Big Data Analytics in law enforcement– Crime prediction using deep learning methods.

Deepthy Prasad
Department of Computer and Systems Sciences
Stockholm University
Stockholm, Sweden
depr5821@student.su.se

***Abstract*—The adoption of Big Data Analytics (BDA) methods such as Machine learning, Artificial intelligence, Natural Language Processing, etc by law-enforcement agencies is increasing as the traditional methods are less supportive in the analysis of heterogeneous data. This research is a study to investigate the various deep-learning-based BDA methods available for crime prediction and forecasting using historical data. The basis of this research paper is a short literature review including only state-of-the-art deep learning methods used worldwide in forecasting crime occurrences. The deep learning methods Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) are found to be the most popular techniques with high performance in crime data analysis. In order to support future studies, the effectiveness of each method found in the selected articles is also discussed in this study according to their precision and accuracy.**

Keywords—BDA, Machine learning, Deep learning, law enforcement, Crime prediction.

I. INTRODUCTION

The evolution of technology made drastic changes in human life and the internet became an indispensable part of our day-to-day life. A large amount of data became accessible through the internet from anywhere, anytime. As the amount and sources of information increased, the law enforcement organizations such as Police, Federal agencies, and various private and public investigation agencies also started to benefit from technological advancements in order to improve the efficiency of their services. These sources include structured and unstructured data which requires large memory such as video footage, audio files, chat logs, geodata, weblogs, etc. This type of data is known as Big Data and more accurately Big Data refers to large and complex datasets [1]. The traditional tools and approaches became incapable of handling this kind of heterogeneous data which lead to the need for new tools and methods [2]. 'Big data analytics (BDA)' is an effective and fast way to manage the multi-sourced, unstructured, and huge volume of data compared to the traditional methods.

A. Background

This research study is based on the Big Data Analytics ontology shown in Figure 1 [3]. The Big Data Ontology represent how state-of-the-art data processing methodologies and tools can be used by ensuring the proper utilization of available Big Data in any domain. The ontology has three layers which include the components of BDA, various data analysis methods, different models and algorithms available for the data analysis and popular big data management systems.

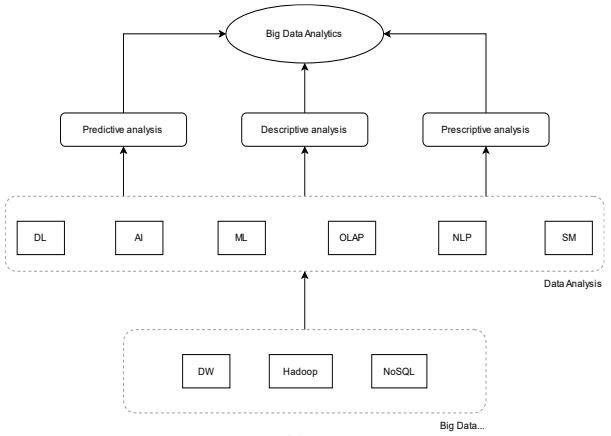

Figure 1 Big Data Analytics Ontology

The main three components of BDA are descriptive, predictive, and prescriptive analytics. Descriptive analytics shows what has already happened, predictive analytics is used to predict future outcomes based on existing big data and prescriptive analysis shows what should happen with the best outcome [4] [5] [6]. All these components can be used together to produce the best results [7]. The middle layer of the ontology has some state-of-the-art data analytics methods for Big Data which include AI – Artificial Intelligence, DL - Deep learning, ML - Machine Learning, OLAP - Online analytical processing, NLP – Natural Language Processing and SM - statistical modelling. The bottom layer includes various data management systems such as DW - Data Warehouse, Hadoop and NoSQL databases which can handle a large amount of heterogeneous data.

B. Research Problem

Nowadays, law enforcement agencies often come across large volumes of crime data from multiple sources and they have realized that this can be utilized for investigation and crime prediction [8]. The traditional centralized data storage method also has been changed to distributed storage which supports large and complex datasets known as big data and its analysis. The better performance and cost benefits are some additional reasons that encourage the change from traditional to big data analytics [2]. Exploring the applications of state-of-the-art BDA methods in law enforcement area for investigation and crime prediction is a vast area of research.

