and reduced memory consumption when compared to other algorithms [15]. Chen et al. [18] examined three forms of travel data: PNR, the share of wallet, and Webtrends to better understand airline passenger behavior. The authors provided a summary and insights on individual passengers' websites and mobile usage by fusing this Webtrends data with additional information sources. Additionally, they showed how to gain a thorough insight into passenger travel behavior and social networks [18]. Homayounfar et al. [16] created a data mining program based on ANPR data for vehicle convoy analysis. In order to discover and identify unique patterns and abnormal behavior of multi-vehicle activities, the authors employed data clustering methods to extract pertinent traffic patterns from the ANPR data [16]. Sun et al. [17] suggested a method for categorizing vehicle activity patterns using ANPR data obtained from embedded ANPR cameras. They suggested a centroids-based activity categorization method after using data-mining tools to analyze vehicle activity patterns. Automatic aberrant vehicle behavior that results in the identification and alerting could speed up the surveillance and processing of emergency events while minimizing the impact on traffic flow [17].

To trace suspects or criminals, Kumar et al [20] suggested a technique for analyzing mobile phone conversations made by suspects. In order to solve the case, the authors used graph analysis software to examine CDR data belonging to the suspect that was obtained from the service providers in a variety of formats [20]. An ANPR system was presented by Chaithra et al. [19] to identify automobiles that disregard traffic signals by obtaining their license plate from digital photos. Additionally, it immediately sends the owner of the vehicle a violation SMS. The number plate region was extracted using image segmentation algorithms. The segmented image was compared to the template images using correlation techniques. The obtained information was then used to compare with the database's records. The car license plate was extracted from the digital image using MATLAB software, and a violation SMS was sent to the offender a minute after the offense with the help of an Arduino and GSM module (SIM900) [19]. Khan et al. [1] proposed an idea for a system that accepts cellphone number(s) as input and extracts pertinent CDRs, so producing several databases of CDRs. After that, it examines these databases and discovers numerous connections between different suspects (mobile numbers), producing the results of its study as output. This database may include contact information like phone numbers, call times, and locations [1].

In this study, we primarily concentrate on information gathered through CDR, ANPR, and API/PNR, which are used by legal authorities worldwide in their inquiries. Because it contains location information and depicts interactions or communications between people, CDR is a useful data type for legal authorities [1, 21]. The cameras placed at road crossings produce the ANPR data, which include details such as the license plate numbers of the passing cars and the time of passage. The role of vehicles in criminal actions is the primary justification for our choice of the ANPR data type. The use of ANPR is also possible for a number of other purposes, including monitoring traffic signal infractions and classifying vehicle behavior [22]. The other form of data is API/PNR data, which is mostly used by airlines for business-related purposes or other research projects like predicting passengers' nationalities using. Since terrorists and combatants often fly to their destination and our study is forwarded for security purposes, we deployed PNR data [23]. It is important to note that this paper is based on the work of [24].

3. Model Requirements

We argue that in order to prevent crime, we must identify those people who are most likely to perpetrate it. In order to estimate the risk to a person or an object (e.g., a car), we must integrate archival data (criminal records and other confidential data) with real-time data (such as contacts and location data) analysis. Following the risk assessment, we will alert the appropriate authorities so they may take action. This section explores the requirements and parameters of the proposed model.

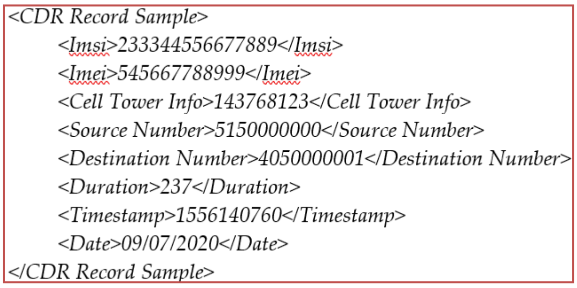
3.1. Model Parameters

Several parameters were developed to meet the requirements of the proposed model. These parameters are introduced below and discussed in more detail in Section 4.

- Suspicion Rate (SR): It's a term we offer for the model, which is represented by a numerical value. There is an SR attached to everything, whether it is a person or an object. It has a zero beginning and never ends. By examining its activities, we continuously raise or lower the SR of the person or object. We may alert legal authorities when the SR exceeds a certain specified level.
- Personal Suspicion Factor (PSF): We process the data collected that we produce for the model to develop the PSF data. PSF has a numerical value between 0 and infinity. Each person's criminal history is examined in order to compute the PSF. Varying types of criminal activities and the number of criminal records will have different effects on the PSF value. We presume that a person's previous criminal behavior and potential future criminal behavior are strongly correlated. Legal authorities must pay closer attention to a person's activities if their PSF is higher. PSF will have an impact on SR; a greater PSF value corresponds to a higher SR value.
- Vehicle Suspicion Factor (VSF): We process the data collected for the model to produce the VSF data. It has a numerical value between 0 and infinity. Criminal records are examined first in order to determine the VSF. The vehicle's VSF value is raised if it is involved in criminal activity. The criminal history and the nature of the illegal activity have distinct effects on the VSF rating. As criminal records change, the VSF value must be updated. A greater VSF value corresponds to a higher SR value.
- Action Value (AV): We also define and suggest this term, which is also a numerical value. It begins at zero and continues till infinity. Everything has an AV attached to it, whether it be a person or an object. It may be predefined or modified on a regular basis in response to variations in PSF or VSF. When a person or an object's SR equals or exceeds the predefined AV level, it indicates that there is a significant chance that the person or thing in question will be a victim of criminal activity.
- Location and Time Risk Factor (LOTRF): The LOTRF data have to do with how much risk a place presents at a certain moment. LOTRF data are made up of the place, time, and value (e.g., location: Ulus/Ankara, time: noon, value: 9). The data are compiled by looking through police records of criminal activity. We presume that there is a connection between potential criminal activity and criminal activity in a given place in the past. Each period at a particular location will have a numerical value in the data. The range of the LOTRF value is zero to infinity. The risk of criminal activity increases as the value increases. The LOTRF values are one of the most crucial pieces of information that the model will use to change SR, thus we must update them regularly to be effective.

