Review On Cybercrime Prediction with Data Mining Techniques

Farhat Bibi[1]& Dr. Hamid Ghous[2]

Abstract-Cybercrimes identify cases of alleged crimes and federal crimes involving computers or communication instruments as goals, commissioned tools, and correlated with the prevalence of new technologies. Child porn, cyberbullying, identity theft, cyber fraud, credit or debit card theft, cybercrime, privacy violations, graphic violence, malware and other cyber hacking could be common types of cybercrimes. These types of cybercrimes often lead to infringement of the privacy of users, breach of security, loss of business, money laundering, or harm to public and government assets. Therefore, this paper discusses cybercrime identification and prevention strategies thoroughly. This paper discusses, the latest cybercrime prevention and detection approaches with supervised learning, unsupervised learning and Hybrid techniques. It addresses the attributes objectively and evaluates the weaknesses of each approach critically. As a potential review, the paper presents guidelines for the implementation of a cybercrime classification algorithm in which, compared to current techniques, cybercrime can be efficiently detected.

Keywords: Cybercrime detection methods, Data mining, K-mean clustering, Machine learning, Neural Network, Cyber Security, cyberbullying.

I. INTRODUCTION

Cybercrime can be defined like any crime carried out using a computer or other communications platform to give people fear and alarm, or to hurt, damage, and destroy property. Cybercrime can be defined in two way, one of them is computer-assisted and the other is computer focused. Crimes including in computer-assisted are child porn, theft, cyberbullying and money laundering. Whereas website defacement, hacking and phishing are included in focused cybercrimes[1].

In many ways, it is difficult to find correct and official cybercrime data because there are undeclared events, social hurdles and lack of information about the crime. In these cases, police force plays a major role as it regulates the amount of information that is published.

First cyberattack on a computer was made in 1960, where computer modules have been reproduced. . Many fraudulent activities and theft by deception are revealed after 1970, when over $1.5 million is defrauded from consumer accounts by a bank teller at New York's Union Dime Savings Bank. The first network with packet switching technologies and the TCP/IP protocol is a creeper virus created by Bob Thomas in 1971 to infect the Advanced Research Project Agency Network (ARPANET) networks. Imperial Chemical Industries (ICI) servant stole a lot of information from the company's computer and their backups in early 1971. The servant demanded 275000 pounds as a ransom. The first computer worm is created by Robert T. Morris in 1988 at the Massachusetts Institute of Technology via a computer (MIT). In 1994, in Russia, Finland, Israel, Germany, the United States, the Netherlands, and Switzerland, Russions hackers used this method and transferred a heavy amount from city bank to other accounts[1].

This paper literature review highlights research that have been performed to improve approaches to be used in cybercrime prevention and detection that have many kinds of techniques. Statistical techniques are used in this method, focused on evaluating and extracting data, themain purpose of this method is to evaluate the data, extracting data from study data in order to successful method for detecting cyberattacks. These methods also used some techniques of machine learning given input data is necessary for its predicting outcomes. Another group which locate the cybercrime and gave its solution. These developed techniques are also good for fuzzy logic and genetic algorithm. These techniques can overcome the situation of false alarm during cybercrime attack. Priori algorithm is another strategy which is also used for data mining algorithms to detect cybercrime. Many researches have been made to examine and evaluate these methods. These methods developed for cybercrime attacks. However, the latest analysis research focused exclusively on investigating methods of detection that are restricted to one or more cybercrimes, such as cyberbullying, fake profiles, phishing, email spam, or botnet [50].

SIGNIFICANCE

A critical evaluation of cybercrime detection strategies using different methods of classification is presented in this paper. This study discusses cybercrime identification and prevention strategies thoroughly. This paper discusses the latest cybercrime prevention and detection approaches with supervised learning, unsupervised learning and Hybrid techniques. This paper offers a thorough analysis of various methods for detecting cybercrimes that have been carried out. To check its validity and efficiency

according to accuracy response time and disadvantages of the evaluated techniques are compared and analyzed. The paper presents some suggestions to enhance the efficacy of the detection and to improve the methods that restrict the precision of the prediction.

II. CRITICAL LITERATURE VIEW

In past many researchers have worked on cybercrime detection by using different methods of supervised and unsupervised machine learning algorithms. Supervised learning provides Support Vector Machine (SVM), Linear Regression, Naïve Bayes (NB), Decision Tree (DT), K-Nearest-Neighbor (KNN), Logistic Regression, Linear Discriminant Analysis (LDA), Neural Network (NN) and Deep Neural Network (DNN). Unsupervised learning algorithms provides K-mean Clustering, Apriori algorithm, Principle Component Analysis (PCA), Singular value Decomposition and Independent Component analysis (ICA).

A. SUPERVISE LEARNING

Mrs. Prithi S. et al. have developed a model using a training dataset that has go through data cleaning, data transformation and reduction of data by using sampling and correlation is the principle of machine learning.The study predicts accuracy by comparing the results of various supervised algorithms for machine learning. Data cleaning and preparation, missing meanings, experimental analysis and finally model creation and evaluation begins the analytical process by using python. To predict a value, the Logistic Regression(LR) algorithm also uses a linear equation withprediction model. After that, a comparison is made with other methods like Logistic Regression (LR), RF, KNN, SVM and DT. The Logistic Regression achieved higher precision prediction resulted by comparing the better accuracy [2].

The aim was to provide requisite broad awareness of cybercrime attacks to the existing framework in a community, to allow them to recognize the possible threats of such attacks and to prevent cybercrime offenses from being exhibited. Feature selectionis used for pre-processing the data. The major objective of the work is to detect crimes that take benefit of security weaknesses and use machinelearning techniques to analyze these threats. Three classification algorithms are used: RF, SVC and NB. Accuracy rate of these algorithms are: Logistic Regression 0.9938%, Linear SVC 0.9923%, Multinomial NB 0.9895% [3].

The rise in financial, psychological, cultural, social, political and security harm is caused by cybercrimes. The results of these studies show that in order to use artificial neural networks, where cybercrime is directly linked to the rise in crime in society, the distance between the theory and implementation must be minimized, in particular, in the police area. This research can be very useful and realistic, as the evidence used in the field of cybercrime using artificial intelligence has been defined. Three classification algorithms are used: SVM, RF and DT [4].

Data analysis algorithms provided the best recovery score of 31.71%, which is really bad of crime dataset. To pre-process the dataset, the Python Library Sklearn is used. Therefore, study split the thirty nine classes into two categories: One is a normal, and the other is an un-common category. Study used methods of over-sampling and under-sampling to resolve the imbalanced problem. After that, a comparison is made with other methods like DT, KNN, Adaboost and RF. The best decision making training set was performed by RF with an accuracy of 99.16% compared to other machine learning agents [5].

The findings indicate that Point of Interest (POI) characteristics are incredibly helpful for attack detection, allowing lower and higher areas to be accurately distinguished.Four classification algorithms are used: SVM, Logistic Regression, DT and RF. The positive results obtained with classification techniques get an opportunity to evaluate other forms of modelling using crime reports and POI features, such as logistic regression for crime count estimation or clustering to identify similarity trends between micro-areas with regard to the occurrence of POI and crime [6].

The objectives of this research could be used to build grid-based crime prediction methods and data design features for classification learning and to facilitate the modelling of police department's knowledge and theory of criminology.The pre-processing of the data is done by using feature extraction F1 Score. Study used different supervised learning algorithms: DNN-tuning, SVM, RF, and KNN. Accuracies of these machine learning algorithms are: DNN-tuning 0.8376, SVM 0.8810, RF 0.8197 and KNN 0.8706 [7].

In the Indonesian area, if it has a criminal database which can be evaluated since the features used in this study are attributes that also exist in Indonesia, the development of Linear Regression for predictions can be applied.The aim is to use a linear regression algorithm for crime analysis data to produce predictions for crime, which shows a very accurate outcomes. These findings suggest that the algorithm for Linear Regression is successful and could be used for prediction [8].

The DT technique can determine that populations are indeed likely to be vulnerable by cybercrime, since it requires firstly, a process of learning knowledge, in this case cybercrimes that have already been performed in different populations in the US, and then identifies trends of greater and lesser recurrence through its own algorithm.

