

**SCHOOL OF COMPUTING AND ENGINEERING**

**DEPARTMENT OF DATA SCIENCE**

**DISSERTATION RESEARCH PROJECT 2022/2023**

**Crime Forecasting and Prevention:**

**Examining the role of machine learning models**

**OLUWATOYIN ODENIYI**

**S4115181**

**SUPERVISOR: DR. BHUPESH MISHRA**

**31st May 2023**

Table of Contents

[Chapter 1- Introduction 1](#_Toc136390397)

[1.1. Overview 2](#_Toc136390398)

[1.2. Problem Statement 3](#_Toc136390399)

[1.3. Research Aim 4](#_Toc136390400)

[1.4. Research Objective 4](#_Toc136390401)

[1.5. Scope 5](#_Toc136390402)

[1.6. Ethical Issues 5](#_Toc136390403)

[1.7. Conclusion 6](#_Toc136390404)

[Chapter 2-Literature Review 7](#_Toc136390405)

[2.1. Overview Of Machine Learning Models for Forecasting and Preventing Crimes 7](#_Toc136390406)

[2.2. Theoretical Foundations Supporting Machine Learning Models in Forecasting Crime 9](#_Toc136390407)

[2.3. The Historical Evolution of Machine Learning Models to Forecast Crimes 11](#_Toc136390408)

[2.4. Principal Machine Learning Models Used to Forecast Crimes 13](#_Toc136390409)

[2.5. Machine Learning Models to Forecast and Prevent Definitive Crimes 15](#_Toc136390410)

[2.5.1. Terrorism 15](#_Toc136390411)

[2.5.2. Cybercrime 16](#_Toc136390412)

[2.5.3. Financial crime 17](#_Toc136390413)

[2.5.4. Urban crime 18](#_Toc136390414)

[2.6. Evaluation of Machine Learning Models for Predicting and Preventing Crimes 18](#_Toc136390415)

[2.7. Ethical Concerns and Potential Biases in Application of Machine Learning Crime Models 20](#_Toc136390416)

[2.8. Future Research and Development for Machine Learning Models Utilised in Forecasting Crimes 22](#_Toc136390417)

[2.9. Comparison Of Machine Learning Models and Established Techniques for Crime Predictions 23](#_Toc136390418)

[2.9.1. City Policing Plans Driven by Crime Prediction ML Models 23](#_Toc136390419)

[2.9.2. Community Policing Crime Prevention Programs Incorporating ML Models 24](#_Toc136390420)

[2.10. Case Studies and Practical Applications of Machine Learning Models for Crime Forecasting 25](#_Toc136390421)

[Chapter 3 - Research Methodology 28](#_Toc136390422)

[3.1. Machine Learning Modelling 37](#_Toc136390423)

[Chapter 4- Result Evaluation and Discussion 39](#_Toc136390424)

[Chapter 5-Conclusion 42](#_Toc136390425)

[5.1 Future Research 43](#_Toc136390426)

[References 44](#_Toc136390427)

[Appendix A 53](#_Toc136390428)

[A.1. Decision Tree Modelling 53](#_Toc136390429)

[A.2. Evaluation Metrics for Decision Tree 54](#_Toc136390430)

[A.3. Entropy 54](#_Toc136390431)

[A.4. Gini index 55](#_Toc136390432)

[A.5. Logistic Regression Modelling 57](#_Toc136390433)

[APPENDIX B 59](#_Toc136390434)

[B.1. K-MEANS CLUSTERING 59](#_Toc136390435)

**LIST OF FIGURES**

[Figure 1. Research methodology workflow. 28](#_Toc136125826)

[Figure 2.Occurrence of each crime type 30](#_Toc136125827)

[Figure 3. Average number of crimes daily and monthly. 32](#_Toc136125828)

[Figure 4.Average of highest and lowest crimes per month. 32](#_Toc136125829)

[Figure 5. Average of crime categories per month. 33](#_Toc136125830)

[Figure 6.Average of crime categories per month. 34](#_Toc136125831)