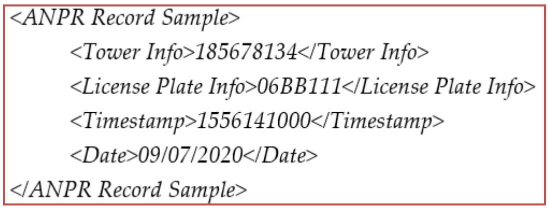
3.2. Data Sources

Data sets represent an action of an object or a person, or other processed data make up the inputs for this model including CDR, ANPR, API/PNR, and background data (e.g., criminal records, etc.). Because we needed a court order or a warrant, we were not able to collect real data to test our model. However, we are aware of the data fields and how to use them. The data that the model will use is described below [24]. We can use other sorts of data in addition to the ones we specified in the model.



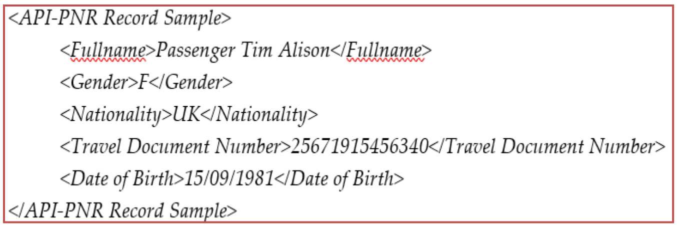
```
<CDR Record Sample>
        <Imsi>233344556677889</Imsi>
        <Imei>545667788999</Imei>
        <Cell Tower Info>143768123</Cell Tower Info>
        <Source Number>5150000000</Source Number>
        <Destination Number>4050000001</Destination Number>
        <Duration>237</Duration>
        <Timestamp>1556140760</Timestamp>
        <Date>09/07/2020</Date>
</CDR Record Sample>
```

Fig. 1a. CDR data sample



```
<ANPR Record Sample>
        <Tower Info>185678134</Tower Info>
        <License Plate Info>06BB111</License Plate Info>
        <Timestamp>1556141000</Timestamp>
        <Date>09/07/2020</Date>
</ANPR Record Sample>
```

Fig. 1b. ANPR data sample



```
<API-PNR Record Sample>
        <Fullname>Passenger Tim Alison</Fullname>
        <Gender>F</Gender>
        <Nationality>UK</Nationality>
        <Travel Document Number>25671915456340</Travel Document Number>
        <Date of Birth>15/09/1981</Date of Birth>
</API-PNR Record Sample>
```

Fig. 1c. API/PNR data sample

3.2.1. Call Detail Record

CDR is a record that includes specific details about a telecom transaction, such as the opening time, elapsed time, length, call participants, mobile IDs, source number, destination number, cell tower information, and requested URLs. The telecom sector uses CDRs for essential purposes like billing and charging, but our goal in this study is to detect movement and examine connections. CDR qualifies as a big data source since it possesses the three 3Vs of big data (i.e., volume, velocity, and veracity) [25]. Figure 1a shows a CDR data sample.

3.2.2. Automatic Number Plate Recognition

CDR is a record that includes specific details about a telecom transaction, such as the opening time, elapsed time, length, call participants, mobile IDs, source number, destination number, cell tower information, and requested URLs. The telecom sector uses CDRs for essential purposes like billing and charging, but our goal in this study is to detect movement and examine connections. CDR qualifies as a big data source since it possesses the three 3Vs of big data (i.e., volume, velocity, and veracity) [25]. Figure 1a shows a CDR data sample.

3.2.3. Advance Passenger Information, Passenger Name Records

The entire name, date of birth, gender, country, booking date, the number of travel documents, etc., of a passenger, are all included in the electronic data interchange system known as API/PNR. PNR is a collection of data generated when a booking is made for travel. It is produced by airlines or approved agents like travel agencies [23]. For instance, passengers heading to Turkey must give API before checking in. In accordance with international standards, travel companies must also provide passenger data about the countries they visit before the travelers arrive. Airlines use passenger data for various goals, including identifying influential passengers, providing superior customer service to foster customer loyalty, and identifying customer behavior. Because API data possesses big data qualities, we can think of it as big data. Our proposed model uses passenger data to update the SR value after detecting a movement. Figure 1c shows an example of API/PNR data.

3.2.4. Togetherness Analysis Suspicion Factor (TASF)

For the model, we process and produce the TASF data. In terms of numbers, it has a value between 0 and infinity. The proposed model employs CDR and ANPR data to generate the TASF. By default, legal authorities should obstruct the movement or behavior of many people or vehicles if all or some of them have high VSF or PSF values because they are quite likely to engage in an offense. The model employs CDR data since it provides location information and shows a relationship between people, which helps it determine connectedness among a group of people. The TASF will be the total of all PSF values for the persons within the togetherness and, accordingly, update their SR.

On the other hand, the model uses ANPR data to determine whether a group of vehicles moves together because ANPR data contains information about the time and location of the vehicles [16]. Following the detection of togetherness, each vehicle's VSF value will be evaluated. The model will then estimate the TASF (TASF is the total of all the group’s vehicle VSF values) and adjust each vehicle's SR correspondingly. Criminals (such as smugglers) frequently place a vehicle in front of the actual delivery automobile, known as the "pioneer," that checks for police vehicles. When the "pioneer" vehicle sees or detects legal authorities, it alerts the actual delivery automobile to avoid the police, for instance. To perpetrate robbery, criminals collectively follow vehicles like bank cash carriers [26].