Eventually, it includes a prediction method that takes these recurrence trends and, with its own method, sets the probability that cybercrime may affect a society [9].

RF Classifier giving the most structured results for the prediction of Per Capita Violent Crimes functions in terms of accuracy, recall and F1 score out of three models. RF model takes several trees into account and produces an average result that has proven to be suitable for this data form. As it had values similar to the RF Classifier, NB proved to be a balancing quotient for this crime data. Classification, accuracy, generalization and error reduction increases efficiency by providing appropriate preparation and evaluating samples that seemed to assist in this study by providing accurate and reliable efficiency. Four classification algorithms are used: DT, RF, NB and Logistic Regression. Accuracy of clean data by DT 75.90%, RF 83.39%, NB 77.64% and Logistic Regression 64.72%. Accuracy of dirty data by DT 76.77%, RF 81.35%, NB 75.42% and Logistic Regression 66.93% [10].

Research also shows that a cybercrime classification method can be developed using features from the processing of natural language and cybercrime related psychological factors. To determine characteristics that provide better pre-processing variability, study also use Principal Component Analysis (PCA) and normalization. Three machine learning algorithms are used SVM, NB and IBK. Experiments show that to identify texts describing cybercrime, it is possible to use text-based fraud detection. Accuracies of various machine learning algorithms are: SVM is 40%, NB is 60%, and IBK is 50%. In future researcher, can create model in various web genres and make assumptions, the detection of fraud in texts and emails in other web domains [11].

Fateha Khanam Bappee et al. focuses on datasets from the real world, another significant issue is the topic of data discrimination. Data discrimination refers to prejudice that exists due to inconsistencies between various sources of data. Using four distinct types of crime, the new features are tested using only the details contained in the UCR forms as characteristics for a classification as the benchmark. Four classification algorithms are used: LR, RF, SVM and Ensemble. The findings demonstrate that when the recently designed features are applied to the validated classification methods, significant enhancements in accuracy and AUC are identified. For alcohol and motor vehicle offences dependent on based features, the Ensemble method produces an AUC score of 82.5 percent & 69.4 percent [12].

The tool authors have developed and provided a platform for visualizing and analyzing crime structures, using Google Maps and different R-packages with two different machine learning algorithms: NB and KNN. By means of different data visualization, the project allows crime researchers to analyze these crime structures. Collective and visual function technologies will be beneficial in tracking and exploring the nature of crime. It is possible to consider and compare several classification models in the study. It is clear that law enforcement agencies will take great advantage of the use of algorithms for machine learning to combat crime and save humanity. The study want to upgrade data as soon as possible, using current developments such as internet and device, for better performance. Chance of reporting is 37.5% and chance of Burglary is 10.7% [13].

Ms. Vrushali Pednekar et al. focused on a special day, the designed methodology predicts crime prone regions in India. If the research consider a specific province, it will be more precise. Another issue is that the system will not predict the time at which the crime occurs. As some time is a critical aspect in crime, the study must expect not only the areas that are vulnerable to fraud, but also the right time. Taking into account the methods proposed for predictive modeling, it demonstrates that specifications such as the impact of anomalies in pre-processing data mining, the quality of data training and validation, and the value of attributes have not previously been discussed. The study used KNN algorithm for crime prediction [14].

Study have suggested two objects, a structure for data processing and a model of classification. The study have performed an ex-ante assessment of the accuracy of their classifiers and a former assessment of its execution using model implementations. The study used NB algorithm for the proposed model. In order to have a better insight into the possible drawbacks, future work might classify terms and threats by the sector, and it might try to find the network impact [15].

Hamid ZolfiEt al. addressed and introduced pre-processing and normalization. SVM, NB, DT, and LR are used for the execution of the techniques together. Each of these techniques are applied and in various modules the results obtained are given. SVM is the best classification technique with 99% accuracy, which provided reasonable accuracy for cybercrime identification in cyber threats. Accuracy of various algorithms are: NB 84%, DT 80%, Logistic Regression 63%, SVM 99%. Hence SVM provided best accuracy [16].

Performance measures of machine learning KNN and increased DT are deployed and when predicting crime in Vancouver, a crime predictive performance of 39% to 44% is achieved. For several methods and algorithms, the precision, efficiency, and practice time of algorithms are completely different. While as a prediction model, this model has low accuracy, it offers a scope management and includes the processes for further studies. Accuracy and

practice time for method 1 is 41.9% and 903.63 sec respectively, and with 459.26 sec preparation time, method 2 is 43.2% accurate [17].

As mentioned in the research, a variety of classification techniques are studied and the results in the evaluation phase from which the study preferred to use the J48 method along with its success in applying it to the data gathered. Using the Waikato Framework for Information Analysis WEKA Tool Kit, the study developed and trained a J48 classifier on a preprocessed crime dataset. From the observational data, the J48 methods have been applied with 94.25287% accuracy of the unspecified classification of crime reports, which is correct just enough to focus on the system for crime prediction and also takes minimal time to process, compared to other classification techniques [18].

Mehmet Sait Vural Mustafa Gok developed framework can be used in criminology to help security forces find the offender of the crimes due to the simplicity induced by the NB assumptions of independence. In addition, the model is relatively compatible with the apparent inconsistency in that the product meets an 80% rate reduction in the death list. The investigational findings indicate that with its average of 78.05% chance of getting with its emerging technologies for both the generation of crime datasets and the decision making method, the study encourages further work on the criminal prediction issue [19].

CNN is indeed a "black box" in which the mechanisms of the neuron relation are not based on forecasts. Research developed a dynamic model for feature selection based on dynamic CNNs. Consequently, instead of going to wait for the "black box" day when this study is open, desperate to take a quick step to apply deep CNNs to the analysis of spatial and temporal crimes for the early detection of crimes present, the study assume that this work only contains the substrate of what is feasible in this position, and there are several mechanisms for more investigation, such as console application types of crimerisk, effectively promoting social and economic characteristics to stop crime. The developed scheme had the best classification efficiency [20].

The crime type categories are: aggression, offences, theft and crime clustering using K-means to capture the information [43]. In order to get the data across the web, this approach is faster. Successful web mining is to get the unstructured data into structured data. The study assume that there is a great career for crime data mining to increase the efficiency of criminal and intelligence analysis. Four classification algorithms are used: SVM, DT, ANN and NB. For instance, inquiry methods for crime patterns and network visualization can be developed for more visual and intuitive crime and intelligence [21].

NB are used to predict the probability while association rule mining are used to split data. Three classification algorithm are used: J48, NB and JRIP. The ability to extract useful information from large databases is data mining. So for this reason, data mining can be used. The choice of appropriate methods for data mining has a greater impact on the results achieved. This is the primary explanation behind the comparing results and the evaluation of the top performing algorithms for data mining [22].

Six machine learning methods are formulated and solved to predict the incidence of crime hot-spots in a town in China's southeast coastal region [47]. These six algorithms are: KNN, NB, CNN, RF and SVM. The LSTM model's prediction accuracy is higher than that of the other models. Further enhancing the prediction accuracies of the LSTM model is the inclusion of urban buildingperformance regression. Utilizing historical crime data on the experimental data is higher than those of the original model. Compared to other methods, the research models have enhanced prediction accuracy. Accuracy of six machine learning algorithms is improved from 46.6% to 52.3% and the accuracy of LCTM model from 57.6% to 59.9%. LCTM model is better than others [23].

Evaluating the accuracy rate of 4 data mining techniques with different initial conditions in a structured way, the research have established a formal organizational IM authorship research framework. The research used C4.5, DT and K-mean for proposed model. The study also established a holistic categorization of the IM specific attributeset that can be conveniently used in future studies. For Dataset 1 “19 authors” and Dataset 2 “25 authors” are taken simultaneously, the experiments realize authorship recognition prediction precision of 88.42 % and 84.44 % respectively. By extending the datasets, developing other different classifiers and using author analysis methods, study continued this research to narrow the area of suspects in a crime detection [24].

LIMITATIONS

After reviewing the literature of supervised machine learning algorithms, some limitations are found that are: some papers have not good accuracy, some need to check on real time.

