IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 24, 2021, accepted May 2, 2021, date of publication May 6, 2021, date of current version May 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3078117

Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques

WAJIHA SAFAT[1], SOHAIL ASGHAR[1], (Member, IEEE), AND SAIRA ANDLEEB GILLANI[2]
[1]Department of Computer Science, COMSATS University, Islamabad 44000, Pakistan
[2]Department of Computer Science, Bahria University Karachi Campus, Karachi 75260, Pakistan

Corresponding author: Wajiha Safat (wajihasafat165@yahoo.com; fa16-rse-017@isbstudent.comsats.edu.pk)

ABSTRACT Crime and violation are the threat to justice and meant to be controlled. Accurate crime prediction and future forecasting trends can assist to enhance metropolitan safety computationally. The limited ability of humans to process complex information from big data hinders the early and accurate prediction and forecasting of crime. The accurate estimation of the crime rate, types and hot spots from past patterns creates many computational challenges and opportunities. Despite considerable research efforts, yet there is a need to have a better predictive algorithm, which direct police patrols toward criminal activities. Previous studies are lacking to achieve crime forecasting and prediction accuracy based on learning models. Therefore, this study applied different machine learning algorithms, namely, the logistic regression, support vector machine (SVM), Naïve Bayes, k-nearest neighbors (KNN), decision tree, multilayer perceptron (MLP), random forest, and eXtreme Gradient Boosting (XGBoost), and time series analysis by long-short term memory (LSTM) and autoregressive integrated moving average (ARIMA) model to better fit the crime data. The performance of LSTM for time series analysis was reasonably adequate in order of magnitude of root mean square error (RMSE) and mean absolute error (MAE), on both data sets. Exploratory data analysis predicts more than 35 crime types and suggests a yearly decline in Chicago crime rate, and a slight increase in Los Angeles crime rate; with fewer crimes occurred in February as compared to other months. The overall crime rate in Chicago will continue to increase moderately in the future, with a probable decline in future years. The Los Angeles crime rate and crimes sharply declined, as suggested by the ARIMA model. Moreover, crime forecasting results were further identified in the main regions for both cities. Overall, these results provide early identification of crime, hot spots with higher crime rate, and future trends with improved predictive accuracy than with other methods and are useful for directing police practice and strategies.

INDEX TERMS LSTM and ARIMA based crime prediction, analysis and forecast.

I. INTRODUCTION

Criminality is a negative phenomenon, which occurs worldwide in both developed and underdeveloped countries. The criminal activities can severely strike the economy as well as affect the quality of life and well-being of residents, thus leading towards social and societal issues [1]. The crimes and criminal acts can incur costs to both the public and private sectors [2]. Public safety is a considerable factor for secure environments when people travel or move to new places [3]. In reality, different kinds of crimes may be associated with distinct consequences [4]. Overall, crimes take place due to various circumstances including specific motives, human nature and behavior, critical situations and poverty [5]. Furthermore, multiple factors such as unemployment, gender inequality, high population density, child labor, and illiteracy, can cause an increase in violent crimes [6]. The growing and populated cities also have a strong correlation with higher crime rates associated with multiple types of environments such as commercial buildings and municipal housing areas [7]. A socially sustainable community heavily relies on minimizing crime so that people can live peacefully and actively, while corrupt societies cannot prosper both socially and economically in the absence of peace.

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

Consequently, analyzing the crime reports and statistics are essential to improve the safety and security of humanity while maintaining sustainable development.

Crime prediction has gained popularity in recent years because it supports the ability of investigation authorities to handle crime computationally. There is a need for better predictive algorithms, which direct police patrols toward criminals [8]. Several studies have been carried out to predict crime types, crime rates and hot spots of crime by using crime datasets for different areas, for example, in South Korea, and the U.S. (including Portland) [9], [10]. Furthermore, different pilot projects are also extended to identify crime geographical locations such as residential and commercial using the Canada dataset [11]. Research has been dedicated to implementing innovative methodologies such as machine learning and deep learning techniques to predict crimes as a rigid approach and maintain a safe and secure environment [4]. Recent examples of machine learning and deep learning algorithms for successful crime prediction and analysis are the Naïve Bayes, random forest, SVM, decision tree, and regression techniques [12], [13].

Accurate crime prediction is complicated but necessary for the prevention of criminal acts. The accurate estimation of the crime rate, types and hot spots from past patterns creates many computational challenges and opportunities. Crime prediction based on machine learning is the current mainstream for prediction analysis; however, only a few studies systematically compare different machine learning methods. The ability of machine learning algorithm in processing non-linear rational data has been confirmed in many fields, including crime prediction. It can handle very high-dimensional data with faster training speed and can extract the characteristics of the data [14]. Despite considerable research efforts, the literature lacks the relative accuracy for crime prediction from large datasets for multiple cities; such as Los Angeles and Chicago datasets have been used rarely. Recent literature further suggest that the challenges concerned with the accuracy of prediction and forecast of violent acts mainly in high crime density areas by implementing different models [15]. Given that, the crime data is usually based on time series data, which shows the data seasonality, and suggests the potential significance of crime activities evolved in the years. Therefore, time series analysis is required to generate visual patterns along with a deep learning algorithm specifically LSTM, which provides the better classification of crimes over time based on adequate measures [16]. Additionally, forecasting the crime trends through ARIMA model is highly recommended in recent research [17].

Therefore, this study aims to analyze crime prediction in the Chicago and Los Angeles datasets [18], (1) improving the predictive accuracy compared to results in the recent literature by implementing the Logistic Regression, SVM, Naïve Bayes, KNN, Decision Tree, MLP, Random Forest, XGBoost algorithms, (2) time-series analysis by LSTM, (3) creating a visual summary through exploratory data analysis, and (4) crime forecasting for the crime rate and high intensity crime areas for subsequent years by using an ARIMA model. The Chicago and Los Angeles datasets have been collected throughout the years; it is no surprise that machine learning and deep learning methods may be useful in the prediction of crime types and forecasting future benefit [19]. The overall crime rate forecasting results would benefit the police by using identified alleged crime areas to allocate additional resources and protective measures against criminals.

This study reports an improved efficiency for accurate crime prediction as compared with previously achieved with further analysis based on different machine learning algorithms. Besides crime prediction accuracy, the LSTM for time series analysis was reported using different performance metrics. Moreover, the study also provides a visual summary through exploratory data analysis to portray crime types and count. Finally, the future crime rate and crime density areas for the next five years were examined through ARIMA. The structure of this paper is organized as follow: *Section 2* discusses the literature review related to crime prediction. *Section 3* presents preliminary classification methods, prediction and performance evaluation measures. *Section 4* introduces the data and preprocessing. *Section 5* explains the major findings with a detailed comparative analysis of Chicago and Los Angeles datasets about crime prediction and future forecasting. *Section 6* covers the discussions and future directions with additional considerations and key points about models. Finally, concluding remarks are given in *Section 7*.

II. LITERATURE REVIEW

The recent literature regarding crime prediction can be categorized in different research domains [20]. For example, several studies highlight the ecological factors like education, income level, unemployment to name a few, behind crimes, while spatial-temporal crime event has also been focused [21], [22]. The recent literature also suggests that crime prediction and analysis are based on new types of data taken from online forums such as Twitter and mobile phone data [23]. Nevertheless, all these studies mainly focus only on the cause of crimes followed by their consequence [24]. Herein, we particularly emphasize the implementation of multiple techniques to achieve substantial accuracy on two large datasets.

The literature review section specifically reveals the related studies on crime prediction based on Chicago and Los Angeles datasets. This section further highlights the classification, prediction and forecasting of crimes. Different aspects of crime detection have been analyzed by different research methods. However, the overall prediction depends directly or indirectly on the information available within the given dataset for crime prediction. Chicago and Los Angeles both are populous and iconic cities of the U.S. and their datasets are available publically at authorized repositories, relating multiple traits that have been a great source of attraction for analysts. With a specific goal to the brief, there have been

different studies in recent years based on these datasets to predict accuracy and hotspot crime regions by applying multiple machine-learning algorithms, and kinds of expectation accomplished. Some of the recent studies on both cities are summarized below.

A. CHICAGO

Chicago is the third most populous city of the U.S., and crime rates are more often distinct as compared to less populated area. Most crimes are associated with location, properties and distribution of people, rather than patterns of past crimes. Some recent studies for crime analytics from the Chicago city dataset are discussed below:

Kang *et al.* used environmental context information to improve the prediction of models by proposing a feature-level data fusion method on deep neural networks [6]. This study used four multiple demographic datasets (City of Chicago Data Portal, American FactFinder, Weather Underground, and Google Street View) for the year 2014 and showed improved results after exercising area under the curve, precision and recall. Stec and Klabjan utilized the neural network idea by merging two techniques; convolutional neural network (CNN) and recurrent neural network (RNN), and achieved 75.6% accuracy [10]. The study was conducted on multiple datasets including; Portland, public transportation, weather census and Chicago dataset with 6 million records. It predicted the top three crime types namely violence, theft and narcotic crimes for Chicago; after implementing Feed Forward with 71.3%, CNN with 72.7% and RNN with 74.1% accuracy respectively. Another recent study conducted on the Chicago dataset from the year 2001 to 2019 also forecasts future crimes by using the ARIMA model [15]. They proposed their own model LFSNBC and achieved 97.47% accuracy along with SVM 67.01%, deep neural network (DNN) 84.25% and kernel density estimation (KDE) 66.33%. Najjar *et al.* used 12,000 satellite images to inquire about crime rates from data and reports gathered by the police department [12]. Their finding predicts 79% accuracy by executing CNN using the deep learning concept. Wang *et al.* implemented Linear Regression Negative Binomial Regression to figure out the MAE and mean relative error (MRE) for two Chicago datasets; the point of interest data (POI) and taxi flow [13]. POI was applied to aid the demographic features, while taxi flow was used as a hyperlink to help the neighbors by seeking geographical knowledge. Results anticipate the rapid decline in the overall crime rate.

