Spatio-temporal crime predictions by leveraging artificial intelligence for citizens security in smart cities

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 1, 2021, accepted March 19, 2021, date of publication March 23, 2021, date of current version April 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068306

Spatio-Temporal Crime Predictions by Leveraging Artificial Intelligence for Citizens Security in Smart Cities

UMAIR MUNEER BUTT[1], SUKUMAR LETCHMUNAN[1], FADRATUL HAFINAZ HASSAN[1], MUBASHIR ALI[2], ANEES BAQIR[3], TIENG WEI KOH[4], AND HAFIZ HUSNAIN RAZA SHERAZI[5], (Senior Member, IEEE)

[1]School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Malaysia
[2]Department of Management, Information and Production Engineering, University of Bergamo, 24129 Bergamo, Italy
[3]Department of Environmental Sciences, Informatics and Statistics, Ca'Foscari University of Venice, 30123 Venice, Italy
[4]Department of Software Engineering and Information System, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Seri Kembangan 43400, Malaysia
[5]Tyndall National Institute, University College Cork, Cork 021, T12 E8YV Ireland

Corresponding authors: Umair Muneer Butt (umair@student.usm.my), Sukumar Letchmunan (sukumar@usm.my), and Fadratul Hafinaz Hassan (fadratul@usm.my)

This work was supported in part by Universiti Sains Malaysia (USM) through Fundamental Research Grant Scheme (FRGS) under Grant FRGS/1/2020/TK03/USM/02/1, and in part by the School of Computer Sciences USM.

ABSTRACT Smart city infrastructure has a significant impact on improving the quality of humans life. However, a substantial increase in the urban population from the last few years poses challenges related to resource management, safety, and security. To ensure the safety and security in the smart city environment, this paper presents a novel approach by empowering the authorities to better visualize the threats, by identifying and predicting the highly-reported crime zones in the smart city. To this end, it first investigates the *Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN)* to detect the hot-spots that have a higher risk of crime occurrence. Second, for crime prediction, *Seasonal Auto-Regressive Integrated Moving Average (SARIMA)* is exploited in each dense crime region to predict the number of crime incidents in the future with spatial and temporal information. The proposed HDBSCAN and SARIMA based crime prediction model is evaluated on ten years of crime data (2008-2017) for *New York City (NYC)*. The accuracy of the model is measured by considering different time scenarios such as the year-wise, (i.e., for each year), and for the total considered duration of ten years using an 80:20 ratio. The 80% of data was used for training and 20% for testing. The proposed approach outperforms with an average Mean Absolute Error (MAE) of 11.47 as compared to the highest scoring DBSCAN based method with MAE 27.03.

INDEX TERMS Citizen security, smart cities, crime prediction, artificial intelligence, safe city.

I. INTRODUCTION

The smart city's primary objective is to improve its citizens' quality of life by efficient utilization of the city's resources. The unprecedented transformation of urban areas has a significant impact on the socio-economic development of the cities [1]. Because of technological advancements, smart cities infrastructure has been introduced that mainly focuses on the quality of citizen life, better management of urban population issues, and sustainability in every aspect of their life [2], [3].

Smart cities have empowered human life by exploiting technology to address socio-economic challenges such as education, health, transportation, economy, and public safety. However, the fairly increasing population in cities is posing challenges such as resource planning, public safety, and an enormous amount of data generated from sensors, cameras, and tracking devices [4].

The collaboration among researchers, technology developers, government officials, industry, and citizens is paramount

The associate editor coordinating the review of this manuscript and approving it for publication was Lorenzo Ciani.

to present and develop ideas to cope with smart city challenges to achieve smart city goals. One of the crucial challenges is to provide a secure and safe environment [5]. The availability of the enormous amount of data from the past few years has motivated researchers to pursue research in the areas of criminal investigations [6]. The law enforcement agencies must consider the crime trends and patterns for effective policymaking towards better and peaceful communities [7].

The crime is a disorder in behavior, and it is a complex phenomenon of multiple dimensions closely associated with various factors such as spatial, temporal, societal, and ecological. Considering the crime trends and patterns of a locality is critical for someone while deciding to relocate to a new city. Similarly, it enables to avoid traveling to locations, and places with higher safety concerns [8]. Based on historical data, forecasting crime rate has recently got significant attention by the research community. Hence, a number of different methods for discovering a range of aspects related to crime prediction have already been proposed [9].

Numerous data mining and machine learning techniques [10] have been proposed in the literature to analyze and predict the crime rate and criminal incidents such as association rule mining [11], classification [12], and clustering [13]. While constant growth in urbanization has led a number of challenges for the city administrations, crime spikes concerning seasonality are a serious threat to be dealt with utmost care [14]. Figure 1 shows different types of crime spikes over the years in the New York City [15].

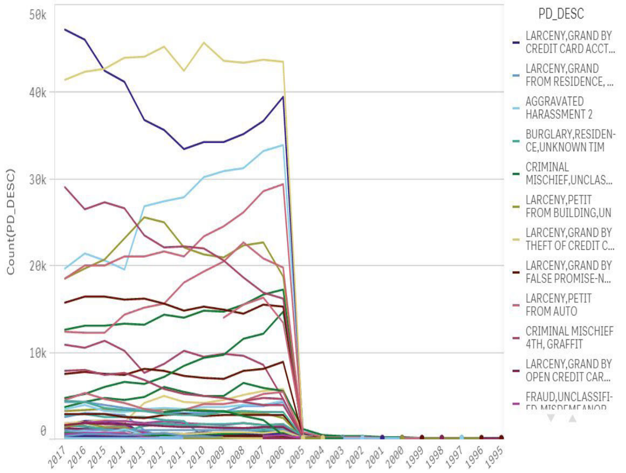


FIGURE 1. **The number of different types of crime spikes over the years in NYC.**

To overcome the crime spiking into different spots, strict yet effective policies need to be enforced. For this purpose, predictive policing is one of the few ways based on statistical prediction techniques to analyze the likelihood of the number of crime occurrence in the near future [16]. The higher crime rates have induced increased complexities over the period and the technological innovations enabled the authorities to analyze and understand crime trends and patterns [17]. Hence, identifying the hotspots with greater crime threats would help the authorities to plan accordingly. In this regard, mapping crime-dense areas have emerged as a sublime analytical method for identifying areas with greater crime risk for efficient and effective allocation and mobilization of police resources [18].

Moreover, the efficient resource utilization is also crucial while computing the exponentially growing data which requires fairly high computation resources to process the data for desired results. Therefore, among other goals, one of the main tasks was to utilize a clustering algorithm demanding fewer computational resources than the baseline study [17] which used Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Specifically, our study aims to contribute towards achieving the following goals.

1) First, given a raw dataset, pre-processing techniques are applied to remove outliers so that crime hotspot detection and prediction approaches can be exploited efficiently.
2) Second, crime hotspots are detected from the processed dataset where criminal events have a density higher than other locations.
3) Finally, the number of crimes in each dense crime region at a given timestamp is predicted for the authorities to plan preventive actions.

The rest of the paper is organized as follows: Section II discusses previously available literature on Spatio-temporal crime hotspot detection and prediction. Section III presents the proposed method with algorithmic details. Section IV discusses the dataset being used for the experimental evaluations. Section IV-C and IV-D describe the proposed crime hotspot detection and prediction algorithms' parameter level detail. Section V presents a discussion on the comparative analysis of the proposed technique with state-of-the-art. Finally, section VI concludes the paper by highlighting various aspects of the proposed approach and possible future directions.

II. RELATED WORK

This section focuses on the role of predictive policing in the identification and prediction of the likely crime locations by exploiting different "state-of-the-art" methods [16]. Specifically, crime hotspot detection and prediction approaches are discussed that use spatial and temporal information for predictive analysis. Several attempts have been made for crime predictions due to their importance for law enforcement agencies [19]. However, the advancement in geographical information systems and the inclusion of spatial and temporal information in crime datasets make the predictions more efficient and reliable [6], [7]. This section focuses on the most significant research carried out in the domain of crime hotspot detection and prediction.

A. SPATIO-TEMPORAL CRIME HOT-SPOT DETECTION

Crime rates can vary according to geographic location; some locations can be identified as low and high-risk areas.

TABLE 1. Spatio-Temporal Crime Hot-Spot Detection Techniques.

Reference	Techniques	Data	Findings
[32]	Hot Spots prediction model based on mixed spatial-temporal characteristics	Data of Main city zone of Nanchang ranging from 2014 to 2015	Optimal performance can be achieved by the prediction model if crime statistics are conducted on weekly basis
[33]	Spatio-Temporal Neural Network	Call for service data provided by Portland, Oregon Police Bureau for March 2012 through the end of December 2016	81.50% Accuracy
[34]	Kernel Density Estimation (KDE)	Crimes occurred in Manila, Philippines from the year 2012 to 2016	Criminal activities in Manila are at peak around 8:00 PM to 4:00 AM
[35]	spatio-temporal kernel density estimation (STKDE)	Data of residential burglaries in Baton Rouge, Louisiana in 2011	Southwest area of Baton Rogue is identified as the high-risk area
[31]	Spatio-temporal Ordinary Kriging	Crime dataset of Philadelphia from January 2011 to December 2016	90.52% Sensitivity
[17]	DBSCAN	Crimes Dataset of New York city and Chicago	Crime dense regions are discovered

The trends of crimes can change with seasonal patterns and time of the year as well. Performing analysis concerning their spatial properties has grown in the last decade. In this regard, analyzing crime hot-spots is an important and popular approach [20]–[23]. It is evident from the studies that some locations have a greater perception of crimes than the actual risk level [24].