B. UNSUPERVISE LEARNING

Cross-type Correlation leverages heterogeneous broad urban data, such as data on crime complaints, search and strip search data, weather data, data on Point of Interest (POI), data on human mobility and 311 data on public service concerns. With systematic experiments focused on real world urban data from New York City, the study test the structure. Comparison is made with the different

methods like ARIMA, VAR, RNN and Deep ST. The research suggests that various types of crime are significantly associated with each other. In the near future, the proposed system can reliably forecast crime amounts, and cross type and spatiotemporal correlations can improve the prediction of crime [25].

Sergio Pastrana et al. have developed instruments to identify and forecast actors engaged in cybercrime operations. Similar methods quickly locate user accounts that may enable more analyses by online groups tracking law enforcement and security firms and also for early implementation of new measures to protect or modification of new sites. The study have proof of these main actors' online social interactions and the study discovers different traditional positions for these key actors. The study use Logistics Regression and K-mean clustering for proposed method. Use topic analysis extraction and NLP tools for pre-processing the data. The methods used during the whole analysis are available publicly in the github repository [26].

In cluster one, Surigao City is the town in Surigao Del Norte with the highest number of index and non-index crimes reported. The municipalities of Placer, Claver and Dapa have the highest crime rate in Cluster two. In Cluster three, meanwhile, the research identified Monica and Pilar. Theft is classified as the highest number of registered crimes among the index crimes in the province of Surigao del Norte, with a total of 2,565, with the objectives and hypothesis. The serious trauma with the expected values of 2,508 or 26% rise is the highest predicted crime for the year. In the region, the least reported crime is livestock creaking [27].

The list of crime levels over the years had offences involving “visitors”. Crime categories like robbery and stealing have different classes that have a very good linear association when applying the K-means algorithm to the 2015 crime data collection. In comparison, population and crime totals form closely linked categories for the year 2015. There are also non-strongly connected crime categories that form K-classes that do not have a positive correlation [28].

The FCM algorithm operates by grouping an individual data point into several clusters. The official outcome is used to analyze the states, vulnerable to Crime in the US so that it can be avoided by raising the level of protection in those areas. The findings are only useful for the detection of crime, so there is a need to examine the trends of crime that might arise in the future. It is difficult to foresee crimes, but it can be avoided if the time that the crime is going to occur is known. In near future, along with the research structure, the statistical method of inevitable crime will be carried out using K-means clustering [29].

By applying the new data mining algorithms such as K-means, Inspired Association Classifying with Prediction J48 tree, the proposed model produces a superior concept over cybercrime prediction. An enhanced model for association and connection with measured help and confidence measures is the Affected Association Classification. The K-means algorithm bundles item sets from cybercrime datasets. With K-means, Affected Associate Identification with Prediction Tree J48, it is possible to boost the classification convention and accuracy. Interruption instruments should be built anywhere it is feasible and tested on a standard b era [30].

The ARIMA model is widely used to make forecasts, but it is predictable. RNN model for cybercrime violation detection and evaluation of the pattern of the incident. The findings reported in this research are associated with validated attack data on the reliability of the SARIMAX and Compared to the SARIMAX model, the prediction accuracy of the RNN model, both of the SARIMAX and RNN models can achieve a reasonable forecast in terms of actual situations. There are several interesting aspects that are available for serious investigation [31].

K-mean clustering is used for data partition. Data classification is used specifically to distinguish between the types of safety practices to be used for each crime [44]. Different offences need different care and using this technology can be done easily. Investigations indicate that the method is aware of the importance of speed of research, detection of prevalent trends of crime and areas vulnerable to crime for future prediction. In the modern dynamic crime situation, the established system has interesting value and can be used as an effective instrument by Indian police and police forces for the classification and prediction of crime [32].

LIMITATIONS

The limitations of above work are: that many researchers observed single form of behavior that could not detect crime accurately and some focus on external sources to identify key actor that is cause of biased results.

C. HYBRID TECHNIQUE

Supervised based classification is applied by making use of Network Bayesian classifiers like RF, KNN, DT, LR, and K-mean. The study calculated at which iteration the best valid output is obtained using the Bayesian, Levenberg and weighted algorithm on train and test data, and it is observed that the scaled algorithm provided the good result compared to the other two, for the data considered. Statistical analysis is carried out on the basis

of correlation, ANOVA and graphs. Accuracy of prediction model is 168.9142% [33].

Hybrid approach is used for the proposed system. Understanding the connection between the expertise of the analysis and the characteristics of the type of cybercrime will make it easier for investigators to use these methods more efficiently to identify trends and patterns, fix problem areas, and even anticipate potential cybercrimes. To allow the system to deal with the continuously evolving nature of crimes, the detection mechanism should be customizable. Measures of similarity are a significant aspect that helps to find outstanding crimes in the pattern of crime. Only a sample of the different data mining techniques for cybercrime detection in various fields has been presented in this research [34].

Five classification algorithms are used: NB, K-mean, Artificial Neural Network, DT and DNN. The set of data to be extracted is huge, so a vital task is the pre-processing and processing of null values. The Artificial Neural Network and DNN can be useful to detect and analyze crime using past crime databases in order to predict future crimes from the enormous amount of data. Data mining algorithms can be preferred when data is monitored, and deep learning techniques can be used when data is multi model, enormous and unmonitored [35].

Highly developed data mining and Artificial Intelligence techniques are widely accessible to the legal community as computer science and technology have advanced. The current study focuses on creating an Indian case crime optimization algorithm using various data mining techniques that can assist the police agency to tackle violence properly. The proposed instrument helps departments to clean, classify, and evaluate crime data quickly and economically to identify measurable trends and patterns. The previous analysis is made as 83% of the crime. The experimental algorithm measured a crime rate of 89% [36].

Data mining procedures and algorithms are applied for pre-processed data in order to detect or predict fraud and remove noisy, incomplete, missing values. Three classification algorithms are used: DT, K-means and clipping method. This study introduces a descending algorithm for Big Data anomaly detection using K-means algorithms. To detect the anomalies posed in the monitored and unmonitored data collection, the suggested algorithm is used. With the clustering method, using big data analytics reduces the investigative time and helps to recover the secret information[37].

D. MISCELLANEOUS

The proposed approach collects quantitative data, including the London police website, from web sources. This model helps and speeds up a phase in which the rate of crime will rise. Four classification algorithms are used: ARIMA, SVM, Logistic Regression and DT. The best fit model for crime rates was introduced in London in this analysis. The relation between the previous crime rate and the predicted crime rate performed well, and actual scenarios also use this statistical record. Higher risks of crime are indicated by the Upper Trust Limit UCL. The lower confidence point indicates the smallest probability of crime. The best validation performance is 148.5423 at epoch 398 [38].

To find the best paths, the meta-heuristic Genetic Algorithm GA and Cuckoo Search CS are implemented. Using real world datasets, the output of the proposed solution is checked and contrasted with several state of art approaches. This work leads to the creation of an efficient police monitoring strategy to deter potential incidents of crime and emergency situation events. The given model also responds to the unpredictable arrival of emergencies and conditions that a police officer is expected to attend. The results in this study indicate that the planning of path optimization for multiple officers can be achieved more accurately by fusing crime incident prediction and real time emergencies. The ARIMA method's precision is best for predictive modeling. Matching predict value with actual value is 80% correct [39].

In the ST-Cokring algorithm, time series historical crime data are used as the primary outcome measure, while intermediate zones derived from VIIRS nightlight visualization are used as a secondary co-variable to boost street crime forecasting. A user friendly software tool has been developed with the ST-Cokriging algorithm. Effective data models and rapid computational methods have been integrated into the Execution of the process of ST-Cokriging. It is also important to crime risk prediction and hotspots in other cities, offering theoretical and technological support for decision making on the implementation of police forces [40].

In addition, Colorado Springs' expected hotspots display consistent spatial trends with the trends of hotspots observed. Applying the NIJ award winning forecasting algorithm indicates that even if the original algorithm is factual, the new decision making is still valid dependent on the crime data of a particular city. In addition, even if there is a shift in the type of data relative to what is based on the original algorithm, the forecast results show high precision and performance. The interest of P1P Crime hot-spots is expected to increase quickly from 37.6% to 61% and overtake the output of other forms of crime [41].