Statistical analysis was conducted to evaluate violent and non-violent crimes by using Chicago arrest data for the social criminal network [25]. K–S test were executed to model the exact-repeat and near-repeat effects of the arrest. Catlett *et al.* applied two clustering algorithms; DBSCAN and ARIMA to detect high-risk crime regions to forecast future crime trends by using a spatial-temporal approach [26]. Catlett *et al.* proposed an approach that relies on Spatial-temporal to discover the crime in high-risk areas that are mostly urban and dependable trends for crimes forecast in every region while using clustering technique [27]. Christian *et al.* relate the socioeconomic and sustainable development indicators like poverty rate and unemployment toward crime by implementing Linear Regression Analysis from the year 2008 to 2012 foe Chicago [28]. A detailed analysis report cited by Schnell et. al. addressed 359,786 incidents and were geocoded to 41,926 street segments nested within 342 neighborhood clusters, within 76 communities from 2001 to 2014 [29]. There have been multiple studies that are performed by using geographical locations, meta-association rules and specific detection system introduced to examine the crime rate in Chicago [30].

B. LOS ANGELES

Data from Los Angeles corroborate almost identical percentages and indicate the involvement of long-term dependency on additional systems and subsequent higher costs. Different studies highlight more comprehensively the crime prediction among the dually involved population in Los Angeles. Young *et al.* studied the report of the Los Angeles Times and the Data Desk (a team of reporters and Web developers) to inspect the technological changes in the newsroom at the start of the twenty-first century [31]. The contribution of this study recommended the computational schemes appear in a discontinuous advancement of practices, identities and norms. A study conducted by Contreras further analyzes the connection between dispensaries for medical marijuana and crime rates in Los Angeles [32]. Their outcome indicates that dispensaries for marijuana are considered as an assailant of crime. Another similar study conducted by Dierkhising *et al.* reveals an intense female involvement among the sample to predict rearrests rate and child welfare histories [33]. Brantingham *et al.* analyzed the racial biases using arrest for predictive policing experiments. The findings anticipate that the total numbers of arrests by algorithmically predicted locations were numerically higher [8]. Ridgeway *et al.* further evaluate the impact of rail transit on crime from 1988 to 2014 in neighborhoods near transit stations [34]. With permutation tests, results revealed that there was no appreciable crime effect in rail transit. Valasik *et al.* inspect the environmental risk factors in East Los Angeles for the year 2012 that spatially influence gang assaults and gang violence [35]. RTM (risk terrain modeling) was used as an analytic tool that greatly aided the local law enforcement, stakeholders, and policymakers by presenting anti-gang efforts to high-risk areas.

Orsogna *et al.* addressed the complex data analysis issues by using the modeling tools for research, mathematicians and scientists to predict crime and safety measures [36]. Almanie *et al.* used the dataset for the year 2014 to predict the potential crime type and applied the Apriori algorithm, Naïve Bayesian and Decision Tree [37]. The result achieved 54% prediction accuracy with 'robbery' as a major attempted crime. Wang *et al.* predicted the spatial-temporal crime distribution in Los Angeles over the last six months of 2015 [4]. Results provide reliable guidance for crime

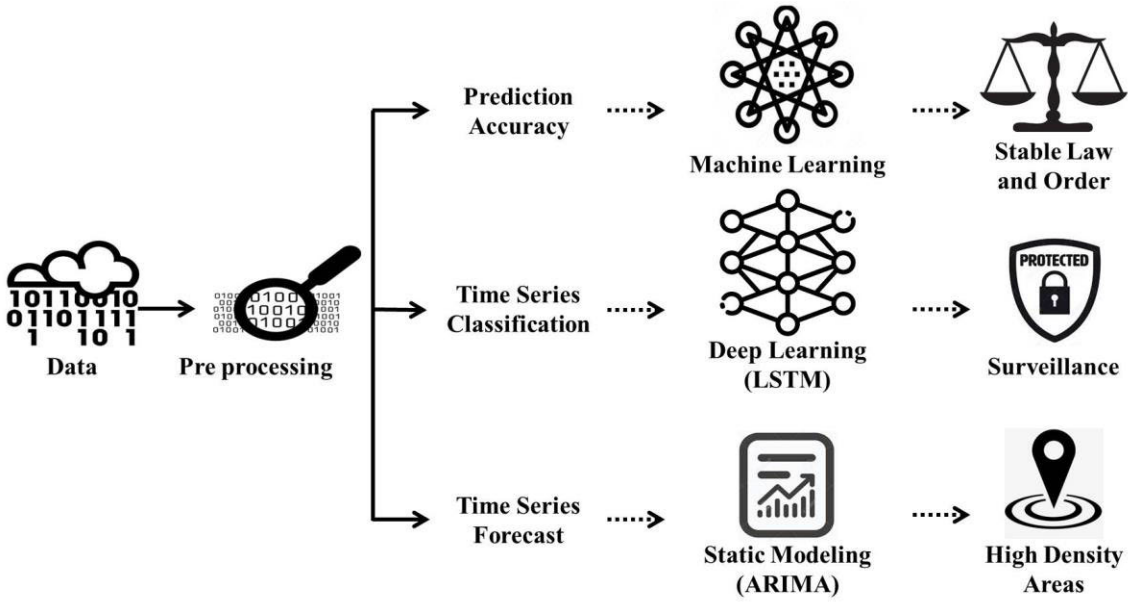

FIGURE 1. Proposed methodology and study framework.

control after applying ST-ResNet and CNN on 104,957 crimes. Sungyong *et al.* analyzed the classification of crime, whether the crime is related to gang-oriented or not through Generative Neural Network (GNN) [38]. The model is capable to classify gang-oriented crimes when complete information is available from 2014 to 2016 in Los Angeles dataset. For crowd-sourcing crime prediction, the Hawkes technique was introduced on Los Angeles crime reports, which requires no previous history [39]. This method illustrate a real-time crime model with an online k-mean type algorithm.

Overall, studies on crime prediction and forecast highlight multiple research aspects, based on multiple cities worldwide. All these studies mainly involve different types of models including socio-economic factor that features education, income level, unemployment to name a few. In addition to socio-economic factors, multiple computational models have been proposed to enhance crime prediction, classification and forecast; and the spatial-temporal models, which specifically assess the hotspot crime regions. Different methodologies have been analyzed for crime prediction in different cities such as South Korea, the U.S. (including Portland), and Canada, and many others [9]–[13]. Significant research effort has been made in different aspects, yet literature is still pointing major concern towards better prediction accuracy, forecast and hotspot in large datasets such as Chicago and Los Angeles cities. The results and discussion part is divided on prediction accuracy, time series analysis and time series forecasting as shown in Fig. 1.

III. PREDICTION AND FORECASTING

Crime prediction and forecasting approaches have transformed dramatically in recent years since the introduction of commercial software packages. Crime prediction refers to the accuracy of reported crimes in the past, whereas forecasting direct towards the future crime trends. However, a quick overview of criminal activities has been achieved by investigation authorities through the available software packages, whereas for deep analysis, only learning approaches may ensure the optimum solution. Therefore, different machine learning techniques can be used to predict crime patterns and thus may assist in further necessary actions based on historical data. Therefore, this study is divided into two sections: i) crime prediction and ii) crime forecasting. Eight different machine learning algorithms are implemented to achieve highly accurate predictions in both the Chicago and Los Angeles datasets. The machine-learning algorithms implemented in this study were namely logistic regression, decision tree, random forest, MLP, Naïve Bayes, SVM, XGBoost, and KNN to get the crime prediction accuracy. Detailed information about these machine learning algorithms models architecture is given in the supplementary information (SI) and an experimental flow chart is given in Fig. 2. The prediction results further identify areas with high crime density, all crime types and the crime rate over the past years. Additionally, the statistical model ARIMA for time series analyses was applied to foresee future crime trends and analytics.