Researches have been dedicated enough in the past years to provide suitable measures that could lead to the prevention and hence reduction of crimes. In this regard, Adelson *et al.* [25] proposed MLP, KNN, and Random forest based method for crime hotspot detection for achieving the smart city goal of public safety in the Natal city of Brazil. Sankar and Gopi [26] exploited deep learning for improving accuracy and time complexity in crime hotspot detection. The authors collected crime data of Los Angeles and California from police records from 2010 to 2018 to assess the algorithm. Experimental results show that the deep learning based approach can significantly improve the crime hotspot detection accuracy.

Shino and Narushige [27] proposed a network-based approach to detect crime hotspots at street level. The algorithm was evaluated on robbery, burglary, and drug data of Chicago. Empirical analysis shows the effectiveness of the method in capturing detailed information on different types of crimes. Cheng *et al.* [28] applied FP-Growth algorithm to discover abnormalities in purchasing, buying, and traveling behavior of criminals. Later, the DBSCAN algorithm is applied to detect rampant crime regions using generated associated rules. The algorithm showed promising results with an accuracy of 73.9 %. Ravi and Bharti [29] present a detailed analysis on big data approaches for crime hotspot analysis and proposed an algorithm for crime prediction. They used the Naive Bayes classifier for crime prediction and criminal identification on Cheltenham, United Kingdom crime data collected from police data portal. The algorithm can provide significant trends and patterns of crimes to police forces.

Yiqun and Shashi [30] highlighted theoretical limitations of existing crime hotspot detection approaches and proposed a robust Non-deterministic Normalization (NN) Scan Statistic algorithm for hotspot detection. It was achieved by presenting a novel Dynamic Linear Approximation algorithm which can significantly improve the computational complexity problem. The enhanced NN-Scan evaluated on crime data of Minneapolis, USA, collected from City police. Experimental results showed the effectiveness of the algorithm as compared to state-of-the-art techniques.

Reference [31] proposed a Spatio-temporal ordinary kriging model that used minimal features like location of crime, its time and type, and their correlation to predict future crime locations as well, which helped improve the accuracy. To apply this model, a crime dataset from Philadelphia from January 2011 to December 2016 was used. The proposed method achieved 90.52% sensitivity and 88.63% specificity. The summary of the techniques previously proposed by various authors on hot-spot detection techniques is mentioned in table 1.

B. SPATIO-TEMPORAL CRIME PREDICTION

To prevent crimes from happening in the future, several methodologies have been proposed in the last few years to provide efficient and compelling insights to law enforcement agencies' for better resource allocation [36]–[38]. Jason and Anthony [39] present a correlative analysis of the disastrous pandemics such as Spanish flu and Covid-19 and its effect on a country's economic growth. They find a strong correlation between unemployment and crime. The authors proposed an Auto Regressive Integrated Moving Averages (ARIMA) based crime forecasting model to predict the next six months of crime in Queensland, Australia. They collected violent crime data from the Queensland police in March 2020. The violent data consists of assaults, domestic violence, and sexual offense as major crime types. Experimental results showed a 95% confidence value in determining possible crimes across Queensland using ARIMA based approach.

Sohrab *et al.* [44] proposed a supervised learning-based crime prediction algorithm using spatial and temporal information of the crime. Decision tree and K-Nearest neighbor algorithms have been used to train the crime prediction model. Moreover, Random Forest and AdaBoost algorithms

TABLE 2. Spatio-Temporal Crime Prediction Techniques.

Reference	Techniques	Data	Findings
[40]	Probabilistic Model	Crime records ranging from June 2013 to May 2014 in Dhaka, Bangladesh	79.24% sensitivity
[19]	Cluster - Confidence - Rate - Boosting (CCRBoost)	Ranging from January 2006 to December 2009, From a Police department in a city from northeastern, US	80% Accuracy
[41]	GA-BP neural network model	Crimes that occurred from 2008 to 2012, at city in South China	The accuracy results is based on the accuracy of input data
[34]	Naive Bayes	Gun shooting crimes incurred from the year 2012 to 2016 in Manila	77.78% Accuracy
[42]	Autoregressive Integrated Moving Average model (ARIMA)	Daily police data provided by the Public Security Bureau (PSB) of a city in China	Prediction results meet the expected requirements and are more accurate
[43]	Random Forest, Neural Network, SVM, Logistic Regression Model	The crime event records of Queensland, Australia from 01/2013 to 09/2013 and New York City from March, 2012 to February, 2013	With the inclusion of dynamic features across diverse types of criminal events, crime prediction performance can be significantly improved
[17]	Seasonal ARIMA	Crimes Dataset of New York city and Chicago	Average MAE is 27.03 for dense crime regions

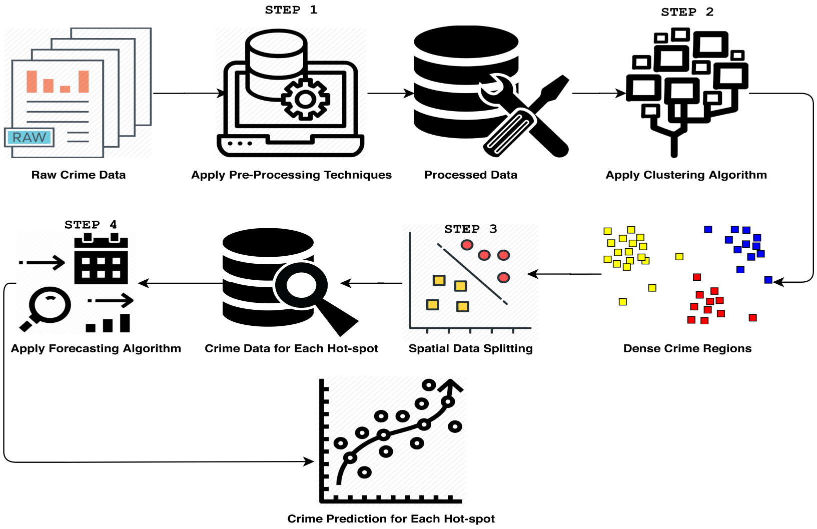

STEP 1
STEP 2
RAW
Raw Crime Data
Apply Pre-Processing Techniques
Processed Data
Apply Clustering Algorithm
STEP 4
STEP 3
Apply Forecasting Algorithm
Crime Data for Each Hot-spot
Spatial Data Splitting
Dense Crime Regions
Crime Prediction for Each Hot-spot

FIGURE 2. Steps of Spatio-Temporal Crime Prediction.

are used to improve crime prediction accuracy. The algorithm was evaluated on 12 years of crime data of San Francisco and outperformed compared to state-of-the-art techniques.

Li *et al.* [42] implemented Auto-Regressive Integrated Moving Average (ARIMA) model on daily police data provided by the Public Security Bureau (PSB) of a city in China. The prediction results met the expected requirements and were more accurate. Reference [43] used Random Forests, Neural Network, SVM, Logistic Regression Model on the crime events recorded in Queensland, Australia from 01/2013 to 09/2013 and New York City from March 2012 to February 2013. It was concluded that with the inclusion of dynamic features across diverse criminal events, crime prediction performance could be significantly improved. The summary of the techniques previously proposed by various authors on hot-spot detection techniques is mentioned in table 2.

III. PROPOSED CRIME PREDICTION METHODOLOGY

Studies show that incidents of criminal events are not equally distributed within a city [17]. Because of the unequal distribution of crimes, it can be considered as a location-oriented feature as some places can exhibit a greater risk of crime to be committed than others [45]. Crime rates can change with respect to the geographic location of the area. Hence, resource allocation and counter-crime are challenging for law enforcement agencies when high-risk regions are unknown. Therefore, an accurate model needs that can efficiently detect crime hot-spots and reliably forecast crime time and location.

In this section, we have explained the proposed crime hot-spot detection and prediction methodology and the number of steps required to perform to be able to detect hot-spots and forecast the number of crimes. The proposed model is based on the number of steps depicted in Figure 2.

A. THE ALGORITHM

As the work to be done in this study is based on two tasks, (1) Detect the crime-dense regions (2) Forecast the number of crimes in each crime-dense region. Therefore, to perform these two tasks, the proposed algorithm is a combination of two different algorithms (first to detect regions where crime density is greater, second to predict the number of crimes in each dense region). The pseudo-code of the *"Spatio-Temporal Hot-Spot Detection and Crime Prediction algorithm (STHDCP)"* is depicted in Algorithm 1.

Algorithm 1 Spatio-Temporal Hot-Spot Detection and Crime Prediction

```
Input: Raw Crime Dataset (RCD)
Output: $\mathcal{HS} = \{HS_1, \dots, HS_K\}$: a set of K hot-spots;
$\mathcal{CP} = \{CP_1, \dots, CP_K\}$: a set of K crime predictions for
  each hot-spot;
Method:
STEP 1: Execute DataPreProcessing(RCD) method to
handle missing values and remove outliers from the
targeted variables to get Processed Data (PD);
STEP 2: Execute DiscoverHotSpots(PD) method to get
$\mathcal{HS} = \{HS_1, \dots, HS_K\}$;
STEP 3: Execute SpatialDataSplitting(PD,HS) to get
$\mathcal{SDS} = \{\mathrm{SDS}\frac{l}{1}, \dots, \mathrm{SDS}\frac{l}{K}\}$: a set of K crime data for
each hot-spot with spatial information;
STEP 4:
while (for each k = 1, ..., K) do
        $CP_i \leftarrow \mathrm{DiscoverCrimePredictor}(\mathrm{SDS}\frac{l}{i})$;
        $\mathcal{CP} \leftarrow \mathcal{CP} \cup CP_i$;
end
return $\{\mathcal{HS}, \mathcal{CP}\}$;
```

As discussed in the proposed model, crime hotspots need to be detected first before performing a time series analysis. Therefore, hotspots are detected using cluster analysis, and then time series prediction is performed for each cluster to predict crime. Figure 2 depicts the processes of the proposed model of this study. For the *"detection of dense crime regions"* from the dataset, the aim is to discover the areas where the occurrences or frequency of crime are greater than other adjacent areas. They are to be discovered automatically without a-priority defining division in areas. This task can be modeled as a geospatial clustering instance using a clustering algorithm. This study exploited HDBSCAN that processes spatial data after it has been filtered for its temporal characteristics. The final output of these processes is K number of clusters which corresponds to a dense crime region.