Research finds that both regular activity locations and Bluetooth adapter regions demonstrate a positive and significant impact, although the effect size observed for

pathways is greater. By improving the interpretation of eyes on the street activity and seeking conclusive proof of crime pattern, the study contributes to crime theory. In addition to crime forecasting, the study demonstrated the predictive capacity of human movements from online location tracking. Exponential distribution coefficients and the related 95 % confidence intervals from the PGLM model are visualized [42].

The main and theoretically most significant consequences of the study, particularly for police professionals, are best summarized in a single statement. If the researcher want to try to predict crime locations, the use of the model in Microsoft Excel has shown comparable levels of efficiency and precision, with lower economic or fiscal commitments and greater clarity than other more expensive ones. P1P Crime produces better outcomes than P1V Crime. In both efficiency and precision, P1P crime shows the best forecasting results. High accuracy range is between 82% and 88% [43].

Closely related to the other forms of conventional crime where a plethora of prevention and detection methods have been used for a long time, it is time to start looking at cybercrime [19]. One of them is geographic profiling, a methodology initially created in criminology, where the implications of regional profiling of cybercrimes for the prediction of serial crime locations have been tested in this study. The profiling process included the generation of the geographical profile from incident reports, model verification and configuration, model parameter responsiveness checking, measurement of accuracy and graphical visualization [44].

Study used classification and others algorithms like SVM, Ground Truth and SMOTE. Then again, for many of the capacities, processes and services identified in well-known supply chain, there is no availability. The research have evaluated and summarized the developments in the standardization of cybercrime on anonymous internet markets in terms of the generalizability of the results. Moreover, the outcomes only indicate expensively that the trend towards standardization may not be quite as systematic as it has been claimed internationally. The results only indicate imprecisely that the trend towards increased availability might not be as systematic as it has been described everywhere [45].

The study revealed that, by exploring the Prophet model, a neural network model, and the deep learning system LSTM, both the Prophet model and the LSTM method performed better than conventional machine learning algorithms. The research plans to conduct more realistic case studies in the near future, further evaluating the effectiveness and reliability of the different models in the method [46].

Study used Deep Random Forest and Deep learning for proposed model. The study used meaningful model network databases with several forms of attacks in this study, which are highly dimensional. The analysis became twofold after the study addressed the classification problem by using the re-sampling method. For each type of attack, thestudy individually designed unique predictive model and developed the best models with the required accuracy. The research further configured the templates to have the best and most realistic results for accuracy [47].

To classify crime hot-spot places, crime visualization and analysis tools are used. To evaluate the hot-spot position more roughly, implementations of the Radial Basis function and Triangular with Linear interpolation approach are combined. This combined strategy allows police staff to quickly analyze the hot-spots in a computerized way to more effectively secure the frequent areas of crime. The Linear Discriminant Evaluation tool can be used to accurately assess the most modern criminal hot-spots in future work [48].

TABLE

Title	Author	Dataset	Preprocessing	Methods	Results	Future Direction
"GUI BASED PREDICTION OF CRIME RATE USING MACHINE LEARNING APPROACH"	"Mrs. Prithi S, Aravindan S, Anusuya E,Ashok Kumar M" 2020	Crime dataset Obtained from Indian police department	-K fold cross-validation -Outlier remove and variable conversion have to be done. -Sampling -Correlation for data reduction and transformation	-RF -Logistic Regression -DT -KNN -SVC	To predict a value, the Linear Regression method also uses a linear equation with independent prediction. The logistic regression model is a higher precision prediction result by comparing the best accuracy.	In future, to optimize the work to be carried out in the environment of Artificial Intelligence. The future references that can be made are.
"Computatio	"Rupa Ch,	Cybercrime	Feature	-RF	Accuracy rate of	In the future, by

nal System to Classify Cyber Crime Offenses using Machine Learning"	Thippa Reddy Gadekallu , Mustufa Haider Abidi,and Abdulrah man Al-Ahmari" 2020	dataset collected from Kaggle and CERT-In 2000 records Attributes: Incident, harm, year, location, offender, victim, age of the offender and cybercrime.	extraction by using TFIDF or tf–idf vector method -Chi—squared for correlation -F1 Score	-SVC -Linear Regression	various algorithm Linear Regression=0.9938 % SVC=0.9923% Multinomial NB=0.9895% RF=0.8069 Proposed model accuracy is 99%.	using Deep learning methods in the forecast of crime cases region wide, the characteristics of the framework can be improved.
"Process Modeling and Extraction of Patterns of Computer Crimes Using Data Mining"	"Abbas Karimi, Saber Abbasaba dei, Javad Akbari Torkestani , Faraneh Zarafshan " 2020	Training dataset	-Feature extraction through text mining -Classification of textual documents using neural network, Specialty of matric structure	-SVM -Decision Tree -RF	Best validation performance is 0.24748 at epoch 11.	Writing their future theses and papers on cybercrime psychology, this study is useful these details
"Crime Prediction Using Spatio-Temporal Data"	"Sohrab Hossain, Ahmed Abtahee, Imran Kashem, Mohamm ed Moshiul Hoque, and Iqbal H. Sarker" 2020	Crime dataset, provide San Francisco open data. Contain 8,74,049 rows	Feature extraction by using Principal Component Analysis (PCA)	-DT -KNN -Adaboost -RF	After using these two techniques, machine learning devices can be extremely benefited. RF observe the best decision making classifier with a precision of 99.16 % than other machine learning agents. Accuracy: Oversampling=73.8 9% Under sampling=99.16%	In the future, the researcher want to enhance the accuracy in crime prediction. Moreover, seeking to combine the prediction of cybercrime with the prediction of real world crime.
"A Proposed Model for Cybercrime Detection Algorithm Using A Big Data Analytics"	"Hossam Abdel Rahaman" 2020	Cybercrime database Collected from various News feeds, articles, blogs, and police	-Association Rule mining -K Mean partition clustering for transformation	-K-mean Clustering Algorithm, -Clipping Method -DT	With the clustering method, using Big Data Analytics reduces the investigative time and helps to recover the secret information.	N/A

		department websites over the web internet.				
"Urban Crime Risk Prediction Using Point of Interest Data"	"Paweł Cichosz" 2020	-Crime Dataset -POI Dataset Obtained as shapefile available from link1	Feature Extraction By Principle Component Analysis	-RF -DT -LR -SVM	The positive values achieved with classification models provide an opportunity to explore other forms of modelling using crime reports and point of interest attributes.	Some spatiotemporal hot-spot detection methods and appropriate procedures for spatiotemporal prediction evaluation in future directions.
"Exploring Spatio-Temporal and Cross-Type Correlations for Crime Prediction"	"Xiangyu Zhao and Jiliang Tang" 2020	-Crime Complaint dataset collect from complaint frequencies of the aforementioned -Stop-and-frisk dataset collect from decline Urban crime -Meteorological dataset , -Point of Interests dataset -Human mobility dataset	Feature extraction of spatial correlation and temporal correlation by using ADMM algorithm, Feature vector, Power law exponential function.	-ARIMA (Auto-Regression Integrated Moving Average) -VAR (Vector Auto-Regression) -RNN (Recurrent Neural Network)	The results indicate that various forms of crime are inherently associated with each other, and the proposed system can reliably forecast crime quantity and cross-type and spatiotemporal correlations can improve the prediction of crime.	1-Cross-type and spatiotemporal correlations will improve crime prediction in the future. 2- For crime analysis, the researcher would like to incorporate and improve more sophisticated techniques. 3- In addition to the role of crime prediction, the researcher would like to design more complex models to solve the real worlds more realistic security challenges.
"Grid-Based Crime Prediction Using Geographical Features"	"Ying-Lung Lin, Meng-Feng Yen and Liang-Chih Yu" 2018	crime dataset	-F1 Score, -DNN auto Feature Extraction, -Min Max normalization range [0,1] is used to convert features	-DNN-tuning, -SVM -KNN -RF	Accuracy of different algorithms are DNN-tuning=0.8376 SVM=0.8810 RF=0.8197 KNN=0.8706	N/A
"Prediction Analysis Of Criminal Data Using Machine Learning"	"Meiliana, Dedi Trisnawarman, Muhammad Choirul	Crime Dataset from Los Angles Police Department	-Cleansing is done by using Rapid Miner to resolve missing value and noise. Linear regression	LR Algorithm	Accuracy of prediction model is 168.914%	N/A