Crime forecasting based on time series data was also implemented in a later part of this study. A time-series analysis involves forecasting based on a sequence of events or data points that forms a series with respect to time [40]. Research groups around the globe have recently used different approaches, including unsupervised models such as the bilinear model, the threshold autoregressive (tar) model, the autoregressive conditional heteroscedastic (ARCH) and deep learning approaches, to identify future trends [41]. Real-time crime forecasting is always critical; especially in unknown

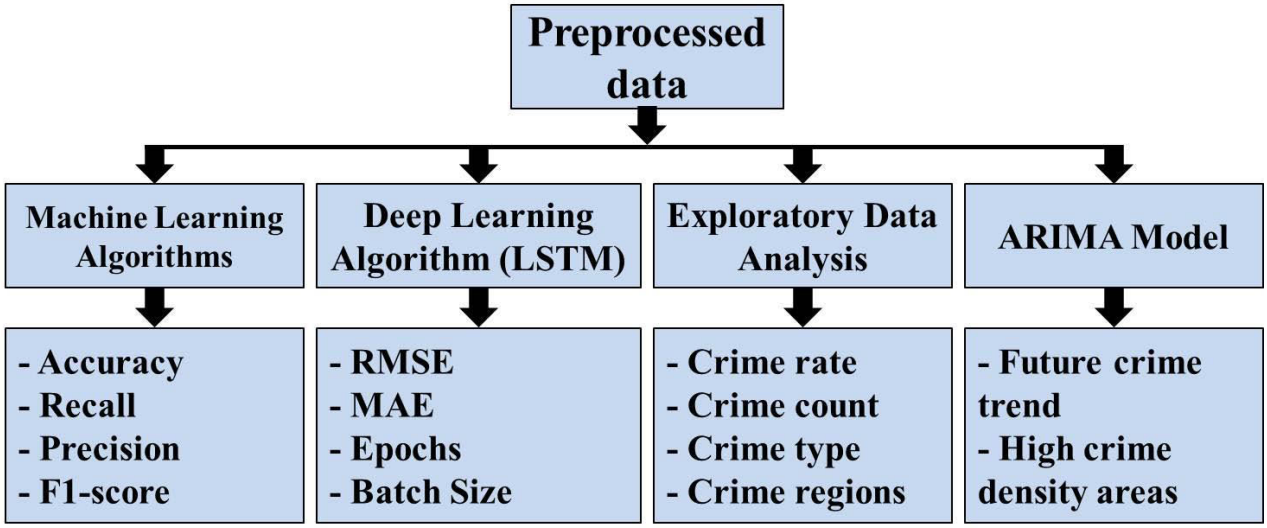

FIGURE 2. Experimental flow chart.

circumstances; when and where the next crime will happen remains difficult to predict accurately [42]. Therefore, we used an ARIMA model for future forecasting and calculated the RMSE to aggregate the magnitudes of the errors and crime predictions. The details of the ARIMA model are discussed in the SI. The forecasting results illustrate future crime trends by highlighting the crime hot spots, top five crimes and overall crime rates until 2024.

IV. DATA AND PREPROCESSING

The data used in this study consists of criminal records for the cities of Chicago and Los Angeles, and is the most decisive part to achieve the crime prediction accuracy. Herein, we used two big datasets namely Chicago and Los Angeles obtained from open access data portals and are easily downloadable in CSV format [18], [19].

The dataset of Chicago city contains the crime history (reports and social factors) from 2001 to November 2019. With 2.7 million population density, Chicago appears to be higher in crime density and the crime rate has been reported to double during the 2005 to 2008 period as compared to the rest of the U.S. where approximately 16% circulation rate was predicted by 2012 [43]. Given that, the situation drives the police officials to revise their policies, which later consequently showed a decreasing trend in recent years. The freshly available dataset contains detailed information regarding time, location (i.e., latitude and longitude), and types of crime, with 22 attributes along with more than 7 million instances. Los Angeles: The dataset of Los Angeles city contains the criminal history from 2010 to 2018. With 3.9 million population density, crime reports have been declined significantly until 2015, but with an increasing trend after 2015. The Los Angeles dataset is reported by the Los Angeles police department, and contains 17 attributes with more than 2.6 million instances. Initially there were 7019734 crime instances within the Chicago dataset, and 16913 crimes were removed due to invalid formatting (missing data, fates, values etc.). The experiment is performed on 7002821 instances of the Chicago dataset. In the Los Angeles dataset, there were 2651233 instances initially and 4770 instances were removed during data pre-processing. Finally, there were 2646463 instances for Los Angeles for experiments. The common attributes were chosen in both datasets for better comparative analysis which were named as ID, date, crime primary type, description of the crime, location, year, zip code, and police district. Both Chicago and Los Angeles datasets have 35 different crime types.

The accumulated raw data from online repositories usually contains irrelevant information and errors. The overall data will likely have noise, inconsistencies, outliers, bungles, and missing qualities or, more fundamentally, data is inconsistent to start method. Therefore, the selection of meaningful data is necessary to eliminate anomalies against the outliers, noise, missing values, and other discrepancies, and thus change over the unfeasible data into possible is manageable to accomplish information handling. Additionally, collection for the more mind-blogging framework is always required keeping in view the current developing rate of data in business, industry applications, science, and research network. The data preprocessing solidifies data planning, exacerbated by mix, cleaning, institutionalization, and change of data; data decrease assignments, thereby reducing the multifaceted design of the data, perceiving or expelling unessential and uproarious components from the data through element assurance, occurrence choice, or discretization frames, and thus finally assists to generate statistically significant data to make accurate crime predictions [44]. Therefore, the bootstrap random sampling method as shown in Fig. 3; an over feature selection method, which is also common since it is the least biased method

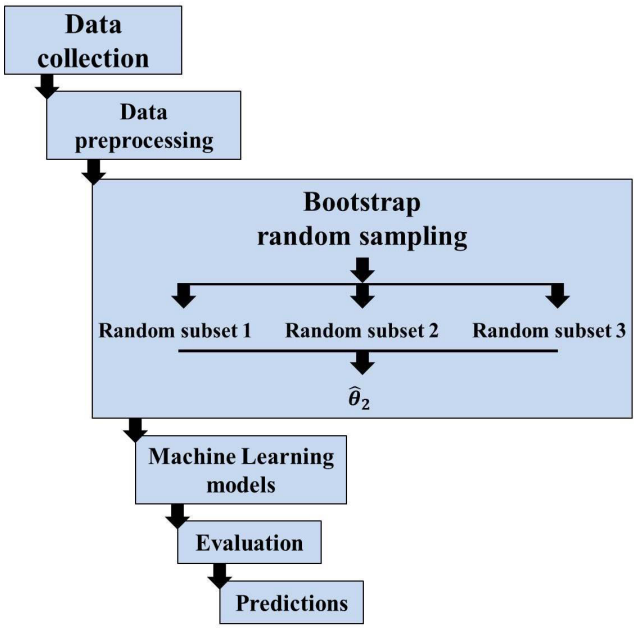

FIGURE 3. Bootstrap random sampling method.

to generate estimates of population parameters; specifically when the dataset is big [45]. Initially the datasets were examined from different sources and to take common attributes. In total, there are 9 common attributes in both datasets, and data cleaning was assured by removing all missing values. For implementation, Python (version 3.6.3) framework was used with different libraries mainly for data transformation e.g., imblearn and sklearn. The final attributes considered for this study were named as ID, date, crime primary type, description of the crime, location, year, zip code and police district. Therefore, the data is divided into test sets (30%) and training set (70%). Finally, there were 7002821 instances for Chicago and 2646463 instances for Los Angeles after preprocessing step. Accuracy, precision, recall and f1-score are the main parameters used for performance evaluation in this study.

V. RESULTS

The results and discussion part is divided into four sections based on methodology as shown in Fig. 1; predictive accuracy, time series analysis through LSTM, exploratory data analysis, and forecasting with an ARIMA model. The experimental results are also shown and discussed in each section. First, the predictive accuracy is discussed based on different algorithms. In the second part, time series analysis was performed thorough LSTM to measure the performance of the model. Thereafter, crime particulars are thoroughly discussed in the exploratory data analysis section, and finally, crime forecasting and future crime trends are shown through the ARIMA model. Different Python libraries were applied including Keras with Tensor Flow, Sk Learn, Pandas, Numpy, Seaburn, Scipy, and many others to generate the results.

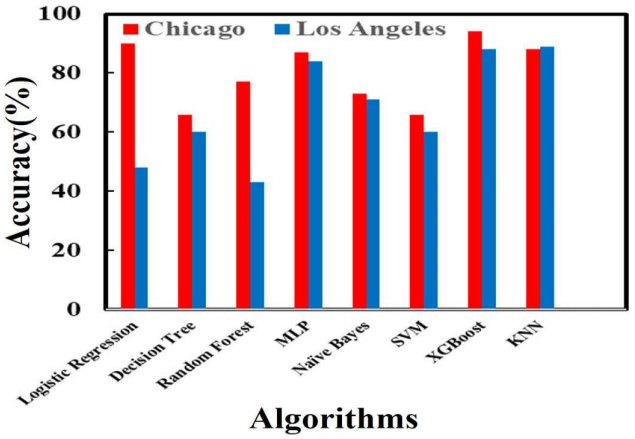

FIGURE 4. Predictive accuracy based on selected algorithms used for Chicago and Los Angeles datasets.