As this study uses HDBSCAN, which doesn't need apriori information regarding the number of clusters to be detected, it is done automatically depending on the data points' density. The algorithm further consists of the steps required to perform *"spatial data splitting"* of the original crime data, depending on the number of clusters found using the clustering model in the previous steps. In other words, data points which points to crime events are occurred belonging to the dataset, allocated to the i^{th} cluster are converted in a time series and gathered in the i^{th} output dataset, for i = 1, ..., K. This step aims to allocate the details of crimes belonging to each cluster and partitioning them accordingly. This step's output is K different time series datasets, each containing the time series of crimes in its associated dense region. The next step is aimed at *"extracting a specific crime prediction model for each crime dense region"*, analyzing the data of crime split in the previous step.

B. DETECTION OF CRIME DENSE REGION

The *"DiscoverHotSpots()"* method (Algorithm 1) performs clustering with respect to its spatial factor and each discovered cluster is a dense crime region. This step in this study is performed by applying HDBSCAN [46], an extended version of DBSCAN developed by [47]. It is extended by converting DBSCAN into a hierarchical clustering algorithm by extracting a flat clustering based on clusters' stability.

Both algorithms have the minimum number of samples parameter, which is the neighbor threshold for a record to become a core point. DBSCAN has the parameter epsilon, the radius those neighbors have to be in for the core to form. HDBSCAN has the parameter minimum cluster size, which is how big a cluster needs to include. This value is more intuitive than epsilon because you probably know how big your clusters need to be to make actionable decisions on them. The absence of epsilon value in HDBSCAN gives it the freedom to discover the number of diverse densities clusters and be more robust to parameter selection than DBSCAN, which falls short of varying densities.

Considering the advantages that come with hierarchical clustering, another clustering algorithm was considered i.e. Hierarchical Agglomerative Clustering (HAC), but it was not suitable for this study because it has a time complexity of $O(n^3)$ and space complexity of $O(n^2)$ [48]–[50]. As evaluated by [7], Hierarchical Agglomerative Clustering required a lot more time and memory as compared to HDBSCAN to perform clustering. Besides, those data objects which are grouped incorrectly at early stages cannot be reallocated to other groups at later stages. Therefore, the dendrogram remains the same and cannot be changed.

Similarly, partition-based clustering algorithms such as K-Means have the notable disadvantage of mentioning the number of clusters beforehand [51]. That can be a crucial problem in large scale data where you do not have the idea of how many clusters are going to be formed or needed. Besides, K-means gives good performance when data is distributed as isotopic clusters and is prone to noise and outliers [52].

In contrast, HDBSCAN possesses the advantages of both hierarchical and density based algorithms. Catlett *et al.* [17] highlighted the algorithm's effectiveness in excluding noise and outliers from the data and supremacy compared to state-of-the-art methods.

C. EXTRACTING SPATIO-TEMPORAL CRIME PREDICTORS

Given a specific dense crime region, the *DiscoverCrimePredictor()* method which is mentioned in Algorithm 1 discovers a forecasting model to predict the number of crimes that will happen in its specific area. In our implementation, this has been performed by the *"Seasonal AutoRegressive Integrated Moving Average"* model (Seasonal ARIMA, or SARIMA) [53]. Which can be defined as a combination of auto-regression, moving average and difference modeling along with seasonality. Briefly, having the time series $\{y_t: t = 1 \ldots n\}$, where y_t is the the value of the time series at the timestamp t, an *ARIMA* p, d, q model can be written as mentioned in (1) below.

$$y_t^{(d)} = c + \phi_1 y_{t-1}^{(d)} + \ldots + \phi_p y_{t-p}^{(d)} + \theta_1 e_{t-1} + \ldots + \theta_q e_{t-q} + e_t \quad (1)$$

where c is a correcting factor, $\phi_1, \ldots, \phi_p$ are the regression coefficients of the auto-regressive part, $\theta_1, \ldots, \theta_q$ are the regression coefficient of the moving average part, $y_{t-1}, \ldots, y_{t-p}$, $e_{t-1}, \ldots, y_{t-q}$ are lagged values of y_t and lagged errors (p+q predictors), and e_t is white noise and takes into account the forecast error. In this study, ***Seasonal ARIMA*** model is exploited, which is an extension of classical **ARIMA**.

Considering the seasonal spikes in the data, i.e., the number of crimes can grow significantly at a certain time of the year. Therefore, to cope with the seasonal element, **Seasonal ARIMA** model is built by including seasonal terms in **ARIMA** model. In the final formula of SARIMA, the additional seasonal terms are multiplied with the non-seasonal terms. A Seasonal ARIMA model is referred as *ARIMA(p,d,q)(P,D,Q)m*, where *m* is a periodicity factor, *(p,d,q)* and *(P,D,Q)* are the orders of the autoregressive, differencing and the moving average part for the non-seasonal and seasonal model, respectively [53].

The problem with ARIMA is that if data is non-stationary and possesses clear trends, it requires a lot of differencing to make it stationary. That might be the reason that the seasonal spikes get ignored, and it may give undesired results. Hence, to deal with the seasonal element of the data, ARIMA is not a suitable approach. Moreover, it is evident from the results extracted by [54], that to perform time series forecasting, ARIMA performs better than state-of-the-art Machine Learning algorithms like SVM, RNN, KNN, and LSTM for time-series forecasting. And because the data has seasonal dips in it, SARIMA is applied to the dataset.

IV. EXPERIMENTAL EVALUATION

For experimental evaluation dataset acquired from [15] is used to check the proposed model's efficiency and effectiveness. The details of the dataset are discussed in section .

A. EXPERIMENTAL DATASET

The data that we used to train the models and perform the experimental evaluation for New York City is housed on [15], a publicly available resource managed by the Mayor's Office of Data Analytics (MODA) and the Department of Information Technology and Telecommunications (DoITT). The focus of this study are all five boroughs of NYC i.e. *(1)* The Bronx *(2)* Brooklyn *(3)* Manhattan *(4)* Queens *(5)* Staten Island.

These regions are one of the most densely urban areas globally, growing in population, business, and mobility patterns. The total area of these Boroughs is 783.83 KM^2. The details of land area and density of the population are mentioned in table 3 extracted from [55].

TABLE 3. New York City's Five Boroughs.

Jurisdiction	Land Area		Density	
	Square Miles	Square KM	Persons/Sq. Mi	Persons/km^2
The Bronx	42.10	109.04	34,653	13,231
Brooklyn	70.82	183.42	37,137	14,649
Manhattan	22.83	59.13	72,033	27,826
Queens	108.53	281.09	21,460	8,354
Staten Island	58.37	151.18	8,112	3,132
Total	**302.64**	**783.83**	**28,188**	**10,947**

As per the information extracted from [55], Queens is the borough with the largest Land area with 281 square KM, whereas Manhattan is the most densely populated borough with 72,033 persons per square miles. The dataset contains the information regarding all crimes that were reported in these boroughs starting from year 01-01-2008 to 31-12-2017.

The dataset[1] contains 4,952,699 rows representing the number of crimes that occurred over the years in various locations. In comparison, the average number of crimes recorded in a week are 10,318. The size of the data size is approximately 2GB. Since Manhattan is the most densely populated borough among all five, it can be assumed and proved from Figure 3 that it is also the densest region for crimes. The density of the crimes in all five boroughs is shown in Figure 3.

The density map in the Figure 3 indicates the crime density of regions, where some parts have a very low density of crimes. Simultaneously, some areas possess greater density because of the greater number of crime occurrences in the past. Starting from blue and going towards red, the colors depict the density of crime from low to highest dense regions where the number of crimes is greater than the other regions. Red represents the densest regions and can be considered for efficient and effective police resource allocation region to control criminal activities there. The crime frequency distribution per quarter, month, and week for each year are shown

[1] https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i/data

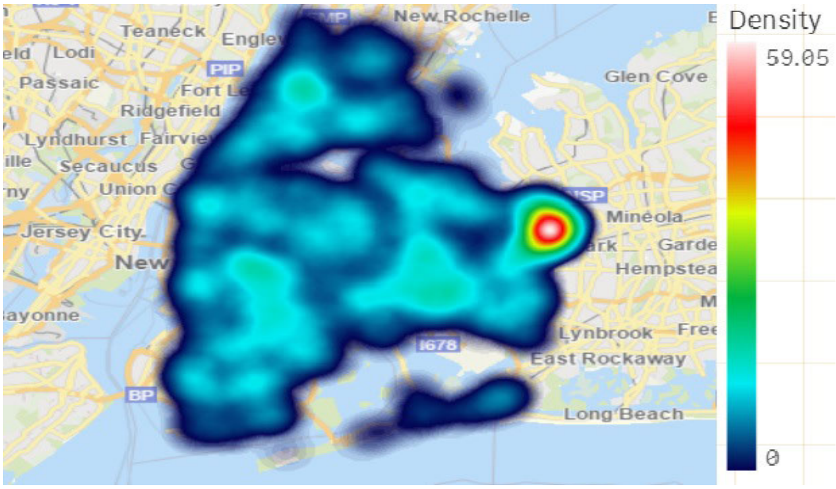


FIGURE 3. Crime Density in the Boroughs of NYC.