4. Proposed Model

4.1. Model Architecture

The objective of this study is to create a model that may operate continuously, irregularly, or in response to an activity (e.g., an ANPR, a phone conversation, or a boarding at the airport). These data are then examined, and the SR values are updated. Big data obtained from various systems and sources, including CDR, ANPR, and API/PNR data, are modeled to be the basis for the model's operation. All actions will result in an increase in the SR of zero or another positive integer number. When there is no rise in the person’s or object’s SR, the SR values will be reduced. Each person or object's SR values may be decreased differently, depending on their PSF or other relevant data, which could be constantly determined. For instance, the decrease rate or quantity may be slower or lower for a person with a high PSF or a car with a high VSF. To choose the best reduction technique, it is necessary to evaluate field feedback and develop a strategy to identify the sweet spot for each person or object. Legal authorities can monitor the zone or check a person’s ID after receiving the alert. The sign does not indicate the existence of actual criminal activity, but it does allow legal authorities to prevent a potential crime by acting proactively. We must emphasize that this does not mean the detected individuals engage in criminal activity. And we think that taking preventive measures can lower crime rates and enhance public safety and well-being. Figure 2 shows the proposed model architecture for the detection model that we have proposed.

- **Real-time data:** The model receives real-time data from various sources and organizations—for example, telecommunications corporations, airlines, and traffic control authorities before sending it to the Memcache unit.
- **Caching ops:** Data retrieved should be analyzed in real time. We recommend saving data for 20 to 30 minutes before destroying it. Any in-memory data structure (e.g., Redis) can be utilized as a transitory database for real-time data. Technology for improved resource management might be employed to host the Memcache virtualization.
- **Archive storage:** This storage compromised PSF, VSF, and LOTRF data. NoSQL or Relational databases, such as PostgreSQL, Oracle, or MongoDB, can be used as a storage technology.
- **Person/object storage:** This storage should include two categories of data to be modified and acted on for each object or person: AV and SR. NoSQL or Relational databases, such as PostgreSQL, Oracle, or MongoDB, can be used as a storage technology.
- **Police/Legal authorities**: We need a continual link to the legal authorities in the model to convey the alerts generated based on the behaviors of persons and objects.
- **SR analysis process**: This is the major data analysis unit. This process comprises the TASF, CDR, ANPR, and API/PNR processes. Any programming language, such as Python or Java, can create all the processes.
    - TASF analysis process: A continuous process that detects connectedness between humans or objects. It makes advantage of the CDR and ANPR data saved in Memcache. After identifying togetherness, the TASF value will be calculated, and the SR values of objects and persons will be updated accordingly.

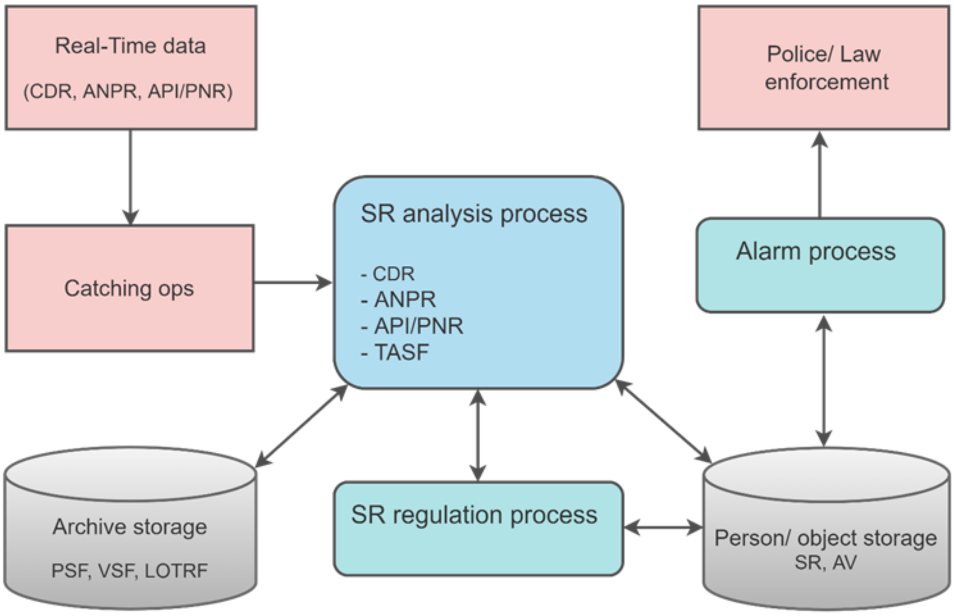


Fig. 2. The proposed model's architecture

  - CDR analysis process: A continuously running process to analyze the CDR data saved in Memcache. This process may access archive storage as well as person/object storage. When a link is detected, it obtains PSF and LOTRF data from archival storage and accordingly updates the person's SR value.
  - ANPR analysis process: A continuously running process to analyze the ANPR data saved in Memcache. This process can access archival storage as well as person/object storage. When a link is detected, it obtains VSF and LOTRF data from archival storage before updating the objects or person's SR values accordingly.
  - API/PNR analysis process: It is a continuous process that reduces the values of SR in person/object storage. The reduction of SR depends on the time duration during which an object's or person's SR does not increase. Any programming language, such as Python or Java, may be used to create it.
- **Alarm process:** It is a continuous process that contrasts the AV and SR values of objects and persons in the person/object storage. The legal authorities will be notified if an object or a person's SR exceeds their AV. Any programming language, such as Python or Java, may be used to create it.
- **SR regulation process:** It is a continuous process that reduces the values of SR in person/object storage. The reduction of SR depends on the time duration during which an object's or person's SR does not increase. Any programming language, such as Python or Java, may be used to create it.

4.2. Call Detail Record

CDR depicts a person-to-person and includes characteristics like the time and place of the conversation. In this model, the PSF for both will be collected after detecting the persons using CDR data, followed by the LOTRF based on location and time. The model will then boost each person's SR by the total of their LOTRF and PSF values. For example, If A phones or texts B, this interaction will produce a CDR record. For our model to be successful, the CDR record should be examined in real or near-real time. Then, the model computes the total of each person's LOTRF and PSF values and then increases the SR values of B and A. Figure 3 depicts the steps for CDR analysis and a sample pseudocode of the CDR process. As previously mentioned, we can argue that a high PSF value indicates a likely proclivity for criminal action. We also presume that persons' locations be considered; being in an area where crime rates have historically been high must be regarded as a potential criminal activity. Following the growth of B and A's SR and reaching their AV level, an alert must be generated for both of them and relevant legal authorities' procedures to be conducted.