[Figure 7.Histogram of crime distribution per day. 34](#_Toc136125832)

[Figure 8.Line plots of each crime per year. 35](#_Toc136125833)

[Figure 9.Number of crimes per weekday. 36](#_Toc136125834)

[Figure 10.Total Crimes Per Day of Each Year. 36](#_Toc136125835)

[Figure 11.Count of crime types. 37](#_Toc136125836)

[Figure 12.Total Crimes in Each Neighbourhood Per Year. 37](#_Toc136125837)

[Figure 13. Decision tree modelling of the Vancouver crime dataset with Entropy. 53](#_Toc136125838)

[Figure 14.Decision tree modelling of the Vancouver crime dataset with Gini Index. 54](#_Toc136125839)

[Figure 15.Bar chart of feature of importance with score. 56](#_Toc136125840)

[Figure 16.Bar chart of feature of importance with score. 57](#_Toc136125841)

[Figure 17. Elbow curve plot using K-Means Clustering 59](#_Toc136125842)

[Figure 18. Plot representing the distinct clusters. 60](#_Toc136125843)

LIST OF TABLES

[Table 1. Occurrence of each crime type 29](#_Toc136212377)

[Table 2. Occurrence of the grouped crime types 30](#_Toc136212378)

[Table 3.Occurrence of grouped crime types per neighbourhood 30](#_Toc136212379)

[Table 4.Occurrence of crimes in Central Business District 31](#_Toc136212380)

[Table 5.Machine learning models with prediction accuracy values 38](#_Toc136212381)

[Table 6.Machine learning models’ performance after hyperparameter tuning 38](#_Toc136212382)

[Table 7.Feature of importance with score 56](#_Toc136212383)

[Table 8. Classification report for Decision Tree 57](#_Toc136212384)

[Table 9.Feature of importance with score 58](#_Toc136212385)

LIST OF EQUATIONS

[Equation 1. Mathematical Equation for Entropy 55](#_Toc136029674)

[Equation 2. Mathematical Equation for Gini Index 55](#_Toc136029675)

GLOSSARY

**ANN**: Artificial Neural Network

**ARIMA**: Auto Regressive Integrated Moving Average

**CART**: Classification and Regression Trees

**CBD**: Central Business District

**CNN**: Convolutional Neural Network

**DT**: Decision Tree

**FFNN**: Feedforward Neural Network

**GNB**: Gaussian Naïve Bayes

**GRU**: Gated Recurrent Unit

**ID**: Identification

**KNN**: K Nearest Neighbour

**LR**: Logistic Regression

**LS/CMI**: Level of Service/Case Management Inventory

**LSTM**: Long-Short Term Memory

**ML**: Machine Learning

**MLP**: Multi-Layer Perceptron

**MSE**: Mean Square Error

**NARX**: Nonlinear Auto-Regressive with Exogenous Input

**NB**: Naïve Bayes

**NNH**: Nearest Neighbour Hierarchical

**RF**: Random Forest

**RNN**: Recurrent Neural Network

**RTM**: Risk Terrain Modelling

**SHAP**: Shapely additive explanation

**SLIM**: Super sparse Linear Integer Model

**STARMA**: Space-Time Autoregressive Moving Average

**STNN**: Spatial-Temporal Graph Neural Network

**SVM**: Support Vector Machine

**UK**: United Kingdom

**USA**: United States

**ABSTRACT**

Reducing crime and preventing criminals from perpetuating their acts, can be helpful in ensuring public safety. However, traditional crime analytical methods have difficulties in managing complex and large crime data, thus limiting effective crime detection. Leveraging on advancements in predictive analysis and machine learning(ML) to enhance crime forecasting and prevention processes, can circumvent this limitation.