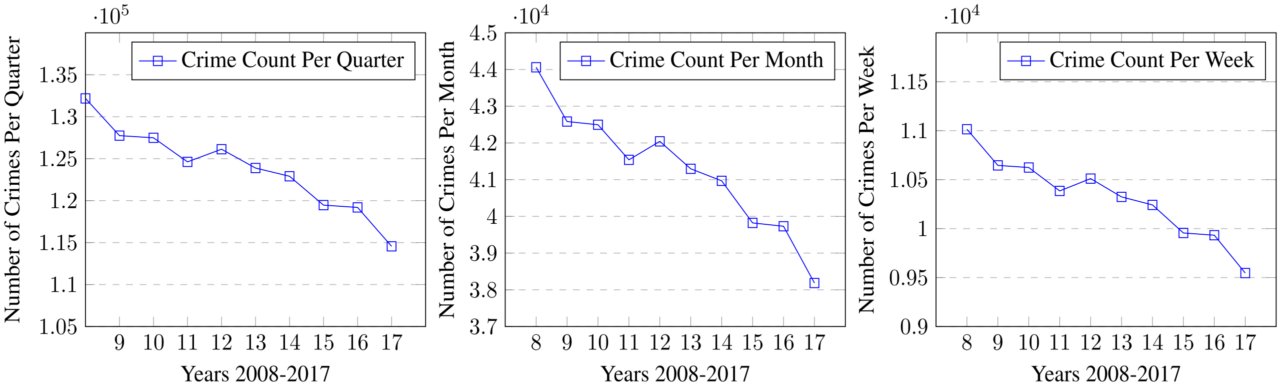


FIGURE 4. Number of Crimes Distribution by Quarter, Month and Week.

in Figure 4 to understand the trends of crimes every year better.

B. PRE-PROCESSING OF THE DATASET

Pre-processing is a cleaning data method to remove noise, handle missing values and deal with outliers, i.e., the values that don't comply with the normal distribution of values. These tasks are performed on the data to make it ready for further processing to get accurate and better results. Data must be cleaned before performing additional calculations. In this study, we have performed the following steps of pre-processing steps to our data to make it ready for further processing. (a) Data Formatting (b) Removing the outliers (c) Attribute Selection (d) Handling Missing Values. This study's main focus was to use the attributes i.e. CMPLNT_FR_DT (Date at which crime was reported), Latitude, and Longitude to detect crime dense areas and perform forecasting in those areas.

The Date is ranged from 01-01-2008 to 31-12-2017. Latitude and Longitude coordinates for NYC are 40.730610, −74.4. First of all, the datatype of CMPLNT_FR_DT was a string in the dataset, which had to be converted into ***Date*** to perform time-series calculations. From ***Pandas*** library of Python, ***"to_datetime"*** was used to convert the data from string to *Date* data type. Afterward, the outliers in Latitude and Longitude attributes were detected and removed. For example, the Latitude values $>= 41.0$ and Longitude values $<= -74.5$ were removed to comply with the normal distribution of values, i.e., representing locations of NYC. Figure 5 depicts the distribution of data points as the Latitudes and Longitudes ***with*** outliers.

Figure 5 depicts the uneven distribution of coordinates because of the outliers. Considering the coordinates' actual range, the data was sorted, and the outliers were removed, which did not belong to the mentioned range. Figure 6 depicts the distribution of data points as the coordinates ***without*** outliers.

To perform clustering, there must not be any missing values in the data. Therefore, to handle the missing values in Latitude and Longitude, ***FillByMean*** method was used. The method computes the mean of all values and fills the missing values with it. Moreover, the mean will remain in the normalized range that was done during the removal of outliers.

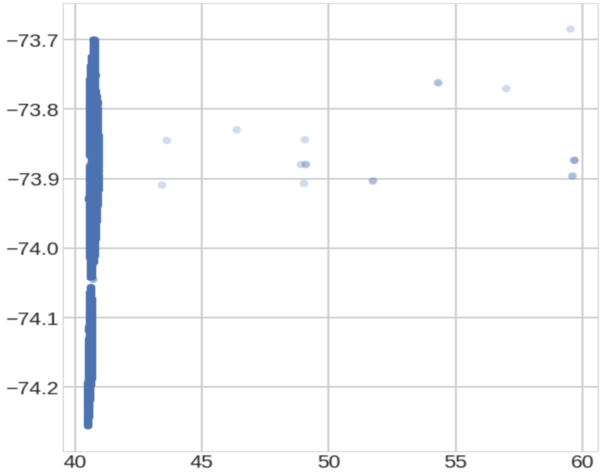


FIGURE 5. Latitude and Longitude ***with*** Outliers.

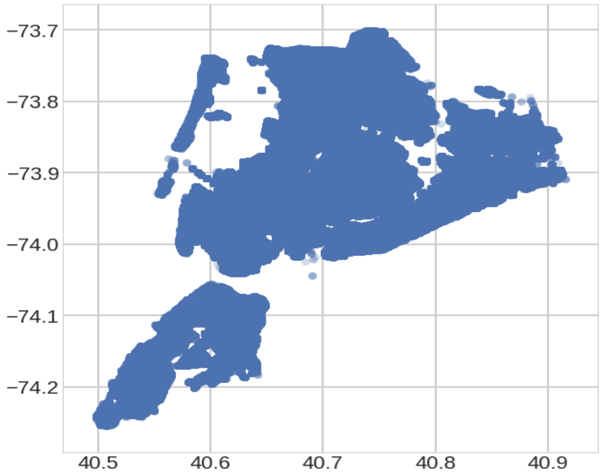


FIGURE 6. Latitude and Longitude ***without*** Outliers.

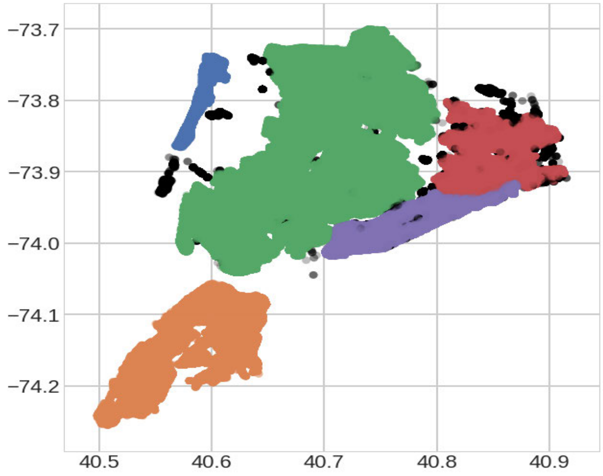


FIGURE 7. Crime Dense Regions in NYC Discovered using Euclidean Distance.

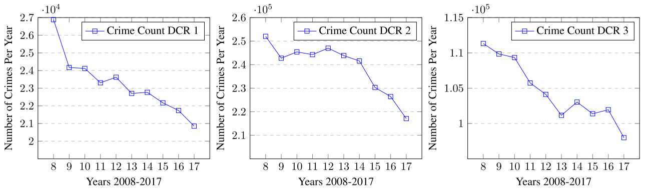


FIGURE 8. Crimes Frequency Distribution for DCR1, DCR2, and DCR3.

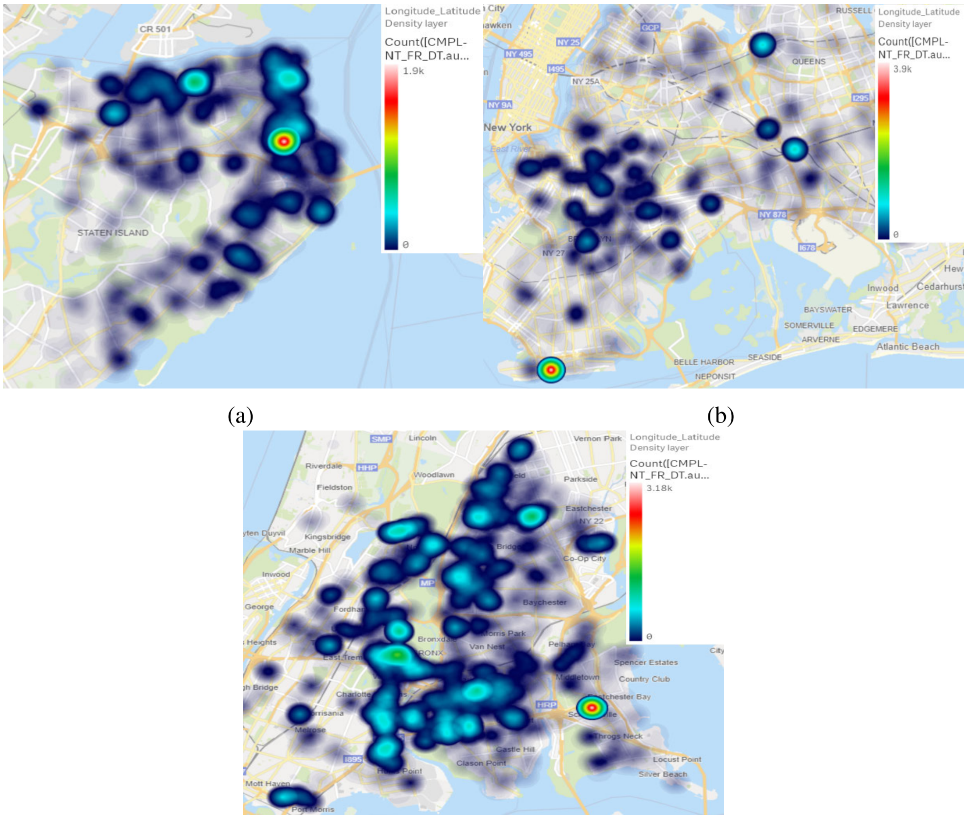
C. DETECTING CRIME HOT-SPOTS FROM DATASET

As discussed in III-B, the prior need to mention the number of clusters before forming clusters can be a major problem, especially when you don't have prior knowledge of how many clusters you should be looking for. Hence, there's a strong need to calculate K number of clusters depending on the number of crimes or density of crimes in an area. This study used HDBSCAN, which automatically forms the number of clusters using the density of data points. The parameters which were used in HDBSCAN are **min_cluster_size** which is an intuitive parameter to select. It is the smallest size grouping that is wished to be considered to form a cluster. This study uses 20000 as **min_cluster_size**. **min_samples** is another parameter that provides a measure of how conservative you want clustering to be. The larger the value, the more conservative the clustering is. The value of **min_samples** was 50 in this study and **distance metric** was *euclidean*.