The current qualitative research is a literature review that explores the state-of-the-art Deep Learning methods used in crime prediction and the future scopes of these techniques in crime investigation.

II. Research Strategy and Methodology

A. Research Strategy

The current paper is a qualitative research to find out how deep learning methods can be used in crime prediction and the future scopes of them. A research strategy should be selected in such a way that the study results are descriptive enough to draw inferences regarding the research question. After the analysis of various research strategies available, 'Record keeping' was found to be the most suitable strategy for the current study. "Record keeping method makes use of the already existing reliable documents and similar sources of information as the data source" [9].

B. Data Collection Method

The corresponding data collection method used for the current study is 'document review' where existing data in form of various documents are collected and analyzed to conduct the research. The sample for the research is collected from Google Scholar, ACM, ResarchGate, IEEE Xplore and Science Direct databases. The literature papers collected from these sources are provided in the reference section. A data evaluation is carried out in order to find the relevant literature and then analyzed the data to draw inferences from pertinent literature.

As the research is based on the BDA ontology and narrowed only to deep learning-based BDA methods, keywords like 'BDA crime prediction', 'Deep learning crime prediction', 'Machine learning', 'crime prediction models', etc are used to search relevant articles. In order to find state-of-the-art methods, search options such as 'papers published after 2018' and 'sort by relevance' were also used. There were more than 3000 papers in each search. The first 20 papers from each search were selected after sorting the search result by relevance. A total of 80 papers were selected. Then glanced through the abstract of each paper and shortlisted 15 out of them which had a practical implementation of a deep learning model for crime prediction or forecast. The variety in the input crime data used and deep learning methods used such as RNN, CNN, GAN, DBN etc were also considered as criteria in the paper selection.

C. Theoretical base and concepts

The research paper is based on Figure 1 Big Data Analytics Ontology and to further narrow down the research area only deep learning methods are considered from the middle layer of the ontology. This section discusses some of the state-of-the-art deep learning methods which are used in the 10 research papers selected for this literature review.

Deep Learning (DL) is part of Machine Learning which gained popularity as data availability and performance of the computers increased. DL algorithms are based on neural networks which process data in a way similar to a human brain [18]. The main types of DL architectures are Fully connected neural network (FCNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Deep Belief Network (DBN) and Generative Adversarial Network (GAN). The DL algorithms can be used for pattern recognition and classification which makes them suitable for crime prediction and analysis such as the identification of fraudulent persons, websites, online predators through chat logs, criminal activities from video footage etc. FCNN can be considered a basic deep-learning neural network in which all the neurons in each layer are connected to the other layer which makes it possible to extract even small features of the input data. Whereas CNN contains convolution layers in which the neurons are only coupled to nearby neurons within the width of the convolutional kernel and are not densely connected. This makes CNN faster than FCNN [10]. CNN are powerful discriminative learning tool used for classification and works well on image data, extracting features which can be used in the face recognition of criminals. In RNN not only the input from the previous layer but also the input from the previous step is used in the current step which makes it excellent in processing text sequential data. Again, CNN is faster than RNN in terms of data processing. RNN is best suitable for time series data and texts such as chat logs, audio, etc. Long Short Term Memory (LSTM) is an extended form of RNN appropriate for generative learning used to make predictions based on time series data. LSTM is a good algorithm for Natural Language Processing (NLP) [11]. DBN is a generative learning model consisting of multiple hidden layers with connections between layers but not between units in the same layer [12, 13]. DBNs are used for image recognition, motion detection etc. GAN is a generative learning model but is also used for discriminative learning [14]. It consists of a generator and a discriminator and makes use of the adversarial machine-learning method [15]. They are mainly used in voice, video and image generation from training data.