This study aims to expore the role of machine learning models such as: Decision Trees, Neural Network, K-Nearest Neighbour, in crime forecasting. An in-depth analysis of different machine learning models, supervised and unsupervised, useful in forecasting crime patterns, criminal behaviours and hot spots (places of high crime risks). There would also be comparison to traditional crime analysis techniques. Also, this research will investigate various factors which can influence the effectiveness of ML models to predict crimes. This will include: crime data source, features and size. Individual, social and environmental factors will also be examined. Ethical issues and biases linked to crime datasets and ML models will be investigated and solutions proffered.

The research results combined with literature reviewed, will contribute to the field of crime detection and forecasting as applicable insights into ML models useful for this purpose, will be provided. This will be helpful to crime analysts, law agencies, police forces and policy makers in developing better crime prevention processes and effective allocation of resources (human and monetary). In addition, benefits, limitations and ethical considerations in ML modelling of the Vancouver crime dataset for prediction analysis will be highlighted.

Keywords: crime forecasting, machine learning models, analysis, ethical issues, decision trees, Neural network, bias, crime analysts, law agencies, prevention, crime detection.

# Chapter 1- Introduction

The utilization of machine learning (ML) models for forecasting crime, presents a potential opportunity for the prevention of criminal activity (Lin, Chen and Yu, 2017). Conducting an in-depth investigation of the different Machine Learning (ML) models which have been applied in past research to analyse crime datasets for prediction. This will be a valuable contribution to crime prevention strategies formulated by security agencies. Therefore, the aim of this study to provide the most effective models which are deployable as crime prevention tools. The practical application of ML models for crime prediction is contingent on their purpose, efficacy and usability by security agencies (Shah, Bhagat and Shah, 2021). Research works comparing the prediction accuracies of ML models for forecasting crimes, have varying objectives. Accuracies of the models can be different, as each is unique in its performance and size of datasets analysed differs. Analysis is sometimes done with unaddressed biases and issues of privacy plaguing the results. Also, lacking in past studies is research which wholly covers relevant objectives and practical insights of applicable and effective ML models in preventing crimes. Some identified gaps to be addressed includes issues of privacy, historical data with challenges of racial and cultural bias, applicability of replicable and interpretable, well-tested ML models for crime prediction. The overall aim of this research is to cover most of these areas as a one-stop knowledge base with relevant and proven answers which can be beneficial in crime prevention. This research also includes processes like data standardization, feature engineering methods for extracting relevant knowledge out of crime datasets and data exploration to find the hidden relationships between features. All of this combined can be instrumental in establishing crime patterns, useful for forecasting crimes (ToppiReddy, Saini and Mahajan, 2018). Also included are the methods to produce better predictive accuracies like hyperparameter tuning of the models.

In some studies, ML models trained on crime data from one urban area with resulting high prediction accuracies, have been successfully applied to other similar locations for the same purpose. This shows the importance of investigating various ML models for crime prediction with similar crime datasets as replication or transfer of the process is feasible; a cost-saving security measure implementable in crime prevention (Cichosz, 2020). Some studies have shown that data attributes and feature selection methods are important in the success of ML models as crime prediction tool (Abdul Jalil, Mohd and Mohamad Noor, 2017). Therefore, this relationship will be investigated alongside some real-world deployment of ML models in crime forecasting. ML model integrated software for crime predictions deployed by security agencies includes PreCobs (in Germany), PredPol (in United States and England) and Hunchlab (in United States) (Mugari and Obioha, 2021).This policing software have allowed a more effective allocation of resources and fundings of law enforcement agencies for crime prevention (Jenga, Catal and Kar, 2023). However, with the innovation of predictive policing software, there are limitations and drawbacks which includes lack of transparency on how predictive ML models work, civil right infringement, flawed data and accountability of the police officers in administration of the software (Meijer and Wessels, 2019). Finding solutions which will ensure accountability and transparency in implementation of the policing software, with regulation of algorithmic crime analysis ML tools, will positively impact the application of ML models and algorithms as crime forecasting and prevention tool (Egbert and Leese, 2021). Therefore, there is a need for thorough investigation of ML models as crime prediction tools, not only to determine which will be productive in forecasting crimes but proposing actionable recommendations addressing these challenges. The methodology approach will involve models such as K-Nearest Neighbour (KNN), Decision Tree (DT), Multi-Layer Perceptron (MLP) and Logistic Regression (LR) for predictive analysis of the Vancouver crime dataset. Evaluation metrics will be applied and hyperparameters will be tuned to improve the prediction accuracies. An overview of challenges and limitations in the analysis of the Vancouver crime dataset will be covered. Future directions for ML models applications in forecasting crimes will also be a contribution to this evolving field of crime prevention.