	Imam" 2020	The dataset has 2.036.897 rows	for data prediction.			
"Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India"	"Shraddha Ramdas Bandekar, C. Vijayalakshmi" 2019	-Crime dataset collected from public domain data national crime records bureau	-Regression is used for data prediction -Visualization of data -Bayesian Neural Networks -Levenberg algorithm -Scaled algorithm Normalized the data.	-KNN -Boosted DT -LR -K-mean clustering -RF	The best validation performance is 148.5423 at epoch 398	The future scope is to develop this work to collect and apply an optimization model to enormous data and to obtain results based on comparative study of various ML algorithms.
"Forecasting Crime Using ARIMA Model"	"Khawar Islam, Akhter Raza" 2020	London crime (LC) dataset collect from 34 borough of London 4 boroughs of London "Barking and Dagenham, Barnet, Bexley and Brent"	-Microsoft Excel for data cleaning and processing -IBM SPSS for Crime data prediction -Linear Regression used for visualization	-ARIMA algorithm -SVM -LR -DT	The ARIMA method's precision is best for predictive modeling. Matching predict value with actual value is 80% correct.	Important data that will be used in the artificial neural network for future Crime prediction study.
"Multi-officer Routing for Patrolling High Risk Areas Jointly Learned from Check-ins, Crime and Incident Response Data"	"Shakila Khan Rumi, Kyle K. Qin, and Flora D. Salim" 2020	Crime Dataset Contain 50% crime data and 50% no crime Collected from sector in Seattle.	-Feature Extraction correlation of POI based feature. -Data sampling is done by encoding scheme -Random Forest for Classification	-Genetic Algorithm -Guided Genetic Algorithm (G-LERK-GD) -Greedy Algorithm -Cuckoo Search	The results in this study indicate that the devising of path optimization for numerous officers can be achieved more safely by combining crime incident prediction and real-time emergencies.	N/A
"A spatio-temporal method for crime prediction using historical crime data and transitional zones	"Bo Yang , Lin Liu , Minxuan Lan , Zengli Wang, Hanlin Zhou and Hongjie Yu"	Residential burglary dataset from Los Angeles, USA	-Kernel density function aggregated data -Transformation is done by covariance function	-ST-Cokriging Algorithm -Aggregation Method -Spatio-temporal covariance model	The outcome of this research is that the precision of crime prediction accuracy is also assessed by PAI and PEI. By modifying the threshold on the crime risk index, hotspots maps are evaluated.	N/A

identified from nightlight imagery"	2020					
"Flag and boost theories for hot spot forecasting: An application of NIJ's Real-Time Crime forecasting algorithm using Colorado Springs crime data"	"YongJei Lee, SooHyun O" 2019	Crime dataset provided by CSPD		-NIJ award winning Forecasting Algorithm, -Theory-driven algorithm	The interest of P1P Crime hot spots is expected to increase quickly from 37.6% to 61% and overtake the output of other forms of crime.	The temperature of each police break can change, measuring the reliability of the temperature. The researcher leave this risk of possible extrapolation to future studies.
"Leveraging Mobility Flows from Location Technology Platforms to Test Crime Pattern Theory in Large Cities"	"Cristina Kadar, Stefan Feuerriegel, Anastasios Noulas, Cecilia Mascolo" 2020	-Foursquare dataset data.sfgov.org, www.opendataphilly.org, data.cityofchicago.org -Crime dataset	Feature generation Check-in Pass-through-flows Computation of pass-through transition	-Schematic Crime Pattern Theory, -PGLM model, -Socio-demographic	Exponential distribution coefficients and the related 95 % confidence intervals from the PGLM model are visualized.	N/A
"A Theory-Driven Algorithm for Real-Time Crime Hot Spot Forecasting"	"YongJei Lee, SooHyun O, and John E. Eck" 2019	-CFS dataset provided by NIJ -Cincinnati crime dataset Select 83 Grid cell from CFS	F1 score	-KDE method - Theory-driven model	P1P Crime produces best outcomes than P1V Crime. In both efficiency & precision, P1P crime identify the best forecasting results. High accuracy range is between 82% and 88%.	In the Forecasting Algorithms, the study recommend that future research integrate these spatial and temporal correlations.
"Analysis to Predict Cybercrime Using Information Technology in a Globalized Environment"	"Segundo Moisés Toapanta Toapanta, Luis Enrique Mafla Gallegos, Bryan Eduardo Cisnero Andrade" 2020	Crime dataset 128 attribute each attribute contain 1994 records, UCI repository,	-Dimension Reduction method -For transformation Applying discretization and numbering method, -Use PCA for Normalization	-DT - Application Tree	The result of this study, is to classify the features that make them vulnerable to cyberattacks within the societies with the greatest effect and thereby be able to eliminate cybercrime.	The key criteria that devote to the success of cybercrime within societies, and which are the potential future societies that will be impacted of more tireless scope.

"Geographic Profiling for serial cybercrime investigation"	"Asmir Butkovica, Sasa Mrdovicb, Suleyman Uludagc, Anel Tanovicb" 2018	Spatial crime dataset	Standard Distance Deviation for visual indication	-Criminal Geographic Targeting(CGT) Algorithm	Outcomes of the implementation of these two geographical profiling methods, measuring the standard distance deviation from the original location of the perpetrator of the supposed anchor point or coordinates point.	N/A
"PREDICTIVE MODELLING OF CRIME DATASET USING DATA MINING"	"Prajakta Yerpude, and Vaishnavi Gudur" 2017	Crime dataset from UCI repository Total 1994 attributes, 128 attributes like population, age and race are used.	-2 Fold-Cross Validation for swap the roles and 10-Fold used for Analysis, -F1 Score -Transforming uses normalization	-DT -RF -NB -LR	Accuracy of clean data: DT=75.90% RF=83.39% NB=77.64% LR= 64.72% Accuracy of dirty data: DT=76.77% RF=81.35% NB=75.42% LR=66.93%	N/A
"A Text-based Deception Detection Model for Cybercrime"	"A.Mbaziira, and J.Jones" 2016	Enron email dataset which hold 500,000 emails. This dataset was made public by the Federal Energy Commission	-Use PCA and normalization to determine feature	-SVM -KNN - NB	Accuracy of various machine learning algorithms are: SVM=40% NB=60% IBK=50%	N/A
"Characterizing Eve: Analysing Cybercrime Actors in a Large Underground Forum"	"Sergio Pastrana, Alice Hutchings, Andrew Caines, and Paula Buttery" 2018	-CrimeBB Dataset That holds information about 572K users accounts	-F1 Score -TF-IDF for feature extraction -Tokenize data -Reduces noisy data -Removing stop word -Punctuation character -Part-of-speech Tagger	-LR Likelihood ratio method -K-mean clustering	Authors have developed instruments to identify and forecast actors operating in cyberattack operations. These sensors help to classify user accounts that may require further exploration via monitoring of online networks and law enforcement and	It is not convenient, even with manual review, to determine whether the predicted actors are actually engaged in criminal occupation.