III. Results and Discussion

A. Literature Review

The basis of the current research study is the most relevant 10 research papers that are shortlisted during the data collection stage. For ease of analysis and readability, the major findings from each paper are presented in a tabular format in Table 1 Literature Review. The table includes the objectives of the selected paper, deep learning methods used in the studies, their implementation to the specific scenarios and the suggested future studies.

B. Analysis of Results

Deep neural networks are good at extracting abstract features from the input data. The analysis of proposed or developed models in each literature review shows that CNN and LSTM are the most popular deep neural network models commonly used in crime prediction and forecast.

The CNNs have demonstrated extraordinarily high accuracy in image recognition and are visible from the results of 7 papers included in the review. CNN is used in them for face recognition of criminals, finding crime situations from videos and identification of crime-related objects such as guns. RNNs are more widely used by researchers in time series data and natural language processing, which mostly involves text sequences. We can observe that 5 papers in the selected papers are using them for different crime data analyses such as crime forecasting from history data, harmful chat identification from chat logs, crime trend prediction, time-based (daily, monthly, quarterly, etc.) crime prediction and taint analysis to find the vulnerabilities in software. The other three deep learning techniques observed in the chosen papers are FCNN, DBN and GAN. GAN is used in a paper where a generative prediction model of the city map is created. GAN gains popularity because of its capability to regenerate image inputs [16]. This is evident from the paper where the GAN model has successfully generated a crime rate-based heatmap of the city. In a paper [17] the next 5-year behaviour

No	*Research Paper*	*Research Objective*	*Deep Learning method(s) used*	*Key points*	*Future Research*
1	Crime activities prediction system in video surveillance by an optimized deep learning framework. [18]	This paper introduces a new deep-learning framework to detect crime activities in video clips from surveillance cameras.	A novel lion-based deep belief neural paradigm (LbDBNP). Convolutional Neural Network (CNN). Recurrent Neural Network (RNN).	LbDBNP model was found to have the best performance compared to RNN and CNN. It uses Deep Belief Network(DBN) and the Lion optimization algorithm for effective image processing and anomaly detection. Three crime datasets from the web are used in the training, testing and validation of the proposed model.	The current study is good at detecting anomalies as objects and activities but inefficient for face detection. So, crime detection based on face recognition is the future work proposed in the study.
2	Crime Prediction Model using Deep Neural Networks. [17]	This paper implements a neural network model on personal criminal charge history data to predict the behaviour of criminals in the next 5 years.	Deep Neural Network with fully connected convolution layer.	A Deep Neural Network with a fully connected convolution layer is implemented using TensorFlow including drop-out layers and with L2 regularization to remove overfitting. The model achieved 99.7% accuracy in future crime prediction and 94% accuracy in the prediction of the level of crime.	There are no specific future works discussed in the paper except that the results can be used to stop criminals from committing crimes in the future.
3	Deep Learning based Facial Detection and Recognition in Still Images for Digital Forensics. [19]	This paper introduces a Face Detection and Recognition in Images (FDRI) open-source software which can be used to automate face recognition and detection in digital forensic software Autopsy.	The FDRI algorithm implements a deep CNN to analyze the images in the dataset.	The performance of the detection algorithm depends on the size of the image as large-size images have high precision. Also, some features of the images such as face orientation, illumination, etc are also affecting the performance of the FDRI software. The training data used affects the accuracy of the model when it comes to Facial recognition in everyday scenarios.	Future work includes using better training sets to make the software more accurate. Also increasing the processing speed by getting the advantage of hardware evolution.
4	Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data. [20]	The paper uses state-of-the-art big data analytics to identify patterns and trends in the crime data from 3 US cities.	Prophet model Long Short Term Memory (LSTM) Multilayer feed-forward neural network model	LSTM and Prophet models were found to be more effective in forecasting crime trends compared to the feed-forward neural network model. This comparison is carried out in terms of RMSE(Root Mean Square Error) and Spearman correlation under the different sizes of training samples.	Proposes a future expansion of the current study using new techniques to draw potential patterns and trends in crime data. Also, more realistic case studies to evaluate created prediction models.
5	A machine learning based forensic tool for image classification - A design science approach. [21]	This research paper proves how pre-trained models without any specific modifications can be used in practice by developing a prototype forensic tool to classify images with guns using various pre-trained ML models.	Keras pre-trained models: Convolutional Neural Network (CNN) model: • Inception V3 • Xception • VGG16 • Residual Network (ResNet)	All 4 models are trained in the ImageNet dataset and the newly designed prediction function is utilized to find the images with the gun. InceptionV3 is selected as the final model according to the evaluation scores and testing results on 3 different balanced and non-balanced datasets.	Future research suggested are evaluating the current prototype tool in real cases, evaluating the automization capability of the tool and improving the model selection by applying multi-criteria decision-making methods.
6	Prediction of crime rate in urban neighborhoods based on machine learning. [22]	This paper uses a neural network algorithm to develop a generative prediction model of a city map based on crime data in Philadelphia from 2006 to 2018.	Generative adversarial network (GAN)	A GAN model is trained with crime data combined with the city map. The input to the GAN model is the map of a city area and the output is a heat map based on the crime rate.	The future proposals are to use the current research as a workflow to evaluate various other urban features and be useful to citizens to choose a living environment.