4.3. Automatic Number Plate Recognition

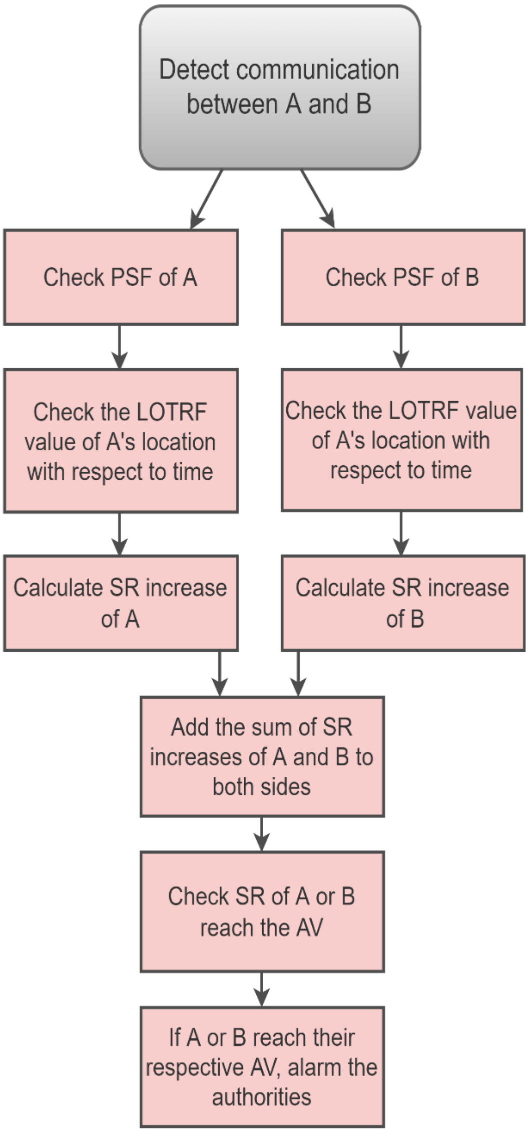
Vehicle motion is shown via ANPR data. After detecting vehicle movement, the model retrieves the VSF of the vehicle, the PSF of the vehicle's owner, and the LOTRF value. The model will then raise the SR of the vehicle and owner by the total of the previously indicated LOTRF, PSF, and VSF values. For example, the ANPR data should be examined in real or near-real time to be successful. Following detection, the model computes the total of the car's VSF, the vehicle location's LOTRF, and the driver's PSF before raising their SR values. The LOTRF value will be derived from the ANPR data's time and location information. Figure 4 depicts the steps and sample pseudocode of ANPR analysis. We consider that a vehicle with a high VSF value has the capability to be employed for potential illegal actions, and its movement should be treated seriously if its SR value approaches its AV level. We presume that the driver and owner of the automobile are the same people. We also consider that location is a significant aspect in dealing with illicit operations involving cars, which is why the LOTRF is included in the model. If, after updating the vehicles and the driver's SR, either the vehicle or the driver's SR reaches their particular AV, both will receive an alert, and necessary legal authorities' action should be taken.

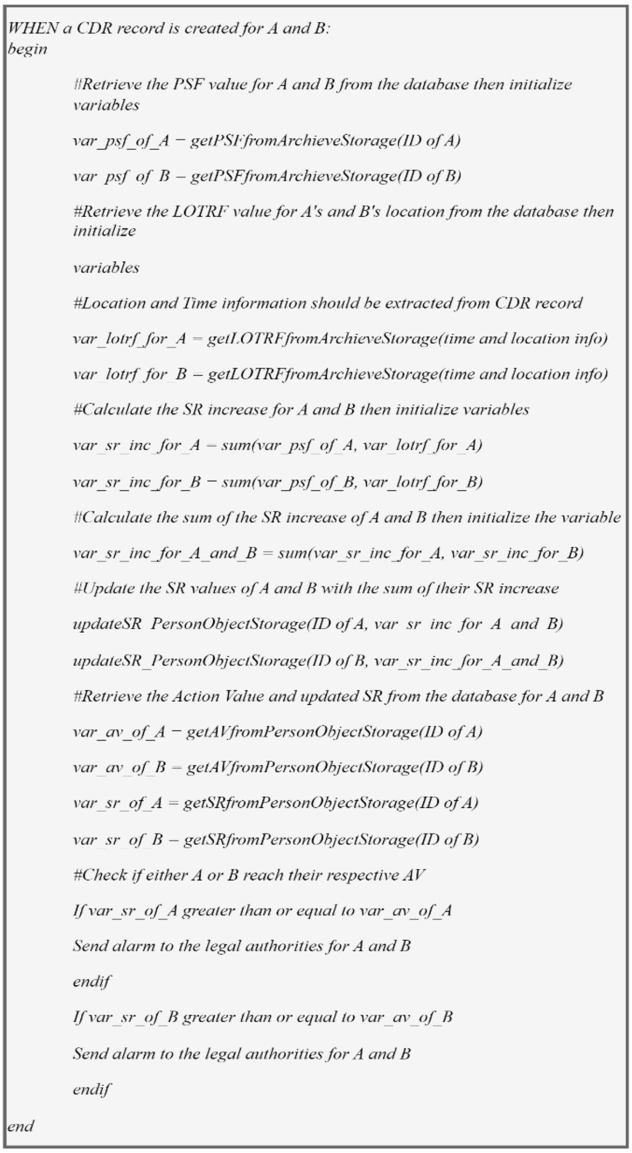
4.4. Advance Passenger Information/Passenger Name Records

API/PNR data reflect the passenger's movements locally and globally; they contain properties such as passenger personal data, booking details, travel papers, and so on. After detecting motion, the model will get the passenger's PSF. The model will then update the SR of the passenger by the PSF value. Figure 5 depicts the steps and sample pseudocode for API/PNR analysis. For instance, if passenger data, such as a PNR or API record, are made and transmitted from airline officials to legal authorities, this data must be supplied and processed by our model in real or near-real-time for the model to be successful. Following detection, the model obtains the passenger's PSF value and increases the SR by the PSF value. We presume that a high PSF passenger poses a possible security risk to the aircraft and the locations the person is flying to. If the passenger's SR hits their AV level, an alarm will be generated, and necessary action must be taken at the airport. This model has the potential to improve and assist passenger and airline security.