					security firms.	
"Predicting Crime Using Spatial Features"	"Fateha Khanam Bappee, Amilcar Soares Junior, and Stan Matwin" 2018	Crime dataset obtained from Halifax regional police department	Spatial Feature selection Geocoder process used	-LR -RF -SVM	The results demonstrate that when the newly integrated technologies are applied to the modified classifier, substantial development in accuracy and AUC are noticed.	The ability to undertake transfer learning from what is learned in NS to other countries is another research path the researcher want to pursue.
"Crime Prediction & Monitoring Framework Based on Spatial Analysis"	"Hitesh Kumar Reddy ToppiRed d, Bhavna Saini, Ginika Mahajan" 2018	Crime dataset of U.K police Department Training dataset That hold 11 attribute,	-Using Google Maps to visualize the data. -Extract location of crime using 3D view -Using Graph and Chart Bar to report the crime frequency	-ML algorithm -KNN -NB	The method that have designed to gives a venue for identifying and analyzing crime networks using Google Maps and different machine learning algorithms.	In the future, a strategy is in progress to apply other classification algorithms to crime data and to improve prediction accuracy.
"Plug and Prey? Measuring the Commoditiza tion of Cybercrime via Online Anonymous Markets"	"Rolf van Wegber, Samaneh Tajalizade hkhoob, Kyle Soska, Ugur Akyazi, Carlos Ganan, Bram Klievink, Nicolas Christin, and Michel van Eeten" 2018	-Soska dataset -Christin's dataset 230,000 Data item include titles, descriptions , advertised prices, item-vendor mapping, Category classificatio n, shipping restrictions and different timestamps.	-Removing stop words -Lemmatize words -Tf-iTD Re-sampling is done -SMOTE -Normalized confusion matrix.	-SVM -Ground truth algorithm -SMOTE Method	The results only indicate imprecisely that the trend towards increased availability might not be as systematic as it has been described everywhere.	N/A
"Predictive Cyber Situational Awareness and Personalized Blacklisting: A Sequential Rule Mining Approach"	"MARTI N HUSÁK, TOMÁŠ BAJTOŠ, JAROSL AV KAŠPAR, ELIAS BOU-	-SABU dataset -DShield dataset -Honeypots dataset Collected from SABU 34 network based IDS,	-Correlation -Top K-sequential rule mining	-Sequential rule mining -Sequential pattern mining -Associatio n rule mining -Hybrid	The finding of descriptive rules from the knowledge using the methods of sequential rule mining and utilize them to predict cybersecurity incidents. Over 60% of the expected	The research work will help future threat management and warning correlation analysis by defining subsets of alerts that provide perspective for specific investigation.

	HARB, PAVEL ČELEDA ” 2020	honeypots.		approach	updates have been recorded to occur.	
“IMPLEMENTATION OF DATA MINING TECHNIQUES FOR CYBER CRIME DETECTION”	“K. Chitra Lekha, Dr. S. Prakasam ” 2018	- Cybercrime dataset	-Probability Density evaluation used Gaussian Metric -C5.0 Decision Tree -Clustering method splits data -Association rule mining -Correlation	-Hybrid Approach -SVM -DT -K-mean clustering	A few of data mining methods have continued and established to be reliable, and can be in the process of creating and enhancing new fraudulent actions to be better implemented.	N/A
“Applying Data Mining Techniques in Predicting Index and non-Index Crimes”	“Allemar Jhone P. Delima” 2019	Crime dataset Total record 7,267 index and non-index crime		-K-mean algorithm	Getting expected an improvement of 26%, the highest predicted physical injury crime. Livestock clatter is the least-reported crime in the province.	N/A
“Crime Rate Prediction using KNN”	“Ms. Vrushali Pednekar, Ms. Trupti Mahale, Ms. Pratiksha Gadhave, Prof. Arti Gore” 2018	Crime dataset Dataset has three dimension crime, criminal and geo-crime	-Correlation patterns AR -Correlation dimensional model	-KNN algorithm	On a specific day, the proposed framework predicts regions vulnerable to crime in India. If the research consider a specific state / country, it will be more precise.	N/A
“Using Data Mining Techniques and R Software to Analyze Crime Data in Kenya”	“Stephen Mangara Wainan, Joseph Njuguna Karomo, Rachael Kyalo, Noah Mutai” 2020	Crime dataset Data extracted from ICT authority website.	Association Rule	-APRIORI Algorithm -K-mean Clustering -Mapping -Shiny App	The daily crime collection is extracted from the database using the APRIORI method for association laws. The APRIORI method demonstrates that multiple crimes are related.	Study of time series can also be used to evaluate crime information as detected crime offences are reported along with the time are detected.
“A VIOLENT CRIME ANALYSIS USING FUZZY C-	“M. Premasundari and C. Yamini” 2019	USArrests dataset Attributes states, murder, assaults,	Correlation between crime attributes x and y co-ordinates. -D-dimensional measure data	-Fuzzy C-mean clustering algorithm	The proposed approach is to calculate the most and least murder, assault and rape arrests in US states.	Near future, along with the research structure, the statistical method of inevitable crime will be carried out using

MEANS CLUSTERING APPROACH"		urban Pop, and rape. Datatype numeric and character.	-Visualization into graphs, sector, plots histogram.			K-means clustering.
"DATA ANALYTICS APPROACH TO THE CYBERCRIME, UNDERGROUND ECONOMY"	"GOKAVARAPU RAGHAVA AVINASH, M S VENUGOPALA RAO" 2020	Malware dataset Obtain from cyber security firm, 53,815 attribute	NB is used to calculate the probability by normalization	-NB algorithm	This research recognizes the value of RAT for detecting underground cybercrime, so these concepts based on RAT are significantly essential parts of the system.	In order to have a better insight into the possible drawbacks, future work might classify terms and threats by sector, and it might try to find the network impact.
"Data mining Techniques in detecting and predicting Cybercrimes in Banking sector"	"K. Chitra Lekha, Dr.S.Prakasam, M.C.A" 2017	Cybercrime dataset Composed from News feeds, articles, and Blogs and Police Department websites and from banking sector.	-Influenced association rule mining algorithm is used to obtain fascinating pattern. -Updated any record by CAR rule. -J48 algorithm evolution of tree and validate the built tree.	-K-mean clustering algorithm, -Influenced Associative algorithm	With K-means, motivated association category with estimation Tree J48, it is possible to boost the classification competition and accuracy.	N/A
"Investigation and classification of cyber-crimes through IDS and SVM algorithm"	"Hamid Zolfi, Hamidreza Ghorbani, M. Hossein Ahmadzadegan" 2019	Cyber-attacks dataset Collected from petrochemical company 27 feature	-Feature selection for modeling -Machine learning algorithm used for classification -Normalization with Rapid Miner	-NB -DT -LR -SVM	Accuracy of various algorithms are: NB = 84% DT = 80% LR = 63% SVM = 99% SVM provide best accuracy	N/A
"Crime Analysis Through Machine Learning"	"Suhong Kim, Param Joshi, Parminder Singh Kalsi, and Pooya Taheri"	-Crime dataset Collected from VPD - Neighbourhood dataset VPD Crime dataset Collected from open data catalog the city of vancouver	-Min max normalization (0,1) - 5 Fold Cross-validation is used for validation process.	-KNN -Boosted DT algorithm	Accuracy and training time for technique 1 is 41.9% &903.63 sec, and with 459.26 sec training time, technique 2 is 43.2% accurate.	While as a classification algorithm, this model has good accuracy, it offers a structure for further studies.

		in GIS				
"Deep learning architectures for crime occurrence detection and prediction"	"Arnav Singh Bhardwaj, Divakar K M, Ashini K A, Devishree D S, Sheikh Mohammad Younis" 2019	Crime dataset Collected manually from Libyan Police Department	Association rule mining -Visualization is done using heat Map or geographic plots	-NB -K-mean clustering -Artificial Neural Network -DT -DNN	This model's performance is the amount of incidents of crime and non-crime. Maps or statistical plots are used to observe the effects of this projected knowledge.	N/A
"Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data"	"MINGCHEN FENG, JIANGBIN ZHENG, JINCHANG REN, AMIR HUSSAIN, XIUXIU LI, YUE XI, AND QIAOYUAN LIU" 2019	-San-Francisco crime dataset, Contain 2142685 incident. -Chicago dataset contain 5541398 records. -Philadelphia dataset contain 2371416 crime incident.	-Correlation for visualization -Random sampling for missing values Normalization Google maps	-State of the Art ML Algorithm -Big data analytics algorithm -DL -Neural Network	When the study analyze and came to know that LSTM algorithm and Prophet model both performed well than traditional neural network and the deep learning algorithm LSTM.	Near future, further test the efficacy and flexibility of the various models in the system, the researcher plan to perform more practical case studies.
"Cyber Intrusion Prediction and Taxonomy System Using Deep Learning And Distributed Big Data Processing"	"Hamzah Al Najada, Imad Mahgoub, Imran Mohammed" 2019	Aggregated dataset	-5 Fold-cross validation -Imbalance data use over sampling technique	-DRF -DL	After MSE DRF=0.21% DL=0.23% Oversampling RMSE DRF=0.45% DL=0.48%	The study will do future research on new and more recent forms of attacks for the future work.
"Analyzing and Predicting Cyber Hacking with Time Series Models"	"C. Soundarya, S. Usha" 2020	Cyber-Hacking dataset Human Health care service application dataset	Transformation is done through normalization N[0,1] -Autocorrelation Function for static data	-SARIMAX -RNN	Comparison to the SARIMAX model, the prediction accuracy of the RNN model, both the SARIMAX and RNN models can produce an offer fast	There are so many different topics that are left for future work. For one, it is difficult to explore how to calculate enormous values and how to deal