Crime dense regions of years 2008 to 2017, discovered using HDBSCAN algorithm are shown in Figure 7. The algorithm discovers five different crime regions visible through different colors. Black color represents noise that will be ignored.

The following information in Figure 8 was extracted from the top three dense regions, which shows the crime frequency distribution for each year. The Figure shows that crime frequency is gradually declining towards the end of 2017. Furthermore, it can be understood that DCR 3 is the densest region as the total number of crimes reported in this region were 2,390,751. In contrast, the total number of crimes reported in DCR1 and DCR2 were 232,328 and 1,045,887, respectively.

The information in Figure 8 shows the number of crimes reported in 2008-2017. However, after detecting the dense crime regions as shown in Figure 7, their density is shown in figures 9a, 9b, and 9c for DCR1, DCR2, and DCR3 for each year respectively. For each year, the density of all dense crime regions is highlighted in the figures. The level of density is represented in the bar using the count of crime for each year. The top three dense crime regions are shown in the figures with the location and density of crimes in those regions based on the clusters formed. One of this paper's main contributions is to detect those regions; hence, patrolling and other efficient strategies can be developed based on this information, leading to a significant drop in the number of crimes. It is to be noted that, as previously stated, not all locations possess the same amount of danger or number of crimes; it is different for each location.


Longitude_Latitude Density layer
Count([CMPL-NT_FR_DT.au...
1.9k
0
STATEN ISLAND
CR 501
Longitude_Latitude Density layer
Count([CMPL-NT_FR_DT.au...
3.9k
0
New York
QUEENS
Longitude_Latitude Density layer
Count([CMPL-NT_FR_DT.au...
3.18k
0

(a) (b)

(c)

FIGURE 9. Crime Density of DCR1, DCR2, and DCR3 respectively.

It is evident from the figures that DCR3 is the densest crime region among those three.

D. TRAINING AND EVALUATING THE REGRESSIVE CRIME MODELS

Among the number of dense crime regions extracted in section IV-C, the top three dense crime regions are selected to predict the number of crimes in those regions. Crime regressive models have been extracted from 2008 t0 2017 for all years for all boroughs in the dataset.

The evaluation of the regressive functions is performed using the 80:20 ratio for New York City's dataset. 80% data is used for training, and 20% data used for testing of the regressive functions for all years mentioned in the table. For *DCR1, DCR2, and DCR3* for years 2008-2017, results were extracted from the data set which are shown in the Figure 10a, 10b, 10c for for *DCR1, DCR2 and DCR3* respectively. The figures shows the forecasting values among the actual values.

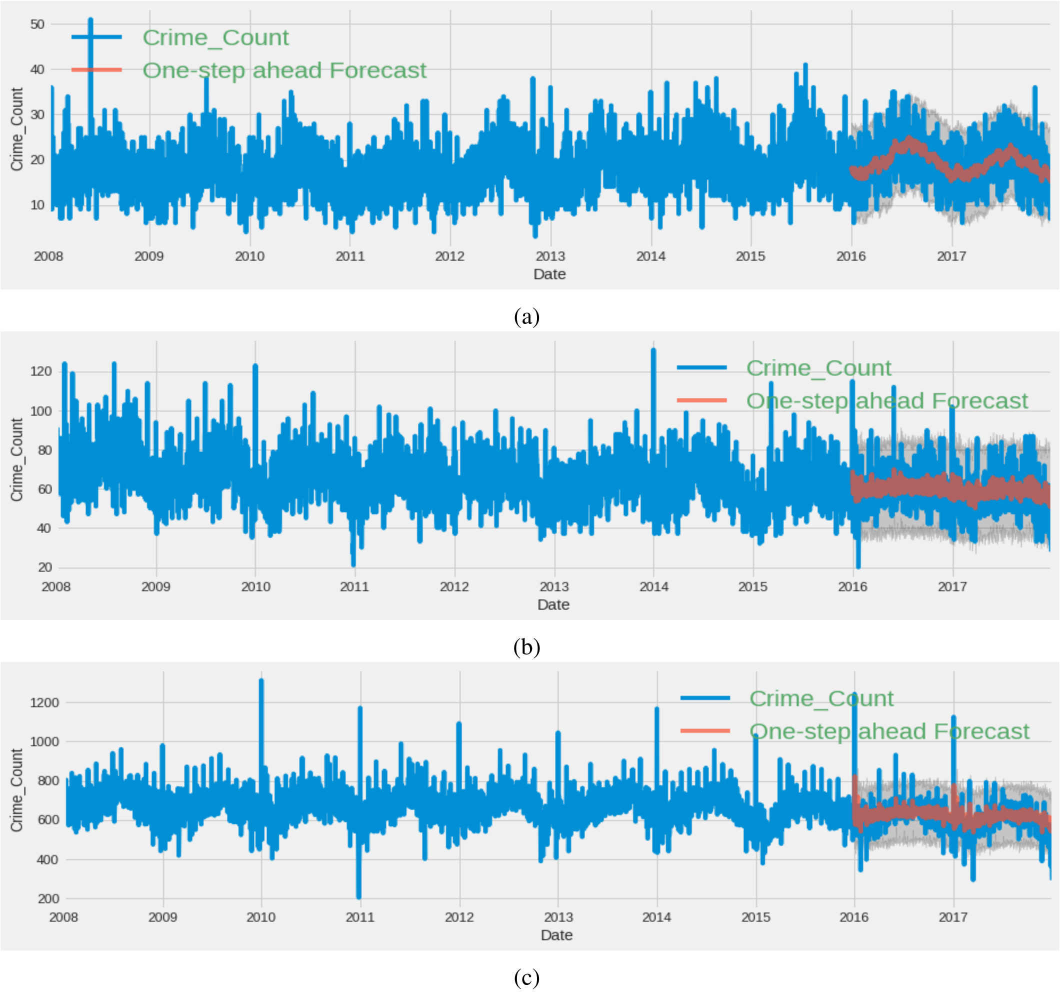
The blue line represents the actual number of crimes each year, whereas the orange line in front of the blue lines represents the predicted number of crimes.

To better understand the forecasting results, *plot diagnostics* are applied to the model used for all three dense crime regions. Four critical parameters in the diagnostic define the quality of results.

1) **Standard Residuals:** There are no obvious patterns in the residuals
2) **Histogram Plus KDE Estimate:**The KDE curve should be very similar to the normal distribution
3) **Normal Q-Q:** Most of the data points should lie on the straight line
4) **Correlogram:** 95% of correlations for lag greater than one should not be significant

The diagnostics of the results depicted in Figure 10a are shown in Figure 11.

As per the four parameters, the diagnostic results are shown in the image 11. It can be concluded from the results in figures for years 2008-2017, *DCR1* is the region that showed better forecasting performance, and the difference between predicted and actual values is minimum compared to other crime dense regions. Whereas the KDE curve is very similar to the normal distribution, most of the data points are lying on the straight line, and correlations for lag greater than are not significant. To evaluate the performance of forecasting on the dataset, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Mean Error (ME), and Root Mean Square Error (RMSE) have been used. The error values are calculated by comparing the predicted values with the actual values using testing and training data with an 80:20 ratio, i.e., 80% training and 20% testing data. The results of applying the proposed model on the dataset are shown in Figure 12.



(a)

(b)

(c)

FIGURE 10. Crime Forecasting for DCR1, DCR2, DCR3 respectively.