4.5. Togetherness Analysis

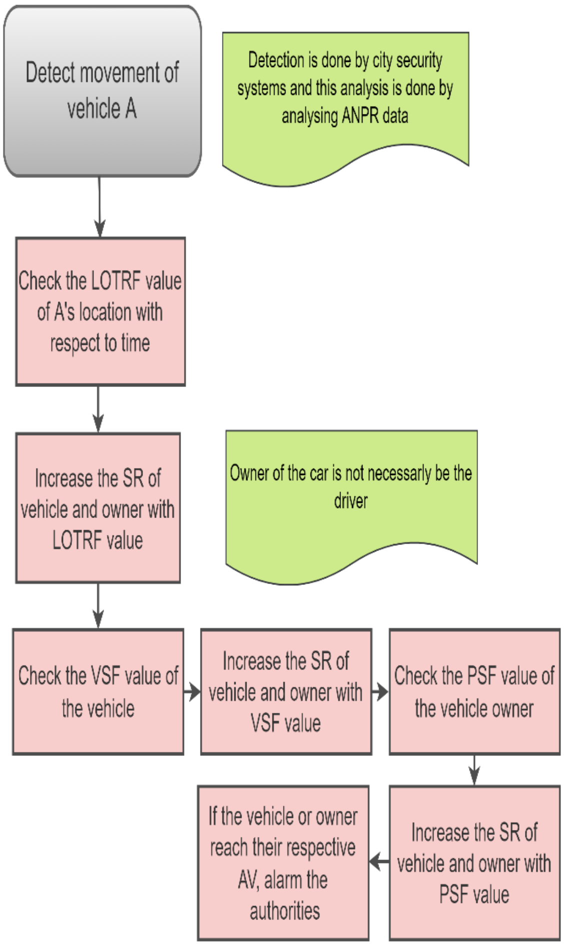
To identify togetherness, the model must continually analyze CDR and ANPR data in real-time. We argue that persons with high PSF or vehicles with high VSF are a signal of potential criminal conduct. The TASF is the sum of the VSF of a group of cars or the PSF of a group of persons moving together. When the model identifies togetherness among persons or vehicles, it increases their SR by the determined TASF value. Suppose a person or vehicle within the identified group hits their AV level. In that case, an alarm must be generated for the whole group of individuals or cars, and legal authorities should be notified. Figures 6 and 7 depict the steps and sample pseudocode for TASF analysis utilizing ANPR and CDR data. Closed Travelling Companion Discovery on the ANPR Data Stream (COINCIDENT) algorithm may rapidly and constantly identify vehicles that move as a group in real-time or near real-time to detect vehicle togetherness. Then, our model may proceed by verifying the VSF of the cars in the group, calculating TASF, and manipulating the SR of the vehicles.


Detect communication between A and B
Check PSF of A
Check PSF of B
Check the LOTRF value of A's location with respect to time
Check the LOTRF value of A's location with respect to time
Calculate SR increase of A
Calculate SR increase of B
Add the sum of SR increases of A and B to both sides
Check SR of A or B reach the AV
If A or B reach their respective AV, alarm the authorities



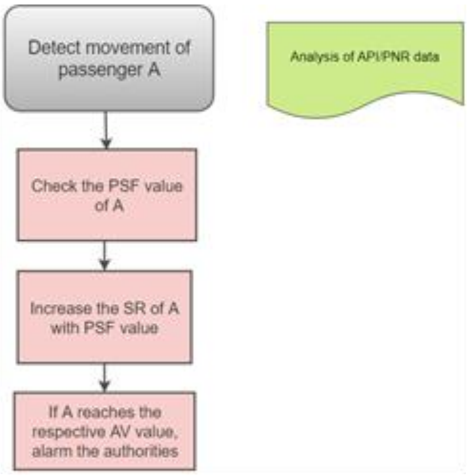
```
WHEN a CDR record is created for A and B:
begin
        #Retrieve the PSF value for A and B from the database then initialize
        variables
        var_psf_of_A − getPSFfromArchieveStorage(ID of A)
        var_psf_of_B – getPSFfromArchieveStorage(ID of B)
        #Retrieve the LOTRF value for A's and B's location from the database then
        initialize
        variables
        #Location and Time information should be extracted from CDR record
        var_lotrf_for_A = getLOTRFfromArchieveStorage(time and location info)
        var_lotrf_for_B – getLOTRFfromArchieveStorage(time and location info)
        #Calculate the SR increase for A and B then initialize variables
        var_sr_inc_for_A = sum(var_psf_of_A, var_lotrf_for_A)
        var_sr_inc_for_B − sum(var_psf_of_B, var_lotrf_for_B)
        #Calculate the sum of the SR increase of A and B then initialize the variable
        var_sr_inc_for_A_and_B = sum(var_sr_inc_for_A, var_sr_inc_for_B)
        #Update the SR values of A and B with the sum of their SR increase
        updateSR_PersonObjectStorage(ID of A, var_sr_inc_for_A_and_B)
        updateSR_PersonObjectStorage(ID of B, var_sr_inc_for_A_and_B)
        #Retrieve the Action Value and updated SR from the database for A and B
        var_av_of_A − getAVfromPersonObjectStorage(ID of A)
        var_av_of_B = getAVfromPersonObjectStorage(ID of B)
        var_sr_of_A = getSRfromPersonObjectStorage(ID of A)
        var_sr_of_B – getSRfromPersonObjectStorage(ID of B)
        #Check if either A or B reach their respective AV
        If var_sr_of_A greater than or equal to var_av_of_A
        Send alarm to the legal authorities for A and B
        endif
        If var_sr_of_B greater than or equal to var_av_of_B
        Send alarm to the legal authorities for A and B
        endif
end
```