A. PREDICTIVE ACCURACY

This study used different parameters to assess the performance of multiple algorithms, which better reflect the real dataset application. Eight different algorithms were applied to the Chicago and Los Angeles datasets to investigate the detailed predictive accuracy of the trained models, as shown in Fig. 4. To the best of our knowledge, these algorithms have not been implemented together for Chicago and Los Angeles datasets. Consequently, the main reason to choose these cities is population density, which reported higher crime rates in the past with big data. The implemented algorithms have different methodologies to refine the data that involves supervised, unsupervised and reinforcement learning approaches. Additionally, Random Forest and XGBoost were also implemented which prompts an ensemble learning approach. Decision Tree layout the significant decisions, while SVM and Naïve Bayes are used for better classification and KNN for advance regression. To handle dependent variables Logistic regression is implemented along with MLP which refers to the network of multiple layers of the perceptron. Since all these mathematical expressions help to seek improved accuracy to the best of their proficiency, with other performance metrics such as precision, recall and F1-score, as listed in Table 1. The accuracy estimates the proportion of instances that are correctly classified to obtain the optimum threshold for crime prediction. XGBoost performs better than other algorithms with 94% and 88% accuracy on both the Chicago and Los Angeles datasets, as multiple innovative algorithms work behind XGBoost. The Naïve Bayes, MLP (with hidden layer sizes of 24, 28, 30, and 34), and SVM algorithms also achieve a better performance on the Chicago dataset than on the Los Angeles dataset with maximum accuracy. The decision tree algorithm achieves an accuracy of approximately 66% (Chicago) and 60% (Los Angeles). The MLP (87 and 84%) and KNN (88 and 89%) algorithms also approach the maximum accuracy on both datasets. The logistic regression model determines the statistical relationship between variables to achieve optimal results; here, it depicts consistent performance with 90% accuracy on the Chicago dataset and

TABLE 1. Performance parameters for Chicago and Los Angeles datasets.

Algorithms	Accuracy (%)		Precision		Recall		F1-Score	
	Chicago	Los Angeles	Chicago	Los Angeles	Chicago	Los Angeles	Chicago	Los Angeles
Logistic Regression	90	48	0.93	0.72	0.90	0.48	0.91	0.56
Decision Tree	66	60	1.00	0.98	0.66	0.60	0.75	0.68
Random Forest	77	43	0.92	0.83	0.77	0.43	0.81	0.54
MLP	87	84	1.00	0.98	0.87	0.84	1.00	0.97
Naïve Bayes	73	71	1.00	0.88	0.73	0.71	1.00	0.79
SVM	66	60	1.00	0.80	0.75	0.55	1.00	0.64
XGBoost	94	88	1.00	1.00	0.91	0.88	1.00	1.00
KNN	88	89	0.88	1.00	0.88	0.89	0.88	1.00

achieves below average results on the Los Angeles dataset. All these reported accuracy results are higher as compared with the literature.

Conversely, the random forest model achieves 77% accuracy on the Chicago dataset, while the Naïve Bayes algorithm achieves almost the same results on the Los Angeles dataset. The accuracy also depends on how often the crime happened in the past, and predicting rare crimes in the population of interest might result in low accuracy. However, the SVM algorithm achieves average results; the random forest model achieves the worst results on the Los Angeles dataset. Overall, the performances of these machine-learning algorithms are more consistent in the Chicago dataset than in the Los Angeles dataset.

The classification quality is usually evaluated on the performance of objective functions such as precision, recall and F1-sore. The recall presents the relevant instances that are retrieved by the classifier, whereas the precision is the percentage of correctly classified samples. Both functions simultaneously and optimize the two objectives with an inverse relationship, whereas the F1-score is the weighted average of recall and precision. The Chicago dataset yielded the highest performance metrics compared with the Los Angeles dataset and suggests better and stable algorithm performance. The general performance parameters, i.e., precision, recall, and F1-score, for the Los Angeles dataset are not stable enough, thereby suggesting moderate performance. XGBoost exhibits better results for precision, recall, and F1-score than the other models.

B. TIME SERIES ANALYSIS THROUGH LSTM

LSTM is an elegant variation in the RNN architecture, which is an approach that can be applied to model sequential data. The structure of LSTM makes it an effective solution to combat the vanishing gradient problem of RNNs. It uses memory capable of representing the long-term dependencies in sequential data. LSTM ensures improved learning for time series by capturing the structure of sequential data more naturally and even performs hierarchical processing for complex temporal tasks. Time series classification tasks are different from traditional classification and regression predictive modeling problems and have been considered challenging in terms of data mining for the last two decades [46]. From electronic health records to cybersecurity, almost all real-world applications require time-series data for classification [47]. A detailed description of LSTM is provided in the SI [48].

Prior to LSTM implementation, the data were preprocessed to reduce noise and then transformed into stationary data. Time series data are usually in non-stationary form and must be transformed into stationary form for easier handling and better classification [49]. Therefore, the Dickey-Fuller test is conducted to check for stationary data in a standard way and to further evaluate the appropriate error scores [50]. The results provide in-depth guidance from data processing and training of the LSTM model for a set of time-series data. For time-series data, different types of errors are usually measured, such as the scale-dependent error and percentage error. Herein, two known scale-dependent error measures were used, namely, the RMSE and the MAE, along with the number of epochs and batch size. The RMSE measures the average magnitude of the errors. Specifically, it is the square root of the average of the squared differences between the predicted and actual observations. Therefore, the RMSE will be more useful when large errors are particularly undesirable. The MAE measures the average magnitude of the errors in a set of predictions, regardless of their direction. Therefore, it is the average across the test sample of the absolute differences between the predicted and actual observations where all the individual differences have equal weight. The performance metrics of LSTM are listed in Table 2, which indicates the performance of the corresponding model in the testing data rather than the training data.

The outcome of the epochs showed the same loss value after the 13th iteration for the Chicago dataset, whereas for Los Angeles, the loss value started repeating after the 18th iteration. There is no evidence training the network with the same dataset more than once would improve the accuracy of the prediction. In some cases, the performance even worsens, indicating that the trained models are overfitting. However, apparently setting the number of epochs to 1 generates a reasonable prediction model [51]. The performance of LSTM seems to be adequate for time series analysis, especially for RMSE and MAE, where it can classify the data focusing on their variations.

TABLE 2. Performance metrics for LSTM.

Algorithms	Chicago	Los Angeles
Number of Epochs	40	40
Batch size	33	31
RMSE	12.66	8.78
MAE	11.70	6

Fig. 5 shows the approximate distribution of the mean crime density areas in different periods after LSTM implementation. The different frequencies include the daily, weekly, monthly, quarterly, and yearly results, as shown in Fig. 5. The mean crime density area for Chicago has an intense variation trend mainly in daily and weekly data, whereas the monthly and quarterly data have moderate variation trends (Fig. 5A). However, the mean crime type for Los Angeles presents some variations initially and then a decreasing trend in recent years, finally becoming stable (Fig. 5B). The overall process involves developing a function that calculates and presents the moving average of the events in the neighborhood of the events. In recent years, the majority of mean crime types demonstrate a downward trend, which suggests a further decline in the majority of forecasts in the overall time intervals (Fig. 5). However, it is not applicable when the historically upward trend is related to other criminal offenses. The time series classification is potentially a direct indicator, but it cannot be treated as an approximation of specific values, but rather as a data-driven model.

C. EXPLORATORY DATA ANALYSIS

This section discusses the detailed periodic insights of the Chicago and Los Angeles statistics. The term crime count refers to the number of crime incidents, while high-intensity crime areas are the hot spot crime regions referring to the district level location. The results are obtained by using the inspection module in Python, where the crime rate is the crime count normalized by the population for time. The study identifies 35 different crime types for Chicago and 39 for Los Angeles. Fig. 6A shows the annual trends for Chicago, showing a significant decrease in the crime rate, while Los Angeles shows an increasing trend in recent years. Previous studies on Chicago also suggest that ecological factors such as harsh weather conditions or the winter season may decrease crime and may favor people and residents [51].

The crime rates declined in February compared to other months in both Chicago and Los Angeles (Fig. 6B). Chicago had the highest crime rates in July and August, starting in March, where it tends to start and after July and August, the crime rates declined. Similarly, Los Angeles had fewer crime events in February in contrast to other months, as shown in Fig. 6B. Furthermore, the crime rates were higher on Fridays in Chicago and lower on Saturdays and Sundays (Fig. 6C), whereas in Los Angeles, the crime rates were higher on Wednesdays and lower on Saturdays and Sundays (Fig. 6C). This study further examined the crime hot spot districts for the crime with their corresponding numbers of crime incidents (Fig. 7). There were 24 crime regions in Chicago and Los Angeles with the highest crime rates with further extensive insights (Table 3). Fig. 7A and 7B display the hot spot regions for Chicago and Los Angeles with their respective crime counts. Additionally, future crime density areas were also studied by using an ARIMA model, which will be discussed in the next section. The crime types and their estimated intensities are even more important to determine the anticipated chance of crime occurrences. The visual frequencies of each crime type with the corresponding crime count are shown in Fig. 8. Theft, battery, criminal damage, narcotics, offense, robbery, motor vehicle theft, deceptive practice, burglary, assault, and theft were the main crimes observed in Chicago (Fig. 8A). Miscellaneous offenses, larceny-theft, assault, narcotics, burglary, grand theft auto, juvenile theft, kidnapping, vehicle loss, vandalism, and accidents were the main crime types in Los Angeles (Fig. 8B). The visual representation allows investigation authorities to take special measures against these violations.