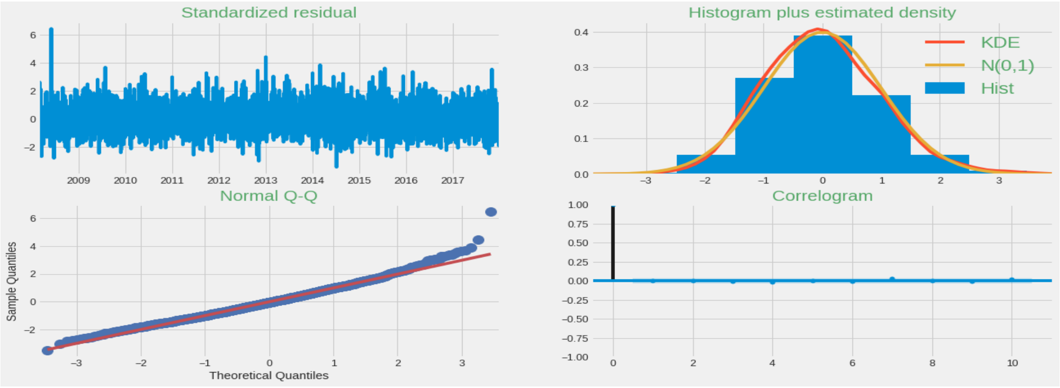
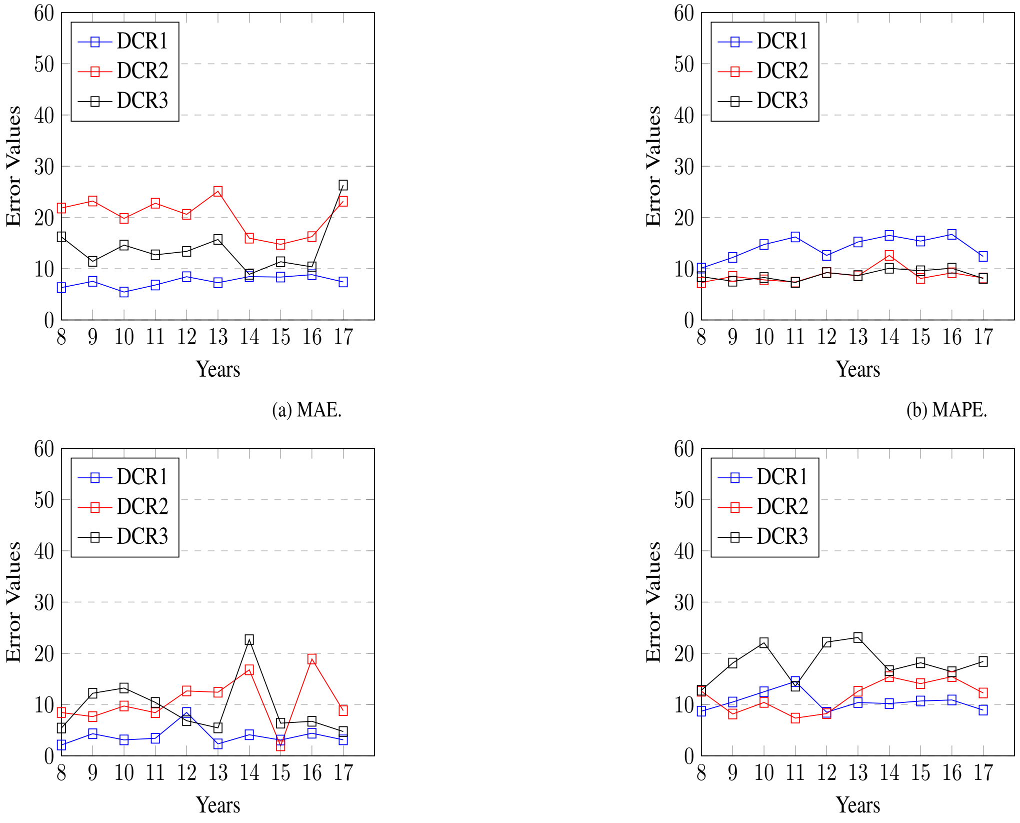


FIGURE 11. Results' Diagnostics of DCR1 Years 08-17.

It can be understood from the figures that DCR1 secured better results than other regions, whereas DCR2 and DCR3 performed less better than DCR1. It can be because DCR1 is the smallest with 232,328 crimes count than the other three regions as mentioned in IV-C. Therefore, it can be understood that areas, where the number of crimes is lower than others, can better forecast results in terms of error values used in this study to evaluate the results.



(a) MAE.

(b) MAPE.

(c) ME.

(d) RMSE.

FIGURE 12. Forecasting Error Values.

V. COMPARATIVE ANALYSIS

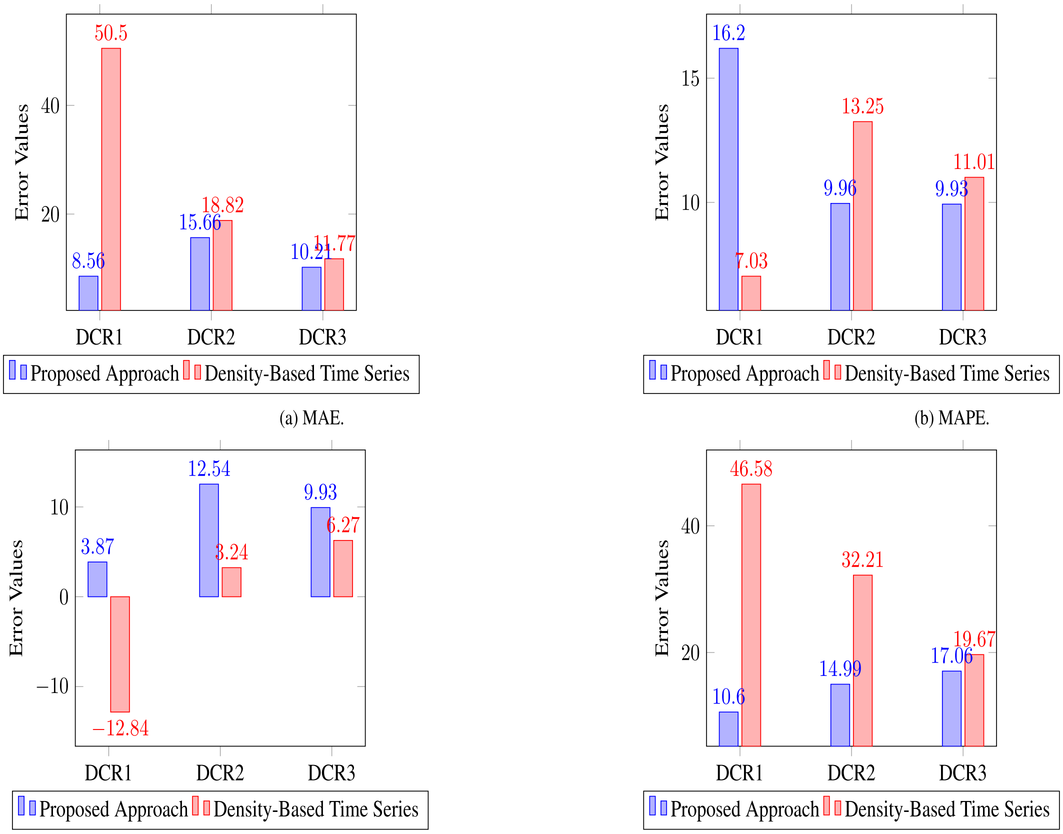
To make our results more accurate and authentic, a comparative analysis is performed in this study with the results achieved by [17] by using DBSCAN for clustering. The comparison of the results of both studies is shown in Figure 13. The models which were used as an input for regressive crime models were different in both studies. Moreover, their forecasting was based on weekly trends; hence the value of *m* was 52. While in this study, as evident from Figure 4, the trend of crime was not non-stationary weekly, whereas an apparent decline in the number of crimes can be seen monthly compared to weekly and quarterly distribution. Therefore, the value of *m* was 12, representing the twelve months in a year.

Furthermore, the results achieved by [17] represents the results of the year 2014-2016 only, whereas this study represents a detailed analysis of the years 2008 to 2017. In both studies, the results are depicted for each crime-dense region for every year. However, for comparative analysis for both studies, the results are shown for common years only, and for all four error measuring parameters, their average is used to compare the results.

It can be concluded from the results that the average performance of the model used in this study extracted better results than extracted by [17]. It is evident from the experimental evaluation that the proposed method outperformed compared to the comparative study in terms of three error evaluation metrics i.e. MAE, MAPE, and RMSE. Only the ME score for *DCR2 and DCR3* of comparative study is better than this study's result. This study extracted forecasting results for the years 2008-2017 for the top three crime-dense regions, which is more than the competitor's study.

Moreover, the results of this study are compared for extraction of crime predictors. Specifically, a comparative analysis of the results achieved using SARIMA in this study is performed against the result extracted by classical regression algorithms such as Random Forest [56], REPTree [57], and ZeroR [58] which is shown in Figure 14.

To compare the results of this study's and the state-of-the-art approaches, predicting the number of crimes in the top three crime-dense regions was evaluated on the dataset used in this study. For each algorithm, results were extracted using suitable and accurate input parameters to get the algorithm's best results. To compare the performance, error evaluation metric MAE is used for two year ahead crime predictions using data of 8 years for training. Figure 14 summarizes the comparison' results, and we can see that SARIMA results are generally better than others. The comparison was performed on the same dataset and using the same timeline. Therefore, the window of comparison is the same. These results confirm



(a) MAE.

(b) MAPE.

(c) ME.

(d) MAE.

FIGURE 13. Comparison of Forecasting Error Values.

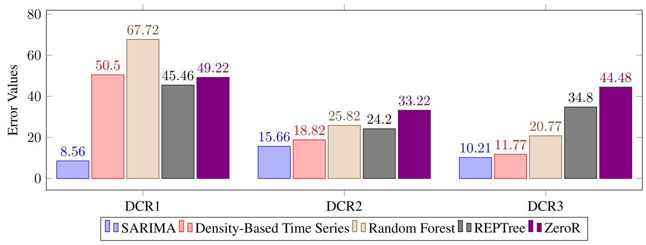


FIGURE 14. Comparison of Forecasting Error Values.

the appropriateness of the autoregressive model and its good performance in the crime prediction domain.

Hence, it can be concluded from the results depicted in the tables and figures above that the proposed model used in this study achieved better results than other studies mentioned in the literature.