Fig. 3. CDR data analysis for suspicion detection (based on [24]


Detect movement of vehicle A
Detection is done by city security systems and this analysis is done by analysing ANPR data
Check the LOTRF value of A's location with respect to time
Increase the SR of vehicle and owner with LOTRF value
Owner of the car is not necessarly be the driver
Check the VSF value of the vehicle
Increase the SR of vehicle and owner with VSF value
Check the PSF value of the vehicle owner
If the vehicle or owner reach their respective AV, alarm the authorities
Increase the SR of vehicle and owner with PSF value

```
WHEN an ANPR record is created for the vehicle A:

begin

        #Retrieve the LOTRF value for A's
        location from the database then initialize the variable
        #Location and Time information should
        be extracted from ANPR record
        var_lotrf_for_A = getLOTRFfromArchieveStorage(time and location info)
        #Update the SR values of A and the Owner with the LOTRF value
        updateSR_PersonObjectStorage(ID of A, var_lotrf_for_A)
        updateSR_PersonObjectStorage(ID of Owner, var_lotrf_for_A)
        #Retrieve the VSF value for A from the database then initialize the variable
        var_vsf_of_A = getVSFfromArchieveStorage(ID of A)
        #Update the SR values of A and the Owner with the VSF value
        updateSR_PersonObjectStorage(ID of A, var_vsf_of_A)
        updateSR_PersonObjectStorage(ID of Owner, var_vsf_of_A)
        #Retrieve the PSF value of the Owner from the database then initialize the
        variable
        var_psf_of_owner = getPSFfromArchieveStorage(ID of Owner)
        #Update the SR values of A and the Owner with the PSF value
        updateSR_PersonObjectStorage(ID of A, var_psf_of_owner)
        updateSR_PersonObjectStorage(ID of Owner, var_psf_of_owner)
        #Retrieve the Action Value and updated Suspicion Rate from the database
        for A and the Owner
        var_av_of_A = getAVfromPersonObjectStorage(ID of A)
        var_av_of_owner = getAVfromPersonObjectStorage(ID of Owner)
        var_sr_of_A = getSRfromPersonObjectStorage(ID of A)
        var_sr_of_owner = getSRfromPersonObjectStorage(ID of Owner)
        #Check if either A or the Owner reach their respective Action Value
        If var_sr_of_A greater than or equal to var_av_of_A
         Send alarm to the legal authorities for A and the Owner
        endif
        If var_sr_of_owner greater than or equal to var_av_of_owner
         Send alarm to the legal authorities for A and the Owner
        endif

end
```

Fig. 4. ANPR data analysis for suspicion detection (based on [24])

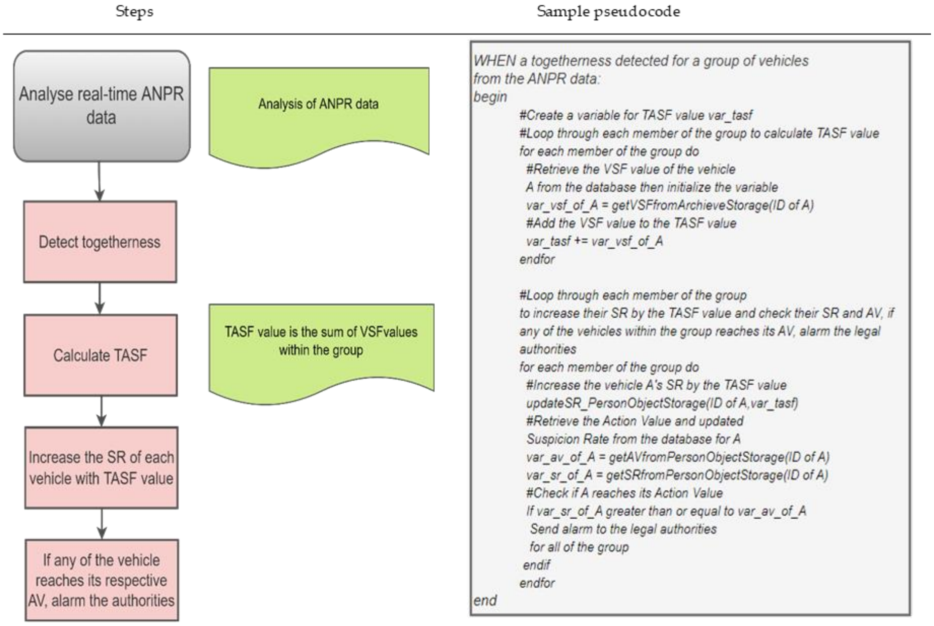

Detect movement of passenger A
Analysis of API/PNR data
Check the PSF value of A
Increase the SR of A with PSF value
If A reaches the respective AV value, alarm the authorities

```
WHEN an API/PNR record is created for the passenger A:

begin

        #Retrieve the PSF value for the A from
        the database then initialize the variable
        var_psf_of_A = getPSFfromArchieveStorage(ID of A)
        #Update the SR value of A with the PSF value
        updateSR_PersonObjectStorage(ID of A, var_psf_of_A)
        #Retrieve the Action Value and updated SR from the
        database for A
        var_av_of_A = getAVfromPersonObjectStorage(ID of A)
        var_sr_of_A = getSRfromPersonObjectStorage(ID of A)
        #Check if A reaches his/her their Action Value
        If var_sr_of_A greater than or equal to var_av_of_A
        Send alarm to the legal authorities for A

    endif

end
```

Fig. 5. APNR/API data analysis for suspicion detection (based on [24])

Instead of the visual inspection performed by legal authorities, our model must do real-time and automated analysis. Clustering approaches on smartphone subscribers based on location data in CDR data may be used for real-time detection, then merging it with conversation data inside the CDR. Then, the proposed model should examine all of the discovered group's PSF values, calculate TASF values, and adjust their SR. We argue this strategy is critical due to the complexities of organized criminal groups. Several forms of organized criminals exist, such as biker gangs involved in various illegal operations [27].