D. FORECASTING WITH AN ARIMA MODEL

Time series forecasting demonstrates its importance in building an effective model, especially in the field of applied sciences [52]. A variety of models are currently available in the literature, and ARIMA models are considered a standard method for time series forecasting [53]. The advantage of ARIMA models is that the seasonal information obtained from other models (e.g., STL) can be incorporated into the predictions. An ARIMA model is a composite model for time series data combining a traditional autoregressive moving average (ARMA) model and autoregressive (AR) moving average (MA) processes [51]. It captures temporal structures using a linear regression-based approach to perform one-step out-of-sample or multistep out-of-sample forecasting. For crime prediction datasets, the algorithm forecasts the time series based on a rolling forecasting origin that focuses on a single forecast and the next data point to predict. The algorithm first splits the data set into training and testing sets (70% and 30%, of the original data, respectively). It then builds two data structures to hold the accumulated added training data-set at each iteration, (history) and the continuously predicted values for the test data-sets, (prediction). The detailed structure of the algorithm is defined in the SI. Stationary series typically have constant values, and the autocorrelation coefficient quickly decays to zero.

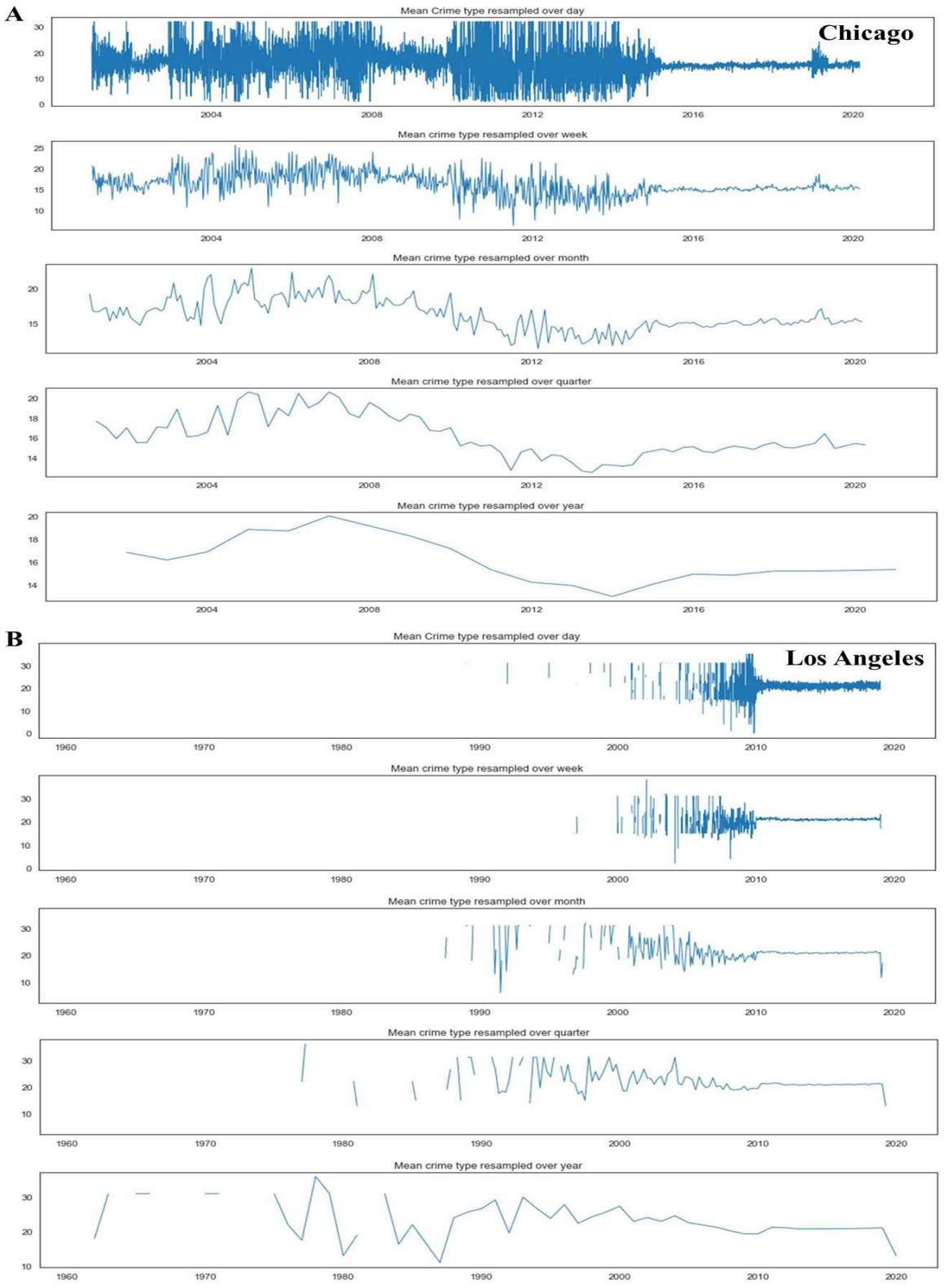

FIGURE 5. Time series analysis with respect to mean crime density area for daily, weekly, monthly, quarterly, and yearly.

Initially, it was assumed that the time series data were stable after differentiation with bounded fluctuation. The ARIMA model was used for forecasting after passing the noise test, and later, a Dickey-Fuller test was conducted to

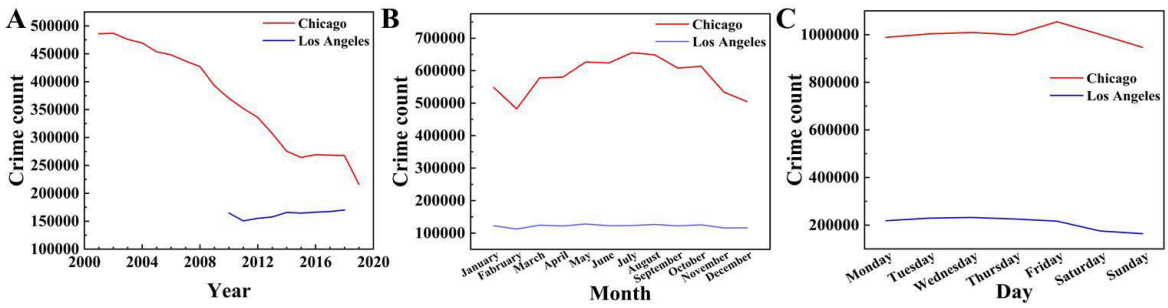

FIGURE 6. Yearly, monthly and daily crime rate trend analysis.

TABLE 3. Crime density areas with crime count and locations.

Chicago			Los Angeles		
Location	**ID**	**Count**	**Location**	**ID**	**Count**
Central	1	277291	Los Angeles	1	351483
Wentworth	2	331648	Lancaster	2	205952
Grand Crossing	3	355235	Compton	3	151790
South Chicago	4	396361	Palmdale	4	148335
Calumet	5	310291	Carson	5	94116
Gresham	6	405138	Norwalk	6	89614
Eaglewood	7	412052	Lynwood	7	86078
Chicago Lawn	8	474460	Bellflower	8	79831
Deering	9	346249	Lakewood	9	63046
Ogden	10	300829	Pico Rivera	10	61665
Harrison	11	449866	Whittier	11	60495
Near West	12	343315	West Hollywood	12	57982
Shakespeare	14	275361	Paramount	13	56657
Austin	15	304400	Castaic	14	46456
Jefferson Park	16	231448	Cerritos	15	45130
Albany Park	17	202423	Rosemead	16	41883
Near North	18	311468	Canyon Country	17	35918
Town Hall	19	312306	La Puente	18	35896
Lincolen	20	121991	Valencia	19	31240
Outskirts	21	4	La Mirada	20	31214
Morgan Park	22	229221	Commerce	21	29327
Rogers Park	24	208895	Santa Clarita	22	28901
Grand Central	25	402326	Diamond Bar	23	28728
Others	31	196	San Dimas	24	28514

examine the stationarity of the data. The prediction results of the ARIMA model for Chicago and Los Angeles are shown in Fig. 9. The objective of an ARIMA analysis is to determine the best predictive performance for the data of interest. The ARIMA model performs favorably to the alternative models. It presents the distribution of the results obtained for each dataset with all architectures depending on the historical window length.