7	Digital forensics supported by machine learning for the detection of online sexual predatory chats. [23]	This paper implements machine learning methods to identify and classify harmful chat logs based on a digital forensic process model.	Logistic Regression (LR) XGBoost Multilayer Perceptrons (MLP) Bidirectional Long Short-Term Memory (BLSTM)	The study uses online data in XML format after Dimensionality reduction and decoding specific symbols to text. Various classification models are trained and tested. LR, MLP and Bidirectional LSTM were found to be the best among used ML techniques. The study was able to classify the data and found some words and phrases in harmful chat messages.	This study can be made quantitative by using more data to train and validate the models proposed to generalize predatory behavior.
8	Convolutional Neural Network Based Criminal Detection [24]	The paper uses deep learning models for crime prediction from facial images.	• Standard Convolution Neural Network (CNN) Pretrained CNNs: • VGG-16 • VGG-19 • InceptionV3	The classification was based on facial emotions and age. The study uses publicly available data to train the models and evaluation metrics such as Precision, Recall, Accuracy and AUC are used. Pretrained CNN models VGG-16 and VGG-19 are found to be most effective with more than 99% accuracy in criminal face identification.	Suggest using a great dataset with face images and adding more personality features for the classification in future studies.
9	Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques. [25]	This paper uses 8 different machine learning algorithms, a deep learning method and a time series method for crime prediction and forecasting with Los Angeles and Chicago crime data.	Time series analysis by LSTM. Other ML models used: • Logistic regression • SVM • Naïve Bayes • KNN • Decision tree • MLP • Random forest • XGBoost	LSTM classify the crimes yearly, quarterly, monthly, weekly and daily. The crime prediction accuracy of LSTM was high compared to the other 8 ML techniques used based on RMSE and MAE.	The proposed future works are to use hybrid models to enhance the performance, build location maps and visual images and use satellite imagery data to expand the study.
10	A deep learning based static taint analysis approach for IoT software vulnerability location. [26]	The paper implements a static taint analysis based on deep learning methods to identify the vulnerabilities in Internet of Things (IoT) software.	Recurrent Neural Network (RNN) Long Short Term Memory (LSTM) Bidirectional LSTM (BLSTM) Convolutional Neural Network-based LSTM (CNN-based LSTM)	Three taint selection principles are used to find the taint and taint weight calculation method for screening. The developed system uses a CNN-based LSTM as it has high accuracy compared to RNN, LSTM and BLSTM. The proposed model is evaluated using Code Gadget Database and achieved an accuracy of 97.32 %	The proposed future work is to embed the developed system in IoT devices as a chip or add it as part of smart homes, smart cars, etc where the IoT equipment is used to ensure safety.