Another example is street gangs, which are involved in thefts, murders, drug trafficking, and gang battles [28]. Accordingly, we argue that when togetherness is discovered for a group of persons or vehicles, we should notify legal authorities if any persons or vehicles within those groups reach their AV level. We can argue that our model will be a valuable tool to legal authorities in combating organized criminal activity and communicating with other machine learning solutions for predicting attacks on social media [29], hence increasing citizens' safety and security and also providing a model and input for machine learning solution by analyzing the data provided by urban residents as city problems to build effective and more secure smart cities by smart text analysis [30].


Steps
Sample pseudocode
Analyse real-time ANPR data
Analysis of ANPR data
Detect togetherness
Calculate TASF
TASF value is the sum of VSFvalues within the group
Increase the SR of each vehicle with TASF value
If any of the vehicle reaches its respective AV, alarm the authorities

```
WHEN a togetherness detected for a group of vehicles
from the ANPR data:
begin
        #Create a variable for TASF value var_tasf
        #Loop through each member of the group to calculate TASF value
        for each member of the group do
          #Retrieve the VSF value of the vehicle
          A from the database then initialize the variable
          var_vsf_of_A = getVSFfromArchieveStorage(ID of A)
          #Add the VSF value to the TASF value
          var_tasf += var_vsf_of_A
        endfor

        #Loop through each member of the group
        to increase their SR by the TASF value and check their SR and AV, if
        any of the vehicles within the group reaches its AV, alarm the legal
        authorities
        for each member of the group do
          #Increase the vehicle A's SR by the TASF value
          updateSR_PersonObjectStorage(ID of A,var_tasf)
          #Retrieve the Action Value and updated
          Suspicion Rate from the database for A
          var_av_of_A = getAVfromPersonObjectStorage(ID of A)
          var_sr_of_A = getSRfromPersonObjectStorage(ID of A)
          #Check if A reaches its Action Value
          If var_sr_of_A greater than or equal to var_av_of_A
           Send alarm to the legal authorities
           for all of the group
         endif
        endfor
end
```

Fig. 6. TASF calculation from ANPR data (based on [24])

5. Conclusions and Future Work

In this paper, we offer a model for identifying suspected activities. The following data types are referenced in this study: CDR, ANPR, and API/PNR. These three data types have a common feature: they all indicate mobility, and the model governs the SR of persons and objects accordingly. Furthermore, these three data represent big data. The proposed model can be used in big data analysis to prevent potential criminal activity and assist legal authorities in improving their strategies and processes. The suggested model is versatile and is not limited to the data types specified above for SR updating. It may be extended and upgraded as needed by incorporating new data formats or technological developments, such as 5G cellular network technology.

Although this paper was prepared with great care and attention to detail, it does have certain limitations. First, due to technical and legal constraints, we were unable to obtain real-world data to evaluate the planned model. To evaluate the model, the accurate and complete construction of the entire model must be evaluated using real scenarios. Furthermore, because such technologies cannot be directly deployed in the open world, implementation is a significant problem [31]. Furthermore, there are a few technological challenges in the model since the model will deal with a large amount of data, which makes data analysis a time-consuming task.

One possible automation of the proposed model would be to develop a system that can forecast and anticipate crime hotspot zones in a city. Our future research will begin with creating a machine learning-based system based on this model capable of anticipating and identifying patterns of such crimes. Although current methods significantly influence crime prevention, this might be the next great strategy that results in a revolutionary movement to decrease the crime rate. In the future, we intend to fulfill our ongoing model by running actual scenarios to assess its efficiency and sustainability. Establishing such a model and finding a means to integrate diverse data into it for crime prevention reasons will assist legal authorities.

Author contributions

Ahmet E. Topcu: Conceptualization, Investigation, Methodology, Validation **Yehia Ibrahim Alzoubi:** Writing-Reviewing and Editing, Methodology **Hüsrev Abdulcelil Karacabey:**Writing-Original draft preparation, Investigation, Data curation, Software.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] S. Khan, F. Ansari, H. A. Dhalvelkar, and S. Computer, "Criminal investigation using call data records (CDR) through big data technology," in *International Conference on Nascent Technologies in Engineering (ICNTE)*, Vashi, India. IEEE, 2017, pp. 1-5.

[2] M. Feng *et al.*, "Big data analytics and mining for effective visualization and trends forecasting of crime data," *IEEE Access*, vol. 7, pp. 106111-106123, 2019.

[3] M. I. Pramanik, R. Y. Lau, W. T. Yue, Y. Ye, and C. Li, "Big data analytics for security and criminal investigations," *Wiley interdisciplinary reviews: data mining and knowledge discovery*, vol. 7, no. 4, p. e1208, 2017.

[4] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, "Crime analysis through machine learning," in *9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Vancouver, Canada. IEEE, 2018, pp. 415-420.

[5] H. Yu and C. Hu, "A police big data analytics platform: Framework and implications," in *1st International Conference on Data Science in Cyberspace (DSC)*, Changsha, China. IEEE, 2016, pp. 323-328.

[6] Y.-L. Lin, M.-F. Yen, and L.-C. Yu, "Grid-based crime prediction using geographical features," *ISPRS International Journal of Geo-Information*, vol. 7, no. 8, p. 298, 2018.

[7] C. Amrutha, C. Jyotsna, and J. Amudha, "Deep learning approach for suspicious activity detection from surveillance video," in *2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, Bangalore, India. IEEE, 2020, pp. 335-339.