Finally, the study forecasts the crime rate and hotspots for both Chicago and Los Angeles to ultimately support proactive policing strategies. The mean crime count is calculated to forecast the five-year crime trend. The RMSEs of the forecasted crime rate for Chicago and Los Angeles were 31.8 and 24.65 and MAE was 29.8 and 20.83 respectively. The Chicago crime rate pattern had intense variations in recent years, and variation will continue to increase moderately in the future, followed by a stable decline, probably in subsequent years, as observed in Fig. 9A. The Los Angeles crime rate has been stable over the last few years, and forecasts suggest a sharp decline in the future (Fig. 9B). After taking the mean of high crime density areas identified as crime hot spot (Fig. 9A and B, x-axis is the number of crimes and y-axis is the years). The Chicago crime intensity for crime density areas as hot spots increased slightly

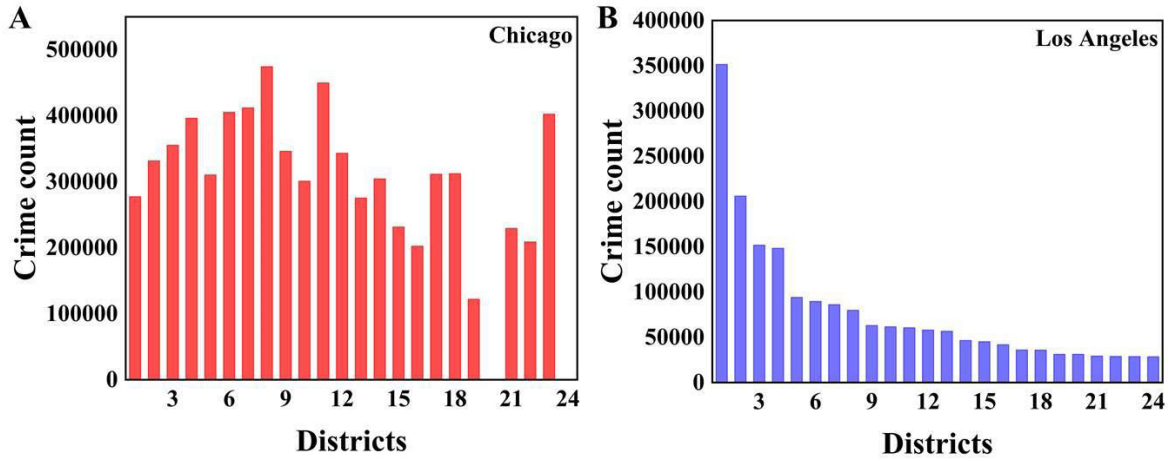

FIGURE 7. Crime density areas with crime count.

(Fig. 9C), where the x-axis is the top ID locations having higher crime rates in the past and the y-axis is the year. The Los Angeles crime intensity for the hot spot declined sharply (Fig. 9D).

VI. DISSCUSSION

Criminality is a phenomenon that occurs seemingly random and multiple research efforts have been made to develop rigorous and independent assessments. However, this study highlights the practical perspective of criminology by introducing predictive analysis through possible methods based on real-time data. Therefore, implementation of different machine learning algorithms were examined including LSTM and ARIMA modeling. First, the performance of different machine learning algorithms namely logistic regression, SVM, Naïve Bayes, KNN, decision tree, MLP, random forest and XGBoost were examined on datasets of Chicago and Los Angeles. The efficiency of prediction accuracy achieved by different algorithms is comparatively better than those reported earlier and suggests better performance. The performance of machine learning algorithms is more consistent for the Chicago dataset as compared with the Los Angeles dataset; where XGBoost achieves improved efficiency for prediction accuracy (around 94% and 88%) followed by KNN (around 88% and 89%) on both crime datasets. Herein, this study reports the better prediction accuracy for Los Chicago and Angeles, which are 94 % and 88% respectively including all types of crimes whereas previous literature report 75.6% accuracy for Chicago by using the dataset until the year 2014 by only three types of crimes namely, violence, theft and narcotic [10]. Also, the Los Angeles dataset is rarely been used and just a few studies were conducted like permutation test and K-S test for gang assaults and gang violence; while recently Almanie *et al.* predicts 54% prediction accuracy with 'robbery' as a major crime [37], [39]. Second, LSTM further classifies the crimes over different periods (yearly, quarterly, monthly, weekly and daily). LSTM performance was evaluated based on RMSE, MAE, number of epochs and batch size. In addition to crime prediction accuracy and LSTM classification, exploratory data analysis provides a visual summary for better comparative analysis between both cities. Results identify the different crime count, crime type, in different classified locations with 35 crime types for both Chicago and Los Angeles. The annual crime trend represents a significant decrease in the Chicago crime rate and Los Angeles indicates an increase in recent years. Furthermore, theft, battery, criminal damage, narcotics and offense were the top five crimes observed in Chicago whereas miscellaneous offenses, larceny-theft, assault and narcotics were the main crime types reported in Los Angeles. Finally, the crime forecasting for crime rate and high-density crime areas for the next five years by using an ARIMA model. ARIMA model suggests that the Chicago crime rate continue to increase moderately in the future whereas suggests a sharp decline for Los Angeles. This study reports the five-year crime trend and high crime density areas until 2024 with ARIMA, as compared with previous reports by using ARIMA. The Chicago crime density in hot spots increased slightly whereas it will sharply decline in Los Angeles. ARIMA model performs better as compared with LSTM based on RMSE and MAE. Overall, the proposed aims and objectives of the study are fulfilled and portray a clear picture of machine learning and deep learning techniques and their implementation with potential for different types of big datasets. All these results could benefit the situational awareness with the help of descriptive graphs that depicts the trend analysis with future forecast. Findings will further assist the law enforcement agencies and investigation departments to determine policies and meaningful insights like high crime density areas and helps the government and city management to ensure public safety. As a future augmentation, we intend to apply hybrid models to expand crime prediction accuracy and to enhance the overall performance. In addition, future work plans to build up visual images and location maps creating effectual anticipation from

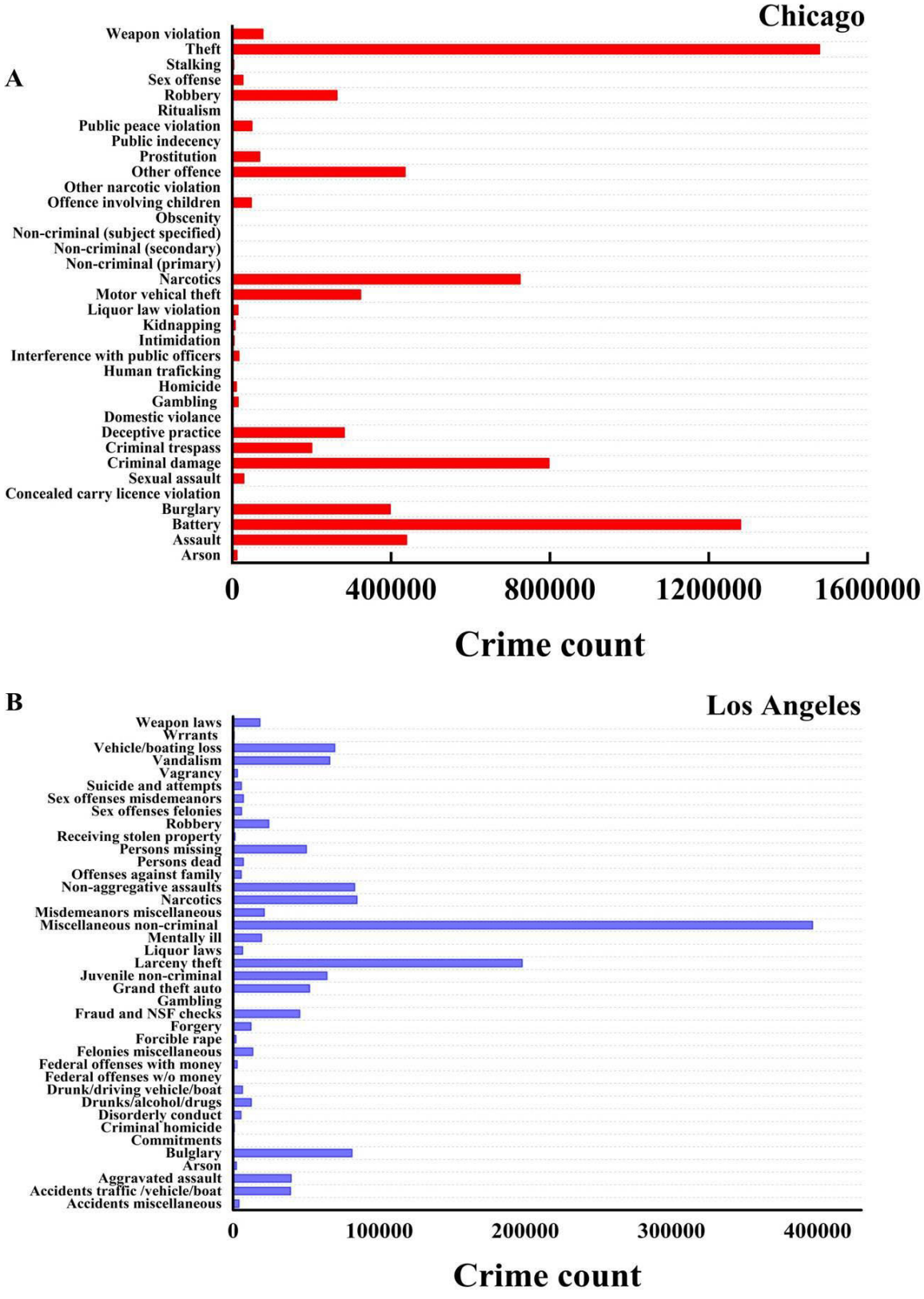

FIGURE 8. Crime rate with respect to all crime types.