		collected from the PRC website Attributes: Incident Dates, Records and Categories.			in terms of real-life concerns.	with random errors.
"Crime Prediction Using Decision Tree (J48) Classification Algorithm"	"Emmanuel Ahishakiye, Elisha Opiyo Omulo, Danison Taremwa, Ivan Niyonzima" 2017	Crime and community dataset UCI machine learning repository website.	Feature selection for data reduction	-DT J48 algorithm	Accuracy of DT J48 algorithm is 94.25287%	N/A
"Criminal prediction using Naive Bayes theory"	"Mehmet Sait Vural Mustafa Gok" 2016	Crime Dataset	Cross Validation is used to train and test data.	-NB algorithm -DT	Accuracy of algorithms NB=81% DT=77%	With its new methods for extraction of crime data and the decision-making structure, the feature requires further studies on the criminal prediction issue.
"Deep Convolutional Neural Networks for Spatiotemporal Crime Prediction"	"Lian Duan, Tao Hu, En Cheng, Jianfeng Zhu, Chao Gao"	Crime dataset from New York city -Felony dataset 653,447 incident -311 dataset This dataset contain information about 10 milion complaint record.	Correlation based feature extraction -10 cross-validation using stratified Sampling method. -F1	-RF algorithm -CNN	F1 and AUC of the presented STCN are now better than other datasets, and the amount of time period exceeded 100, their developed scheme had the best classification efficiency.	In order to improve predictive efficiency in the future, multiple types of data must be assessed.
"Cyber Crime Analysis in Social Media Using Data	"M. Ganesan and P. Mayilvahanan"	Crime dataset	-NB is used to predict probability -K-mean clustering use to	-SVM -DT -ANN -NB	In order to obtain the information across the web, this approach is faster; successful web	In future, methods for crime patterns and network visualization can be developed for more

Mining Technique"			get patterns of structure data.		mining is to get the unstructured data into structured data.	visual and intuitive crime and intelligence.
"Crime Prediction using K Mean Algorithm"	"Vineet Jain, Yogesh Sharma, Ayush Bhatia and Vaibhav Arora" 2017	Crime dataset	K-mean clustering for data partition	-K Mean Clustering	Investigations indicate that the method is aware of the importance of speed of research, detection of prevalent trends of crime and areas vulnerable to crime for future prediction.	N/A
"Crime Investigation using Data Mining"	"S.R.Deshmukh, Arun S. Dalvi, Tushar J .Bhalerao, Ajinkya A. Dahale, Rahul S. Bharati, Chaitali R. Kadam" 2015	Criminal database	-NB is used to predict probability - Association rule to split data	-J48 Algorithm -NB -JRip data algorithm	The number of selected tool for data mining has a great impact on the results achieved. This is the main justification behind the comparing the results and the evaluation of the world's best algorithm for data mining.	N/A
"GIS BASED CRIME HOTSPOT MAPPING AND ANALYSIS USING RADIAL BASIS FUNCTION (RBF) AND INTERPOLATION METHOD"	"S.Sivaranjani A and S.Sivakumari" 2015	Crime Dataset		-Radial Basie Function method -Triangular with Linear interpolation method	Accuracy of existing and proposed method IDW= 72% RBF= 83%	The Linear Discriminant Evaluation tool can be used to accurately assess the most modern criminal hotspots in future work.
"Comparison of Machine Learning Algorithms for Predicting Crime Hotspots"	"XU ZHANG, LIN LIU, LUZI XIAO, AND JIAKAI JI" 2020	-Crime dataset Collected from P-GIS data set of public security	-Normalization [0,1] using Min Max scaler	-LSTM Model -KNN -NB -CNN -SVM -RF	Accuracy of other six machine learning algorithm is improved 46.6% to 52.3%. The accuracy of LCTM model 57.6% to 59.9%. LCTM model is better than others.	N/A
"Evolving Data Mining	"A Malathi	Crime dataset	-Handling missing values	-C4.5 - DT	The previous analysis is made	N/A

Algorithms on the Prevailing Crime Trend – An Intelligent Crime Prediction Model"	and Dr. S. Santhosh Baboo" 2011		by novel KNN-based -K-mean -DBScan for noise	-K-mean algorithm	83% of the crime. The experimental algorithm measured a crime of 89%.	
"Data Mining Instant Messaging Communications to Perform Author Identification for Cybercrime Investigations"	"Angela Orebaugh and Dr. Jeremy Allnutt"	-Dataset 1 Personal IM conversation logs obtained by clients of Gaim and Adium -Dataset 2 Obtained publicly available data from U.S. Cyber watch.	-Feature Extraction by using n-dimensional vector -Cross Validation to split data	-C4.5 -NB -KNN	Accuracy of prediction of authorship recognition of Dataset 1 = 88.42% Dataset 2 = 84.44%	N/A

III. DISCUSSION

After exploring the limitations of each method, some measures can be taken to improve the accuracy of cybercrime detection. We are seeking to combine the prediction of cybercrime with the prediction of real world crime. Some hybrid models with combinations of DP, ML, AI, NN, Cross Type &spatiotemporal correlations will improve crime prediction and some good pre-processing methods can enhance the accuracy. In crime analysis, we would like to incorporate and improve more sophisticated techniques. Threat management and warning correlation analysis by defining subsets of alerts that provide perspective for specific analyses can be useful for crime prediction. Study of time series can also be utilized to evaluate crime information as detected crime offences are reported along with the time is detected. In order to have a better insight into the possible drawbacks, future work might classify terms and threats by sector, and it might try to find the network impact.

IV. CONCLUSION

In this paper, the systematic analysis addressed various methods of detecting cybercrimes and reviewed various studies concerning the detection rates achieved and some of the results. In this paper, the state of the arts represented is analyzed and a comparison is conducted out through some tabulated data as a way to display the findings in order to recognize the respective methods and the results. The inaccessibility of benchmark datasets is an unavoidable outcome of the absence of collaboration in the processing of cybercriminal data between law imposition and authors. The nature of cybercrimes is another concern, as the research can occur on multiple sites, likeYouTube,Twitter,networks, orInstagram involving various kinds of datasets.That is encouraged to implement cybercriminal grading that could be utilized by authors as cybercrime datasets to address the accessibility problem of cybercrime datasets. However, it involves serious coordination between law imposition and authors as well as government sectors to establish cybercriminal profiling. Because the data that could be utilized in cybercriminal grading, which is often captious, confidential and personal, is uncertain. Moreover, the validityof disclosing this data is unclear. For this purpose, authors can discover a way to preserve data protection; by doing so, the study will profit from the data given by law imposition for research motive by cybercriminals, while at the same time protecting their privacy.

REFRENCES

[1] W. A. Al-khater, S. Member, S. A.- Ma, S. Member, K. Khan, and S. Member, "Comprehensive Review of Cybercrime Detection Techniques," vol. XX, 2020, doi: 10.1109/ACCESS.2020.3011259.

[2] M. Computing, "GUI BASED PREDICTION OF CRIME RATE USING MACHINE," vol. 9, no. 3, pp. 221–229, 2020.

[3] R. Ch, T. R. Gadekallu, and M. H. Abidi, "Computational System to Classify Cyber Crime O ff enses using Machine Learning," 2020.

[4] A. Karimi, S. Abbasabadei, and J. A. Torkestani, "Process Modeling and Extraction of Patterns of Computer Crimes Using Data Mining," vol. 28, no. 1, pp. 45–58, 2020.

[5] S. Hossain, A. Abtahee, I. Kashem, M. Moshiul, and I. H. Sarker, "Crime Prediction Using Spatio-Temporal Data," pp. 1–13.