[8] Z. Qin, H. Liu, B. Song, M. Alazab, and P. M. Kumar, "Detecting and preventing criminal activities in shopping malls using massive video surveillance based on deep learning models," *Annals of Operations Research*, pp. 1-18, 2021.

[9] S. Sanober *et al.*, "An enhanced secure deep learning algorithm for fraud detection in wireless communication," *Wireless Communications and Mobile Computing*, vol. 2021, p. Article ID 6079582, 2021.

[10] M. Jullum, A. Løland, R. B. Huseby, G. Ånonsen, and J. Lorentzen, "Detecting money laundering transactions with machine learning," *Journal of Money Laundering Control*, vol. 23, no. 1, pp. 173-186, 2020.

[11] O. Sharif, M. M. Hoque, A. Kayes, R. Nowrozy, and I. H. Sarker, "Detecting suspicious texts using machine learning techniques," *Applied Sciences*, vol. 10, no. 18, p. 6527, 2020.

[12] A. Wibowo and T. Oesman, "The comparative analysis on the accuracy of k-NN, Naive Bayes, and Decision Tree Algorithms in predicting crimes and criminal actions in Sleman Regency," in *Journal of Physics: Conference Series*, 2020, vol. 1450, no. 1, p. 012076: IOP Publishing.

[13] S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker, "Crime prediction using spatio-temporal data," in *Communications in Computer and Information Science*, vol. 1235, N. Chaubey, S. Parikh, and K.

[14] R. Rawat, V. Mahor, S. Chirgaiya, R. N. Shaw, and A. Ghosh, "Analysis of darknet traffic for criminal activities detection using TF-IDF and light gradient boosted machine learning algorithm," in Innovations in Electrical and Electronic Engineering, vol. 756, S. Mekhilef, M. Favorskaya, R. K. Pandey, and R. N. Shaw, Eds.: Springer, Singapore, 2021, pp. 671-681.

[15] R. A. A. Habeeb et al., "Clustering-based real-time anomaly detection—A breakthrough in big data technologies," Transactions on Emerging Telecommunications Technologies, vol. 33, no. 8, p. e3647, 2022.

[16] A. Homayounfar, A. T. Ho, N. Zhu, G. Head, and P. Palmer, "Multi-vehicle convoy analysis based on ANPR data," in 4th International Conference on Imaging for Crime Detection and Prevention (ICDP), London, UK. IEEE, 2011.

[17] Y. Sun, X. Zhou, L. Sun, and S. Chen, "Vehicle Activity Analysis Based on ANPR System," in 12th IEEE International Conference on Embedded and Ubiquitous Computing, Milan, Italy. IEEE, 2014, pp. 89-96.

[18] S. Chen, J. Zhu, Q. Xie, W. Huang, and Y. Huang, "Understanding airline passenger behavior through PNR, SOW and webtrends data analysis," in 1st International Conference on Big Data Computing Service and Applications, Redwood City, USA, IEEE, 2015, pp. 323-328.

[19] B. Chaithra, K. Karthik, D. Ramkishore, and R. Sandeep, "Monitoring traffic signal violations using ANPR and GSM," in International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC), Mysore, India. IEEE, 2017, pp. 341-346.

[20] M. Kumar, M. Hanumanthappa, and T. S. Kumar, "Crime investigation and criminal network analysis using archive call detail records," in 8th International Conference on Advanced Computing (ICoAC), Chennai, India. IEEE, 2017, pp. 46-50.

[21] J. Magnusson and T. Kvernvik, "Subscriber classification within telecom networks utilizing big data technologies and machine learning," in 1st International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications, Beijing, China. ACM 2012, pp. 77-84.

[22] J.-C. Tseng et al., "A successful application of big data storage techniques implemented to criminal investigation for telecom," in 15th Asia-Pacific Network Operations and Management Symposium (APNOMS), Hiroshima, Japan. IEEE, 2013, pp. 1-3.

[23] A. Mottini and R. Acuna-Agost, "Relative label encoding for the prediction of airline passenger nationality," in 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain. IEEE, 2016, pp. 671-676.

[24] H. A. Karacabey, "A model of suspected activity detection and application using big data analytics," Ankara Yıldırım Beyazıt Üniversitesi Fen Bilimleri Enstitüsü, 2019.

[25] S. B. Elagib, A.-H. A. Hashim, and R. Olanrewaju, "CDR analysis using big data technology," in International Conference on Computing, Control, Networking, Electronics and Embedded Systems Engineering (ICCNEEE), Khartoum, Sudan. IEEE, 2015, pp. 467-471.

[26] M. Zhu, C. Liu, J. Wang, X. Wang, and Y. Han, "A service-friendly approach to discover traveling companions based on ANPR data stream," in International Conference on Services Computing (SCC), San Francisco, USA. IEEE, 2016, pp. 171-178.

[27] T. Barker, "American based biker gangs: International organized crime," *American Journal of Criminal Justice*, vol. 36, no. 3, pp. 207-215, 2011.

[28] A. Fraser, R. Ralphs, and H. Smithson, "European youth gang policy in comparative context,*" Children & Society*, vol. 32, no. 2, pp. 156-165, 2018.

[29] A.E. Topcu, Y.I. Alzoubi, E. Elbasi, E. Camalan, "Social Media Zero-Day Attack Detection Using TensorFlow", *Electronic*s, vol. 12, no. 17, p. 3554, 2023.

[30] Y.I. Alzoubi, A.E. Topcu, A.E. Erkaya, "Machine Learning-Based Text Classification Comparison: Turkish Language Context", *Applied Sciences*, vol. 13, no. 16, p. 9428, 2023.

[31] N. Shah, N. Bhagat, and M. Shah, "Crime forecasting: A machine learning and computer vision approach to crime prediction and prevention," Visual Computing for Industry, *Biomedicine, and Art*, vol. 4, no. 1, pp. 1-14, 2021