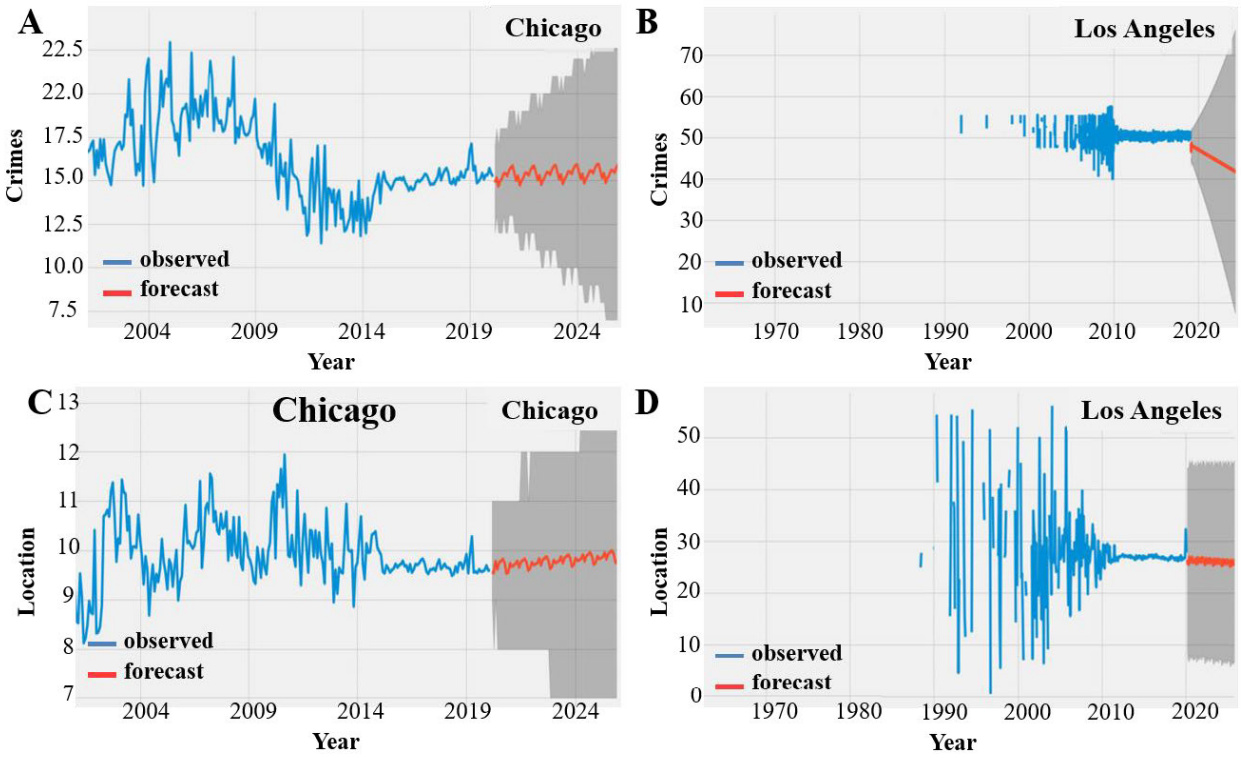

FIGURE 9. Forecast analysis of crime rates and crime density areas by ARIMA model.

the foreseen crime event providing a chance to upgrade the regulation of the patrolling system by police.

VII. CONCLUSION

Crimes are serious threats to human society, safety, and sustainable development and are thus meant to be controlled. Investigation authorities often demand computational predictions and predictive systems that improve crime analytics to further enhance the safety and security of cities and help to prevent crimes. Herein, we achieved an improved predictive accuracy for crimes by implementing different machine learning algorithms on Chicago and Los Angeles crime datasets. Among the different algorithms, XGBoost achieves the maximum accuracy on Chicago datasets and KNN achieves the maximum accuracy on Los Angeles. Data preprocessing was followed by splitting the dataset into training and testing sets, and later the performance parameters were examined. This study further applied a deep learning architecture for time series analysis through LSTM, by which the Chicago crime count had intense variations compared with Los Angeles, as shown by the RMSE and MAE. Also, the exploratory data analysis exhibited extensive visualizations regarding crime particulars, including crime rates in different periods from daily to yearly trends, crime types, and high-intensity areas based on historical patterns. Moreover, the implementation of an ARIMA model to predict the five-year trends regarding the crime rate and hot spots having high crime density suggest moderate variations for Chicago and a decline for Los Angles. For future work, this study will be expanded by using satellite imagery data, and the implementation of different learning techniques with corresponding visual data for different crime datasets.

APPENDIX

The machine-learning algorithms implemented in this study (Logistic Regression, SVM, Naïve Bayes, KNN, Decision Tree, MLP, Random Forest, XGBoost) LSTM and ARIMA models are detailed in SI.

ACKNOWLEDGMENT

Wajiha Safat acknowledges the financial support for M.S. study from COMSATS University, Islamabad. She especially thank Dr. Abdul Ghaffar (Institute of Metal Research, Chinese Academy of Sciences, Shenyang) for fruitful discussions.

COMPLIANCE WITH ETHICAL STANDARDS

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] G. Mohler, "Marked point process hotspot maps for homicide and gun crime prediction in Chicago," *Int. J. Forecasting*, vol. 30, no. 3, pp. 491–497, Jul. 2014.
[2] A. Iriberri and G. Leroy, "Natural language processing and e-government: Extracting reusable crime report information," in *Proc. IEEE Int. Conf. Inf. Reuse Integr.*, Las Vegas, IL, USA, Aug. 2007, pp. 221–226.

[3] V. Pinheiro, V. Furtado, T. Pequeno, and D. Nogueira, "Natural language processing based on semantic inferentialism for extracting crime information from text," in *Proc. IEEE Int. Conf. Intell. Secur. Informat.*, Vancouver, BC, Canada, May 2010, pp. 19–24.

[4] B. Wang, P. Yin, A. L. Bertozzi, P. J. Brantingham, S. J. Osher, and J. Xin, "Deep learning for real-time crime forecasting and its ternarization," *Chin. Ann. Math., B*, vol. 40, no. 6, pp. 949–966, Nov. 2019.

[5] S. Chackravarthy, S. Schmitt, and L. Yang, "Intelligent crime anomaly detection in smart cities using deep learning," in *Proc. IEEE 4th Int. Conf. Collaboration Internet Comput. (CIC)*, Philadelphia, PA, USA, Oct. 2018, pp. 399–404.

[6] H.-W. Kang and H.-B. Kang, "Prediction of crime occurrence from multimodal data using deep learning," *PLoS ONE*, vol. 12, no. 4, Apr. 2017, Art. no. e0176244.

[7] A. Fidow, M. Hassan, M. Imran, X. Cheng, C. Petridis, and C. Sule, "Suggesting a hybrid approach mobile apps with big data analysis to report and prevent crimes," in *Social Media Strategy in Policing* (Security Informatics and Law Enforcement), B. Akhgar, P. S. Bayeri, and G. Leventakis, Eds. Cham, Switzerland: Springer, 2019, pp. 177–195.

[8] P. J. Brantingham, M. Valasik, and G. O. Mohler, "Does predictive policing lead to biased arrests? Results from a randomized controlled trial," *Statist. Public Policy*, vol. 5, no. 1, pp. 1–6, Jan. 2018.

[9] A. Nasridinov and Y.-H. Park, "A study on performance evaluation of machine learning algorithms for crime dataset," *Adv. Sci. Technol. Lett.*, vol. 90, pp. 90–92, Dec. 2014.

[10] A. Stec and D. Klabjan, "Forecasting crime with deep learning," 2018, *arXiv:1806.01486*. [Online]. Available: http://arxiv.org/abs/1806.01486

[11] J. Fitterer, T. A. Nelson, and F. Nathoo, "Predictive crime mapping," *Police Pract. Res.*, vol. 16, no. 2, pp. 121–135, Mar. 2015.

[12] A. Najjar, S. Kaneko, and Y. Miyanaga, "Crime mapping from satellite imagery via deep learning," 2018, *arXiv:1812.06764*. [Online]. Available: http://arxiv.org/abs/1812.06764

[13] H. Wang, D. Kifer, C. Graif, and Z. Li, "Crime rate inference with big data," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 635–644.

[14] X. Zhang, L. Liu, L. Xiao, and J. Ji, "Comparison of machine learning algorithms for predicting crime hotspots," *IEEE Access*, vol. 8, pp. 181302–181310, 2020.

[15] G. R. Nitta, B. Y. Rao, T. Sravani, N. Ramakrishiah, and M. BalaAnand, "LASSO-based feature selection and Naïve Bayes classifier for crime prediction and its type," *Service Oriented Comput. Appl.*, vol. 13, no. 3, pp. 187–197, Sep. 2019.

[16] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[17] Z. Zhang, D. Sha, B. Dong, S. Ruan, A. Qiu, Y. Li, J. Liu, and C. Yang, "Spatiotemporal patterns and driving factors on crime changing during black lives matter protests," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 11, p. 640, Oct. 2020.

[18] *Chicago Data Portal*. Accessed: Nov. 2, 2019. [Online]. Available: https://data.cityofchicago.org/Public-Safety/Crimes-2001-topresent-Dashboard/5cd6-ry5g

[19] *Los Angeles County GIS Data Portal*. Accessed: Nov. 2, 2019. [Online]. Available: http://egis3.lacounty.gov/dataportal/?s=crime

[20] L. Lochner, "Education and crime," in *The Economics of Education: A Comprehensive Overview*, S. Bradley and G. Green, Eds. New York, NY, USA: Academic, 2020, pp. 109–117.

[21] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita, "Self-exciting point process modeling of crime," *J. Amer. Stat. Assoc.*, vol. 106, no. 493, pp. 100–108, Mar. 2011.

[22] J. H. Ratcliffe, "A temporal constraint theory to explain opportunity-based spatial offending patterns," *J. Res. Crime Delinquency*, vol. 43, no. 3, pp. 261–291, Aug. 2006.