Table 1 Literature Review

of the criminals is predicted using FCNN and achieved an accuracy of 99.7% in future crime prediction.

Among the chosen papers, there were some papers which combines different algorithms and create a new model with prediction and forecasting capabilities. A novel lion-based deep belief neural paradigm (LbDBNP) [18] is such a model in which DBN and lion optimization algorithms are combined to achieve both image processing and anomaly detection capabilities together. This model has outperformed basic RNN and CNN models in the recognition of criminal activities from surveillance camera footage. The bi-directional approach to LSTM was also found to be more effective than the unidirectional LSTM as BLSTM attained an accuracy of 98% in the detection of online sexual predatory chats [23]. Another mixed deep-learning model approach observed was in static taint analysis where a CNN-based LSTM model provided an accuracy of 97.32% in vulnerability detection in IoT software. [26].

The paper [20] observed that the LSTM perform better compared to a prophet and a 3-layer neural network model in crime forecasting in 3 cities in the United States (US). The results are shown in the figures Figure 2, Figure 3 and Figure 4:

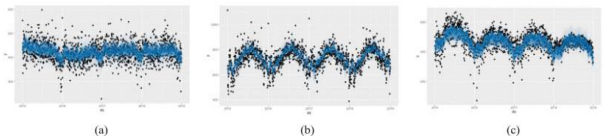

Figure 2 Prophet model a) San Francisco b) Chicago c) Philadelphia

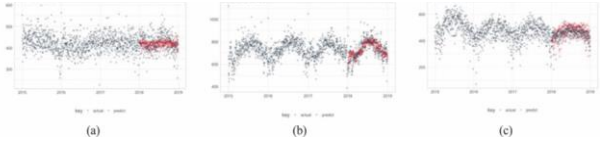

Figure 3 LSTM model a) San Francisco b) Chicago c) Philadelphia

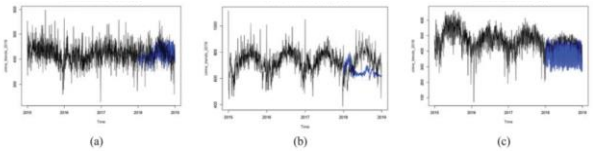

Figure 4 3-layer NN model a) San Francisco b) Chicago c)Philadelphia

We can observe that 6 papers out of the chosen 10 papers have used pre-trained deep learning models with or without modifications. The paper [21] uses only Keras pretrained models without any modifications for detecting images with guns and VGG-16 and VGG-19 among them were found to provide an accuracy more than 99% Figure 5 and Figure 6. This paper proves that the pre-trained models are effective in crime prediction without any modifications and opens an opportunity in image classification for users without deep knowledge [24].

Figure 5 Accuracy_VGG16

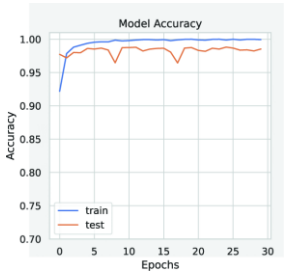

Figure 6 Accuracy_VGG19

IV. Conclusion

Big Data Analytics can make a great impact on the services of law enforcement agencies as it can provide a variety of methods and tools to draw meaningful inferences from the available historical data. The data collection step has proved the variety of applications of BDA in the law enforcement domain. The literature review has explored the applications of Deep Learning models in the law enforcement area and observed that they provide promising results in crime prediction and forecasting using historical data.

Here are the main inferences from the literature review:

- CNN are the most commonly used deep learning method for image classification and face recognition problems with crime data.
- LSTM was found to be well-suited for managing time series crime data and forecasting crime. Bidirectional LSTM provided high accuracy compared to LSTM but it is not recommended for large volume data due to high training time.
- Combined models such as CNN-based LSTM, DBN combined with the lion optimization algorithm, and Deep Neural Network with a fully connected convolution layer has achieved high accuracy and sometimes outperformed the basic CNN and RNN models.
- The literature review shows that digital forensics is the most beneficial area of law enforcement with DL methods as it supports faster evidence collection from huge data such as chat logs, browser history, etc.
- One of the advantages found from the literature review is that there are several deep learning pretrained models, developing tools and libraries available in the market now which provide hope for developers with less DL knowledge to get the benefits of them in their works.