VI. CONCLUSION

One of the crucial challenges of smart city infrastructure is to provide a reliable and secure environment that is addressed by detecting crime hot-spots and predicting the number of crimes in those regions. This information can help concerned stakeholders to offer a safe environment for the citizen in a smart city. Because of the constant growth in smart cities' data, managing and utilizing computational resources can be challenging. The paper proposed a cost-efficient and practical approach to achieve the aforementioned goals. The proposed system is evaluated on a dataset containing ten years of crime reports. The experimental results show that the proposed system outperformed compared to state-of-the-art systems with an average Mean Absolute Error (MAE) of 11.47.

Furthermore, we aim to improve the proposed model by exploiting transfer learning as future work. In this approach, the knowledge of an already learned crime prediction model is utilized to solve related crime regions that can improve both costs and learning performance. The clustering ensemble can also be used in the future for more accurate and robust crime detection and prediction model.

REFERENCES

[1] N. Spencer and D. Butler, "Cities: The century of the city," *Nature*, vol. 467, no. 7318, pp. 900–901, 2010.

[2] F. Cicirelli, A. Guerrieri, G. Spezzano, and A. Vinci, "An edge-based platform for dynamic smart city applications," *Future Gener. Comput. Syst.*, vol. 76, pp. 106–118, Nov. 2017.

[3] H. H. R. Sherazi, R. Iqbal, F. Ahmad, Z. A. Khan, and M. H. Chaudary, "DDoS attack detection: A key enabler for sustainable communication in Internet of vehicles," *Sustain. Comput., Informat. Syst.*, vol. 23, pp. 13–20, Sep. 2019.

[4] R. Iqbal, T. A. Butt, M. Afzaal, and K. Salah, "Trust management in social Internet of vehicles: Factors, challenges, blockchain, and fog solutions," *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 1, Jan. 2019, Art. no. 155014771982582.

[5] Z. Dar, A. Ahmad, F. A. Khan, F. Zeshan, R. Iqbal, H. H. R. Sherazi, and A. K. Bashir, ''A context-aware encryption protocol suite for edge computing-based IoT devices,'' *J. Supercomput.*, pp. 1–20, Oct. 2019.

[6] U. M. Butt, S. Letchmunan, F. H. Hassan, M. Ali, A. Baqir, and H. H. R. Sherazi, ''Spatio-temporal crime HotSpot detection and prediction: A systematic literature review,'' *IEEE Access*, vol. 8, pp. 166553–166574, 2020.

[7] A. Baqir, S. U. Rehman, S. Malik, F. U. Mustafa, and U. Ahmad, ''Evaluating the performance of hierarchical clustering algorithms to detect spatio-temporal crime hot-spots,'' in *Proc. 3rd Int. Conf. Comput., Math. Eng. Technol. (iCoMET)*, Jan. 2020, pp. 1–5.

[8] M. Kaufmann, S. Egbert, and M. Leese, ''Predictive policing and the politics of patterns,'' *Brit. J. Criminology*, vol. 59, no. 3, pp. 674–692, Apr. 2019.

[9] F. Yi, Z. Yu, F. Zhuang, X. Zhang, and H. Xiong, ''An integrated model for crime prediction using temporal and spatial factors,'' in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2018, pp. 1386–1391.

[10] R. Tapia-McClung, ''Exploring the use of a spatio-temporal city dashboard to study criminal incidence: A case study for the mexican state of aguascalientes,'' *Sustainability*, vol. 12, no. 6, p. 2199, Mar. 2020.

[11] E. V. Altay and B. Alatas, ''Performance analysis of multi-objective artificial intelligence optimization algorithms in numerical association rule mining,'' *J. Ambient Intell. Humanized Comput.*, vol. 11, pp. 3449–3469, Oct. 2019.

[12] P. Das and A. K. Das, ''Application of classification techniques for prediction and analysis of crime in India,'' in *Computational Intelligence in Data Mining*. Singapore: Springer, 2019, pp. 191–201.

[13] L. S. Richmond-Rakerd, S. D'Souza, S. H. Andersen, S. Hogan, R. M. Houts, R. Poulton, S. Ramrakha, A. Caspi, B. J. Milne, and T. E. Moffitt, ''Clustering of health, crime and social-welfare inequality in 4 million citizens from two nations,'' *Nature Hum. Behav.*, vol. 4, no. 3, pp. 255–264, Mar. 2020.

[14] P. J. Brantingham, P. L. Brantingham, J. Song, and V. Spicer, ''Crime hot spots, crime corridors and the journey to crime: An expanded theoretical model of the generation of crime concentrations,'' in *Geographies of Behavioural Health, Crime, and Disorder*. Cham, Switzerland: Springer, 2020, pp. 61–86.

[15] NYCOpenData. (2019). *NYPD Complaint Data Historic | NYC Open Data*. Accessed: Jun. 13, 2019. [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Histori%c/qgea-i56i/data

[16] A. Meijer and M. Wessels, ''Predictive policing: Review of benefits and drawbacks,'' *Int. J. Public Admin.*, vol. 42, no. 12, pp. 1031–1039, Sep. 2019.

[17] C. Catlett, E. Cesario, D. Talia, and A. Vinci, ''Spatio-temporal crime predictions in smart cities: A data-driven approach and experiments,'' *Pervas. Mobile Comput.*, vol. 53, pp. 62–74, Feb. 2019.

[18] I. Kawthalkar, S. Jadhav, D. Jain, and A. V. Nimkar, ''Predictive crime mapping for smart city,'' in *Advances in Distributed Computing and Machine Learning*. Singapore: Springer, 2020, pp. 359–368.

[19] C.-H. Yu, W. Ding, M. Morabito, and P. Chen, ''Hierarchical spatio-temporal pattern discovery and predictive modeling,'' *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 979–993, Apr. 2016.

[20] G. Hajela, M. Chawla, and A. Rasool, ''A clustering based hotspot identification approach for crime prediction,'' *Procedia Comput. Sci.*, vol. 167, pp. 1462–1470, Jan. 2020.

[21] C. Xu, X. Hu, A. Yang, Y. Zhang, C. Zhang, Y. Xia, and Y. Cao, ''Crime hotspot prediction using big data in China,'' in *Handbook of Research on Managerial Practices and Disruptive Innovation in Asia*. Hershey, PA, USA: IGI Global, 2020, pp. 351–371.

[22] T. C. Hart, ''Hot spots of crime: Methods and predictive analytics,'' in *Geographies of Behavioural Health, Crime, and Disorder*. Cham, Switzerland: Springer, 2020, pp. 87–103.

[23] A. A. Braga, B. S. Turchan, A. V. Papachristos, and D. M. Hureau, ''Hot spots policing and crime reduction: An update of an ongoing systematic review and meta-analysis,'' *J. Experim. Criminology*, vol. 15, no. 3, pp. 289–311, Sep. 2019.

[24] C. W. Telep and J. Hibdon, *Understanding and Responding to Crime and Disorder Hot Spots*. Washington, DC, USA: Department of Justice, 2019.

[25] A. Araujo, N. Cacho, L. Bezerra, C. Vieira, and J. Borges, ''Towards a crime hotspot detection framework for patrol planning,'' in *Proc. IEEE 20th Int. Conf. High Perform. Comput. Commun., IEEE 16th Int. Conf. Smart City, IEEE 4th Int. Conf. Data Sci. Syst. (HPCC/SmartCity/DSS)*, Jun. 2018, pp. 1256–1263.

[26] S. N. Nair and E. Gopi, ''Deep learning techniques for crime hotspot detection,'' in *Optimization in Machine Learning and Applications*. Singapore: Springer, 2020, pp. 13–29.

[27] S. Shiode and N. Shiode, ''A network-based scan statistic for detecting the exact location and extent of hotspots along urban streets,'' *Comput., Environ. Urban Syst.*, vol. 83, Sep. 2020, Art. no. 101500.

[28] B. Cheng, W. Li, and H. Tong, ''Prediction of criminal suspects based on association rules and tag clustering,'' *J. Softw. Eng. Appl.*, vol. 12, no. 3, pp. 35–50, 2019.

[29] R. Kumar and B. Nagpal, ''Analysis and prediction of crime patterns using big data,'' *Int. J. Inf. Technol.*, vol. 11, no. 4, pp. 799–805, Dec. 2019.

[30] Y. Xie and S. Shekhar, ''A nondeterministic normalization based scan statistic (NN-scan) towards robust hotspot detection: A summary of results,'' in *Proc. SIAM Int. Conf. Data Mining*, Philadelphia, PA, USA: SIAM, May 2019, pp. 82–90.

[31] S. S. Deshmukh and B. Annappa, ''Prediction of crime hot spots using spatiotemporal ordinary Kriging,'' in *Integrated Intelligent Computing, Communication and Security*. Singapore: Springer, 2019, pp. 683–691.

[32] Q. Zhang, P. Yuan, Q. Zhou, and Z. Yang, ''Mixed spatial-temporal characteristics based crime hot spots prediction,'' in *Proc. IEEE 20th Int. Conf. Comput. Supported Cooperat. Work Design (CSCWD)*, May 2016, pp. 97–101.

[33] Y. Zhuang, M. Almeida, M. Morabito, and W. Ding, ''Crime hot spot forecasting: A recurrent model with spatial and temporal information,'' in *Proc. IEEE Int. Conf. Big Knowl. (ICBK)*, Aug. 2017, pp. 143–150.

[34] M. J. C. Baculo, C. S. Marzan, R. de Dios Bulos, and C. Ruiz, ''Geospatial-temporal analysis and classification of criminal data in manila,'' in *Proc. 2nd IEEE Int. Conf. Comput. Intell. Appl. (ICCIA)*, Sep. 2017, pp. 6–11.