[6] P. Cichosz, "Urban Crime Risk Prediction Using Point of Interest Data," 2020, doi: 10.3390/ijgi9070459.

[7] Y. Lin, M. Yen, and L. Yu, "Grid-Based Crime Prediction Using Geographical Features," 2018, doi: 10.3390/ijgi7080298.

[8] I. O. P. C. Series and M. Science, "Prediction Analysis Of Criminal Data Using Machine Learning," 2020, doi: 10.1088/1757-899X/852/1/012164.

[9] S. Moisés, T. Toapanta, L. Enrique, M. Gallegos, B. Eduardo, and C. Andrade, "Analysis to Predict Cybercrime Using Information Technology in a Globalized Environment Analysis to Predict Cybercrime Using Information Technology in a Globalized Environment," no. March, 2020, doi: 10.1109/ICICT50521.2020.00073.

[10] P. Yerpude and V. Gudur, "P REDICTIVE M ODELLING O F C RIME D ATASET," vol. 7, no. 4, pp. 43–58, 2017, doi: 10.5121/ijdkp.2017.7404.

[11] A. V Mbaziira, "A Text-based Deception Detection Model for Cybercrime A Text-based Deception Detection Model for Cybercrime," no. December, 2016.

[12] F. K. Bappee and S. Matwin, "Predicting Crime Using Spatial Features," no. i, pp. 1–7.

[13] B. Saini and G. Mahajan, "ScienceDirect Crime Prediction & Monitoring Framework Based on Spatial Analysis," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 696–705, 2018, doi: 10.1016/j.procs.2018.05.075.

[14] M. V. Pednekar, "Crime Rate Prediction using KNN," pp. 124–127.

[15] G. R. Avinash and P. G. Scholar, "DATA ANALYTICS APPROACH TO THE Complexity International Journal (CIJ)," vol. 24, no. 01, 2020.

[16] H. Zolfi, "through IDS and SVM algorithm," no. August, 2020, doi: 10.1109/I-SMAC47947.2019.9032536.

[17] S. Kim, P. Joshi, P. S. Kalsi, and P. Taheri, "Crime Analysis Through Machine Learning," 2003.

[18] E. Ahishakiye and I. Niyonzima, "Crime Prediction Using Decision Tree (J48) Classification Algorithm," vol. 06, no. 03, pp. 188–195, 2017.

[19] M. S. Vural and M. Go, "Criminal prediction using Naive Bayes theory ¨," 2016, doi: 10.1007/s00521-016-2205-z.

[20] L. Duan, T. Hu, E. Cheng, J. Zhu, and C. Gao, "Deep Convolutional Neural Networks for Spatiotemporal Crime Prediction," pp. 61–67.

[21] M. Ganesan and P. Mayilvahanan, "Cyber Crime Analysis in Social Media Using Data Mining Technique," vol. 116, no. 22, pp. 413–424, 2017.

[22] S. R. Deshmukh, A. S. Dalvi, and T. J. Bhalerao, "Crime Investigation using Data Mining," vol. 4, no. 3, pp. 22–24, 2015, doi: 10.17148/IJARCCE.2015.4306.

[23] X. U. Zhang, L. I. N. Liu, L. Xiao, and J. Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots," vol. 8, 2020, doi: 10.1109/ACCESS.2020.3028420.

[24] A. Orebaugh and J. Allnutt, "Data Mining Instant Messaging Communications to Perform Author Identification for Cybercrime Investigations," pp. 99–110, 2010.

[25] X. Zhao, "Exploring Spatio-Temporal and Cross-Type Correlations for Crime Prediction."

[26] S. Pastrana, A. Hutchings, A. Caines, and P. Buttery, "Characterizing Eve : Analysing Cybercrime Actors in a Large Underground Forum."

[27] E. Bou-harb, "Predictive Cyber Situational Awareness and Personalized Blacklisting : A Sequential Rule Mining Approach," vol. 11, no. 4, 2020.

[28] S. M. Wainana, J. N. Karomo, R. Kyalo, and N. Mutai, "Using Data Mining Techniques and R Software to Analyze Crime Data in Kenya," vol. 6, no. 1, pp. 20–31, 2020, doi: 10.11648/j.ijdsa.20200601.13.

[29] M. Premasundari and C. Yamini, "A VIOLENT CRIME ANALYSIS USING FUZZY C-MEANS CLUSTERING APPROACH," vol. 6956, no. April, pp. 1939–1944, 2019, doi: 10.21917/ijsc.2019.0270.

[30] K. C. Lekha, "Data mining Techniques in detecting and predicting Cyber crimes in Banking sector," 2017.

[31] C. Soundarya and S. Usha, "Analyzing and Predicting Cyber Hacking with Time Series Models," no. 7, 2020.

[32] V. Jain, A. Bhatia, Y. Sharma, and V. Arora, "Crime Prediction using K-means Algorithm," vol. 2, no. 5, pp. 206–209, 2017.

[33] C. Vijayalakshmi, "ScienceDirect ScienceDirect Design and Analysis of Machine Learning Algorithms for the Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India reduction The 9 th World Engineering Education Forum (WEEF - 2019)," *Procedia Comput. Sci.*, vol. 172, pp. 122–127, 2020, doi: 10.1016/j.procs.2020.05.018.

[34] I. Journal, M. Vol, I. Factor, and J. Homepage, "IMPLEMENTATION OF DATA MINING TECHNIQUES FOR CYBER CRIME DETECTION K. Chitra Lekha □ Dr. S. Prakasam □ □ ," vol. 7, no. 4, pp. 607–613.

[35] A. S. Bhardwaj, "Deep learning architectures for crime occurrence detection and prediction," vol. 5, no. 2, pp. 822–

824, 2019.

[36] A. Malathi and S. S. Baboo, "Evolving Data Mining Algorithms on the Prevailing Crime Trend – An Intelligent Crime Prediction Model," vol. 2, no. 6, pp. 1–6, 2011.

[37] H. A. Rahaman, "A Proposed Model for Cybercrime Detection Algorithm Using A Big Data Analytics," vol. 18, no. 6, 2020.

[38] K. Islam and A. Raza, "Forecasting Crime Using ARIMA Model."

[39] S. K. Rumi, K. K. Qin, and F. D. Salim, "Multi-officer Routing for Patrolling High Risk Areas Jointly Learned from Check-ins , Crime and Incident Response Data," pp. 1–21.

[40] B. Yang *et al.*, "A spatio-temporal method for crime prediction using historical crime data and transitional zones identified from nightlight imagery," *Int. J. Geogr. Inf. Sci.*, vol. 00, no. 00, pp. 1–25, 2020, doi: 10.1080/13658816.2020.1737701.

[41] Y. Lee, "Flag and boost theories for hot spot forecasting : An application of NIJ ' s Real-Time Crime forecasting algorithm using Colorado Springs crime data," 2019, doi: 10.1177/1461355719864367.

[42] C. Kadar, S. Feuerriegel, and C. Mascolo, "Leveraging Mobility Flows from Location Technology Platforms to Test Crime Pattern Theory in Large Cities," no. Icwsm, 2020.

[43] Y. Lee, O. Soohyun, and J. E. Eck, "A Theory-Driven Algorithm for Real-Time Crime Hot Spot Forecasting," 2019, doi: 10.1177/1098611119887809.

[44] A. Butkovic, S. Mrdovic, S. Uludag, and A. Tanovic, "Geographic Profiling for serial cybercrime investigation," vol. 18, 2018.

[45] R. Van Wegberg *et al.*, "Plug and Prey ? Measuring the Commoditization of Cybercrime via Online Anonymous Markets," 2018.

[46] M. Feng *et al.*, "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data," pp. 106111–106123, 2019.

[47] H. Al Najada, I. Mahgoub, and I. Mohammed, "Cyber Intrusion Prediction and Taxonomy System Using Deep Learning And Distributed Big Data Processing," *2018 IEEE Symp. Ser. Comput. Intell.*, no. Dl, pp. 631–638, 2018.

[48] "GIS B ASED C RIME H OTSPOT M APPING AND A NALYSIS U SING R ADIAL B ASIS F UNCTION (RBF) AND I NTERPOLATION M ETHOD," vol. 4, no. 5, 2015.