[23] M. S. Gerber, "Predicting crime using Twitter and kernel density estimation," *Decis. Support Syst.*, vol. 61, pp. 115–125, May 2014.

[24] M. Traunmueller, G. Quattrone, and C. Capra, "Mining mobile phone data to investigate urban crime theories at scale," in *Social Informatics* (Lecture Notes in Computer Science), L. M. Aiello and D. McFarland, Eds. Cham, Switzerland: Springer, 2014, pp. 396–411.

[25] P. Kump, D. H. Alonso, Y. Yang, J. Candella, J. Lewin, and M. N. Wernick, "Measurement of repeat effects in Chicago's criminal social network," *Appl. Comput. Informat.*, vol. 12, no. 2, pp. 154–160, Jul. 2016.

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

[27] C. Catlett, E. Cesario, D. Talia, and A. Vinci, "A data-driven approach for spatio-temporal crime predictions in smart cities," in *Proc. IEEE Int. Conf. Smart Comput. (SMARTCOMP)*, Taormina, Italy, Jun. 2018, pp. 17–24.

[28] S. N. Christian, K. R. Majeed, and S. O. Etinosa, "Application of data analytics techniques in analyzing crimes," in *Proc. SAIS*, vol. 40, 2018, pp. 1–7.

[29] C. Schnell, A. A. Braga, and E. L. Piza, "The influence of community areas, neighborhood clusters, and street segments on the spatial variability of violent crime in Chicago," *J. Quant. Criminol.*, vol. 33, pp. 469–496, Sep. 2017.

[30] G. Rosser and T. Cheng, "Improving the robustness and accuracy of crime prediction with the self-exciting point process through isotropic triggering," *Appl. Spatial Anal. Policy*, vol. 12, no. 1, pp. 5–25, Mar. 2019.

[31] M. L. Young and A. Hermida, "From Mr. and Mrs. Outlier to central tendencies: Computational journalism and crime reporting at the Los Angeles times," *Digit. Journalism*, vol. 3, no. 3, pp. 381–397, May 2015.

[32] C. Contreras, "A block-level analysis of medical marijuana dispensaries and crime in the city of Los Angeles," *Justice Quart.*, vol. 34, no. 6, pp. 1069–1095, Sep. 2017.

[33] C. B. Dierkhising, D. Herz, R. A. Hirsch, and S. Abbott, "System backgrounds, psychosocial characteristics, and service access among dually involved youth: A Los Angeles case study," *Youth Violence Juvenile Justice*, vol. 17, no. 3, pp. 309–329, 2018.

[34] G. Ridgeway and J. M. MacDonald, "Effect of rail transit on crime: A study of Los Angeles from 1988 to 2014," *J. Quant. Criminol.*, vol. 33, no. 2, pp. 277–291, Jun. 2017.

[35] M. Valasik, "Gang violence predictability: Using risk terrain modeling to study gang homicides and gang assaults in east Los Angeles," *J. Criminal Justice*, vol. 58, pp. 10–21, Sep. 2018.

[36] M. R. D'Orsogna and M. Perc, "Physics for better human societies: Reply to comments on 'statiscal physics of crime: A review,"' *Phys. Life Rev.*, vol. 12, pp. 40–43, Mar. 2015.

[37] T. Almanie, R. Mirza, and E. Lor, "Crime prediction based on crime types and using spatial and temporal criminal hotspots," *Int. J. Data Mining Knowl. Manage. Process*, vol. 5, pp. 1–19, Aug. 2015.

[38] S. Seo, H. Chan, P. J. Brantingham, J. Leap, P. Vayanos, M. Tambe, and Y. Liu, "Partially generative neural networks for gang crime classification with partial information," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc. (AIES)*, New Orleans, LA, USA, Dec. 2018, pp. 257–263.

[39] G. Mohler and P. J. Brantingham, "Privacy preserving, crowd sourced crime hawkes processes," in *Proc. Int. Workshop Social Sens. (SocialSens)*, Orlando, FL, USA, Apr. 2018, pp. 14–19.

[40] D. S. de O. Santos Júnior, J. F. L. de Oliveira, and P. S. G. de Mattos Neto, "An intelligent hybridization of ARIMA with machine learning models for time series forecasting," *Knowl.-Based Syst.*, vol. 175, pp. 72–86, Jul. 2019.

[41] J. C. B. Gamboa, "Deep learning for time-series analysis," 2017, *arXiv:1701.01887*. [Online]. Available: http://arxiv.org/abs/1701.01887

[42] M. Khashei and M. Bijari, "A novel hybridization of artificial neural networks and ARIMA models for time series forecasting," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 2664–2675, Mar. 2011.

[43] L. W. Kennedy, J. M. Caplan, E. L. Piza, and H. Buccine-Schraeder, "Vulnerability and exposure to crime: Applying risk terrain modeling to the study of assault in Chicago," *Appl. Spatial Anal. Policy*, vol. 9, no. 4, pp. 529–548, Dec. 2016.

[44] T. Altameem and M. Amoon, "Crime activities prediction using hybridization of firefly optimization technique and fuzzy cognitive map neural networks," *Neural Comput. Appl.*, vol. 31, no. 5, pp. 1263–1273, May 2019.

[45] H. Wang, H. Yao, D. Kifer, C. Graif, and Z. Li, "Non-stationary model for crime rate inference using modern urban data," *IEEE Trans. Big Data*, vol. 5, no. 2, pp. 180–194, Jun. 2019.

[46] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: A review," *Data Mining Knowl. Discovery*, vol. 33, no. 4, pp. 917–963, Jul. 2019.

[47] S. L. Hyland, M. Faltys, M. Hüser, X. Lyu, T. Gumbsch, C. Esteban, C. Bock, M. Horn, M. Moor, B. Rieck, M. Zimmermann, D. Bodenham, K. Borgwardt, G. Rätsch, and T. M. Merz, "Early prediction of circulatory failure in the intensive care unit using machine learning," *Nature Med.*, vol. 26, no. 3, pp. 364–373, Mar. 2020.

[48] M. Abdel-Nasser and K. Mahmoud, "Accurate photovoltaic power forecasting models using deep LSTM-RNN," *Neural Comput. Appl.*, vol. 31, no. 7, pp. 2727–2740, Jul. 2019.
[49] P. Filonov, A. Lavrentyev, and A. Vorontsov, "Multivariate industrial time series with cyber-attack simulation: Fault detection using an LSTM-based predictive data model," 2016, *arXiv:1612.06676*. [Online]. Available: https://arxiv.org/abs/1612.06676
[50] M. Alsharif, M. Younes, and J. Kim, "Time series ARIMA model for prediction of daily and monthly average global solar radiation: The case study of Seoul, South Korea," *Symmetry*, vol. 11, no. 2, p. 240, Feb. 2019.
[51] S. Siami-Namini and A. S. Namin, "Forecasting economics and financial time series: ARIMA vs. LSTM," 2018, *arXiv:1803.06386*. [Online]. Available: http://arxiv.org/abs/1803.06386
[52] A. J. Hussain, P. Liatsis, M. Khalaf, H. Tawfik, and H. Al-Asker, "A dynamic neural network architecture with immunology inspired optimization for weather data forecasting," *Big Data Res.*, vol. 14, pp. 81–92, Dec. 2018.
[53] S. Benabderrahmane, N. Mellouli, M. Lamolle, and P. Paroubek, "Smart4Job: A big data framework for intelligent job offers broadcasting using time series forecasting and semantic classification," *Big Data Res.*, vol. 7, pp. 16–30, Mar. 2017.

WAJIHA SAFAT received the undergraduate degree in computer science from Fatima Jinnah Women University, Rawalpindi, Pakistan, in July 2015. She is currently pursuing the M.S. degree in computer science with COMSATS University, under the supervision of Dr. Sohail Asghar. After graduation, she served as a Design Engineer with Emerging Systems Software House. She worked as Graduate Research Assistant during her M.S. Her current research interests include computer science application, knowledge extraction, big data, sustainable development, and societies based on data mining and machine learning.

SOHAIL ASGHAR (Member, IEEE) received the degree (Hons.) in computer science from the University of Wales, U.K., in 1994, and the Ph.D. degree from the Faculty of Information Technology, Monash University, Melbourne, Australia, in 2006. In 2011, he joined the University Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi, as a Director. He is currently a Professor and the Chairman of computer science with COMSATS University Islamabad. He has taught and researched in data mining, including structural learning, classification, and privacy preservation in data mining and text and web mining, big data analytics, data science, and information technology areas. He has published more than 150 publications in international journals and conference proceedings. He has consulted widely on information technology matters, especially in the framework of data mining and data science. He is a member of the Australian Computer Society (ACS) and the Higher Education Commission Approved Supervisor. He has served as a Program Committee Member of numerous international conferences and regularly speaks at international conferences, seminars, and workshops. In 2004, he received the Australian Postgraduate Award for Industry. He is on the Editorial Team of well-reputed scientific journals.

SAIRA ANDLEEB GILLANI received the Ph.D. degree from the Corvinus University of Budapest, Hungary, in 2016. She is currently a Senior Assistant Professor with the Department of Computer Science, Bahria University Karachi Campus. She is working in NLP, text mining, and data mining domain.