This paper can be extended by adding more literature reviews with new applications of DL methods in crime detection and forecasting which can expand the inferences drawn in this paper. Considering papers which use combined deep learning algorithms is also recommended. Similar literature review studies with small areas within the Law enforcement domain are also recommended as it is a very big area of research with many sub-divisions within it.

V. References

[1] "Big data - Wikipedia," [Online]. Available: https://en.wikipedia.org/wiki/Big_data.

[2] "Traditional data Vs Big data," [Online]. Available: https://treehousetechgroup.com/big-data-vs-traditional-data-whats-the-difference/.

[3] Z. Sun, L. Sun and K. Strang, "Big Data Analytics Services for Enhancing Business," [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/08874417.2016.1220239.

[4] "Book," [Online]. Available: https://books.google.se/books?hl=en&lr=&id=HYYaX9dsZsYC&oi=fnd&pg=PR13&dq=Mine.

[5] "Data, information and analytics as services," [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923612001558.

[6] "Book2," [Online]. Available: https://books.google.se/books?hl=en&lr=&id=4JN4CgAAQBAJ&oi=fnd&pg=PP1&dq=Datab.

[7] "types of BA," [Online]. Available: https://online.bath.ac.uk/content/descriptive-predictive-and-prescriptive-three-types-business-analytics.

[8] "SNA," [Online]. Available: https://link.springer.com/article/10.1007/s12117-008-9057-6.

[9] "research methods," [Online]. Available: https://www.questionpro.com/blog/qualitative-research-methods/.

[10] "Neural Networks," [Online]. Available: https://www.baeldung.com/cs/neural-networks-conv-fc-layers.

[11] "What is LSTM?," [Online]. Available: https://www.techopedia.com/definition/33215/long-short-term-memory-lstm.

[12] "DBN," [Online]. Available: https://en.wikipedia.org/wiki/Deep_belief_network.

[13] "DBN," [Online]. Available: https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-deep-belief-network.

[14] C. G. Y. D. Y. L. X. Z. F.-Y. W. Kunfeng Wang, "Generative Adversarial Networks," [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8039016.

[15] "adversarial ML," [Online]. Available: https://viso.ai/deep-learning/adversarial-machine-learning/.

[16] "GAN," [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/generative-adversarial-network-GAN.

[17] V. A. P. S. Y. R. P. V. A. N. R. A. Soon Ae Chun, "Crime Prediction Model using Deep Neural Networks," [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3325112.3328221.

[18] H. V. R. Kadari Kishore Kumar, "Crime activities prediction system in video surveillance by an optimized deep learning framework," [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.6852.

[19] A. F. R. Patricio Domingues, "Deep Learning based Facial Detection and Recognition in Still Images for Digital Forensics.," [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3339252.3340107.

[20] M. Feng, J. Zheng, J. Ren, A. Hussain, X. Li and Y. Xi, "Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data.," [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8768367.

[21] D. M.-R. H. B. L. M. K. H. Joanna Rose, "A machine learning based forensic tool for image classification - A design science approach.," [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281721001827.

[22] H. Z. Jingyi He, "Prediction of crime rate in urban neighborhoods based on machine learning.," [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197621003080.

[23] J. T. V. C.H.Ngejane, "Digital forensics supported by machine learning for the detection of online sexual predatory chats," [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666281721000032.

[24] H. Verma, S. Lotia and A. Singh, "Convolutional Neural Network Based Criminal Detection," [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9293926.

[25] W. Safat, S. Asghar and S. A. Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques," [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9424589.

[26] X. Z. X. D. L. Z. R. C. M. G. Weina Niu, " A deep learning based static taint analysis approach for IoT software vulnerability location.," [Online]. Available: https://www.sciencedirect.com/science/article/pii/S026322411931005X.