[35] Y. Hu, F. Wang, C. Guin, and H. Zhu, ''A spatio-temporal kernel density estimation framework for predictive crime hotspot mapping and evaluation,'' *Appl. Geography*, vol. 99, pp. 89–97, Oct. 2018.

[36] S. Brayne and A. Christin, ''Technologies of crime prediction: The reception of algorithms in policing and criminal courts,'' *Social Problems*, Mar. 2020.

[37] M. H. Bhatti, J. Khan, M. U. G. Khan, R. Iqbal, M. Aloqaily, Y. Jararweh, and B. Gupta, ''Soft computing-based EEG classification by optimal feature selection and neural networks,'' *IEEE Trans. Ind. Informat.*, vol. 15, no. 10, pp. 5747–5754, Oct. 2019.

[38] H. Mushtaq, I. Siddique, B. H. Malik, M. Ahmed, U. M. Butt, R. M. T. Ghafoor, H. Zubair, and U. Farooq, ''Educational data classification framework for community pedagogical content management using data mining,'' *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 1, pp. 329–338, 2019.

[39] J. Payne and A. Morgan, ''COVID-19 and violent crime: A comparison of recorded offence rates and dynamic forecasts (ARIMA) for march 2020 in Queensland, Australia,'' Tech. Rep., 2020.

[40] M. R. Parvez, T. Mosharraf, and M. E. Ali, ''A novel approach to identify spatio-temporal crime pattern in Dhaka city,'' in *Proc. 8th Int. Conf. Inf. Commun. Technol. Develop.*, Jun. 2016, p. 41.

[41] L. Weihong, W. Lei, and C. Yebin, ''Spatial-temporal forecast research of property crime under the driven of urban traffic factors,'' *Multimedia Tools Appl.*, vol. 75, no. 24, pp. 17669–17687, Dec. 2016.

[42] Z. Li, T. Zhang, Z. Yuan, Z. Wu, and Z. Du, ''Spatio-temporal pattern analysis and prediction for urban crime,'' in *Proc. 6th Int. Conf. Adv. Cloud Big Data (CBD)*, Aug. 2018, pp. 177–182.

[43] S. K. Rumi, K. Deng, and F. D. Salim, ''Crime event prediction with dynamic features,'' *EPJ Data Sci.*, vol. 7, no. 1, p. 43, Dec. 2018.

[44] S. Hossain, A. Abtahee, I. Kashem, M. M. Hoque, and I. H. Sarker, ''Crime prediction using spatio-temporal data,'' 2020, *arXiv:2003.09322*. [Online]. Available: http://arxiv.org/abs/2003.09322

[45] R. K. Wortley and L. A. Mazerolle, *Environmental Criminology and Crime Analysis*, vol. 6. London, U.K.: Taylor & Francis, 2016.

[46] L. McInnes, J. Healy, and S. Astels, ''Hdbscan: Hierarchical density based clustering,'' *J. Open Source Softw.*, vol. 2, no. 11, p. 205, Mar. 2017.

[47] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, ''A density-based algorithm for discovering clusters in large spatial databases with noise,'' in *Proc. KDD*, vol. 96, 1996, pp. 226–231.

[48] F. Murtagh, ''A survey of recent advances in hierarchical clustering algorithms,'' *Comput. J.*, vol. 26, no. 4, pp. 354–359, Nov. 1983.

[49] F. Murtagh and P. Contreras, ''Algorithms for hierarchical clustering: An overview, II,'' *WIREs Data Mining Knowl. Discovery*, vol. 7, no. 6, Nov. 2017, Art. no. e1219.

[50] S. Pasupathi, V. Shanmuganathan, K. Madasamy, H. R. Yesudhas, and M. Kim, "Trend analysis using agglomerative hierarchical clustering approach for time series big data," *J. Supercomput.*, pp. 1–20, Jan. 2021.

[51] M. Z. Rodriguez, C. H. Comin, D. Casanova, O. M. Bruno, D. R. Amancio, L. D. F. Costa, and F. A. Rodrigues, "Clustering algorithms: A comparative approach," *PLoS ONE*, vol. 14, no. 1, Jan. 2019, Art. no. e0210236.

[52] S. F. Galán, "Comparative evaluation of region query strategies for DBSCAN clustering," *Inf. Sci.*, vol. 502, pp. 76–90, Oct. 2019.

[53] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*. OTexts, 2018.

[54] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: Concerns and ways forward," *PLoS ONE*, vol. 13, no. 3, Mar. 2018, Art. no. e0194889.

[55] Census.gov. (2019). *United States Census Bureau*. Accessed: Nov. 15, 2019. [Online]. Available: https://www.census.gov/

[56] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[57] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA, USA: Morgan Kaufmann, 2016.

[58] C. Nasa and S. Suman, "Evaluation of different classification techniques for Web data," *Int. J. Comput. Appl.*, vol. 52, no. 9, pp. 34–40, Aug. 2012.

UMAIR MUNEER BUTT received the B.S. degree in CS from GIFT University, Pakistan, in 2012, and the M.S. degree in CS from the National University of Sciences and Technology (NUST), Pakistan, in 2016. He is currently a Ph.D. Student with the School of Computer Sciences, Universiti Sains Malaysia (USM), Malaysia. He has more than six years of teaching and research experience in data mining, machine learning, data science, and image processing. He has served as a research associate for more than five years and worked on different real-world applications. Recently, he secured a Fundamental Research Grant Scheme (FRGS) from the Malaysian Government for crime predictions. He has authored several journal, conferences, and book chapter in well-reputed journals during his career. His current research interests include data science, data mining, and machine learning.

SUKUMAR LETCHMUNAN received the Ph.D. degree in computer science from the University of Strathclyde, U.K., in 2013. Since then, he has been a Senior Lecturer with the School of Computer Sciences, University Sains Malaysia (USM). He has been a Tutor, a Technical Trainer, and served as a Lecturer and a Course Coordinator with private college and private university prior to his Ph.D. studies. His research interests include software engineering, software metrics in web applications, software cost estimation, service-oriented software engineering, and agile project management.

FADRATUL HAFINAZ HASSAN received the Ph.D. degree in computer science (CS) from the School of Information Systems, Computing and Mathematics, Brunel University London, in 2013. She is currently a Senior Lecturer with the School of Computer Sciences, Universiti Sains Malaysia. Her research interests include artificial intelligence (AI) for pedestrian simulation and spatial layout optimization. She has coauthored over 30 publications and secured ten research grants; five as a principal investigator and five grants as a co-investigator. Her current research involved studying pedestrian simulation models in the urban planning domain with the University of Sydney School of Architecture, Design and Planning.

MUBASHIR ALI received the B.S. degree in computer science from Allama Iqbal Open University, Islamabad, Pakistan, in 2011, and the M.S. degree in software engineering from Bahria University, Islamabad, in 2014. He is currently pursuing the Ph.D. degree with the School of Engineering and Applied Sciences, The University of Bergamo, Italy. He was an Assistant Professor with University of Lahore Gujrat Campus, Pakistan, from August 2016 to November 2019. Previously, he worked as a Software Engineer in a research and development based public sector organizations in Pakistan for more than five years. He has authored several articles in prestigious journals and conferences. His teaching and research interests include NLP, machine learning, big data analytics, social media analysis, and software repository mining.

ANEES BAQIR received the B.S. degree in information technology from the University of Gujrat, and the M.S. degree in information technology from The University of Lahore. He is currently pursuing the Ph.D. degree in computer science with the Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy. He has authored several articles in esteemed journals and conferences. His research interests include NLP, machine learning, data science, social media analysis. Previously, he has served as a Lecturer for the Department of Software Engineering, University of Sialkot.

TIENG WEI KOH received the B.Sc. degree and M.S. degree in software engineering and the Ph.D. degree in software engineering from Universiti Putra Malaysia (UPM), in 2004, 2007, and 2011, respectively. He was appointed as the Head of the Software Engineering Research Group (SERG) for four years ended recently. At present, he is the Program Coordinator for Master of Software Engineering degree program, and a Research Associate with the Malaysian Research Institute on Ageing (MyAgeing). He is currently an Associate Professor with the Faculty of Computer Science and Information Technology, UPM. He has participated in number of research projects funded by few government agencies, such as Ministry of Higher Education Malaysia (MoHE) and Ministry of Women, Family and Community Development (KPWKM). His research interests include software measurement, and empirical software engineering.

HAFIZ HUSNAIN RAZA SHERAZI (Senior Member, IEEE) received the B.S. and M.S. degrees in computer science from COMSATS University, Lahore, Pakistan, on Fully Funded Fellowship, in 2011 and 2013, respectively, and the Ph.D. degree in electrical and information engineering from the Politecnico di Bari, Italy, on the Ministerial Fellowship, in 2018. He was an Assistant Professor with GIFT University and an Adjunct Professor with Superior University, Pakistan, in Spring 2020. Previously, he has been with the Department of Electrical and Information Engineering, Politecnico di Bari, Italy, as a Postdoctoral Researcher, from November 2018 to October 2019. He has been a Research Exchange with the University of Glasgow, U.K., from May 2017 to January 2018. He is currently a Researcher with the Tyndall National Institute, University College Cork, Cork, Ireland. Several articles in prestigious conferences and journals are on his credit. His research interests include the Internet of Things, energy harvesting, industrial automation, blockchain technologies, and vehicular networks. He is a reviewer of top ranked journals and a TCP member of renowned conferences. Furthermore, he is also a member of ACM, IEEE Young Professionals, IEEE ComSoc, and IET. Moreover, he is also serving as an Associate Editor for *Internet Technology Letter*, an Area Editor for *Ad Hoc Networks*, a Topic Editor for *Electronics*, and a guest editor for a number of prestigious journals.