The following Overview, Problem Statement, Research Aim, Objectives, Literature review, Methodology and Conclusion will provide augmented view of this research work.

## Overview

The rise in crime worldwide is likely aided by technology development and rapid urbanisation, hence the motivation for this research is to make valuable contribution to the field of crime prevention using knowledge of effective crime forecasting ML models. An exploration of the application of machine learning algorithms in predicting crime can be instrumental in proving which ML models can work successfully and which could be integrated into crime prevention software or processes. The potential benefits of using machine learning techniques will be delved into, mainly to address the information gap in practicable and unbiased crime detection ML models. The existing literature on ML models for crime prediction and prevention will be explored and comprehensively reviewed by illustrating the strengths and weaknesses of various approaches including real-world examples of application.

The methodology to investigate machine learning models to forecast crimes will be based on analysis of the Vancouver crime data which is a publicly available dataset. It was released as the City of Vancouver Open Data Catalogue, provided by the Canadian city’s Police department. Data features include crime types, location, time and location of reported incidents from 2003-2017. Techniques for this analysis will include data cleaning and transformation which will include standardization and feature engineering. Then visualisation to provide better understanding of the relationships between the data features. Testing different machine learning algorithms and evaluating the performances of the models to forecast crimes will also be conducted. Hyperparameter tuning of the models will be implemented to improve the prediction accuracies followed by comparative analysis of the results to determine which will serve successfully to forecast crimes, with focus on run time, accuracy score and absence of bias or privacy issues.

Conclusion will be a discussion of findings and their implications for crime prevention in relation to applicability by law enforcement agencies. Areas for future research will be highlighted as well as the limitations of this current study.

Overall, this research work seeks to contribute to the growing field of ML models used in crime forecasts by providing useful insights into their potential to aid crime reduction, with strategies to avoid bias and violation of people’s privacies.

## Problem Statement

The public's safety continues to be seriously threatened by criminal activities despite significant efforts to lower crime rates. There is a growing need to explore alternate ways to forecast and prevent crimes as traditional methods of crime prevention and law enforcement have limits especially due to large size of crime datasets and evolving technological advancements.

With notable great success, machine learning models have been applied in several industries, including marketing, finance, and healthcare for myriad purposes. Use of machine learning algorithms for crime forecasting has however received relatively less attention. The deployment of these ML models as tools in the criminal justice system, including policing, therefore faces a significant knowledge gap. By examining ML models for crime prediction, this dissertation seeks to address this issue. The research goal is to determine how effective different ML models are in forecasting crime patterns and repeat offenders; in identifying hotspots or potential areas of criminal acts and evaluation of prediction accuracies. Crucial to this study is also an evaluation of supervised and unsupervised ML models in real-world law enforcement settings for crime prevention. This will provide insights into the potential of these models in contributing innovative and effective strategies to forecast crimes.

## Research Aim

This study intends to assess the effectiveness of ML models to forecast and prevent crimes with the goal to improve the performance of these models and provide useful insights in the development of applicable crime fighting strategies.

## Research Objective

1. To critically explore ML models which have successfully predicted crimes in various locations with evaluation of their accuracy scores.

2.To show how data type, size and features determines the choice and performance of ML models for crime prediction.

3.To employ feature engineering and hyperparameter tuning in enhancing performances of ML models to forecast crimes.

4.To address ethical, legal considerations and privacy concerns in deploying ML models

as crime forecasting tool.

5.To provide suggestions on which ML models can be beneficial for crime prediction.

## 1.5. Scope

Investigating how machine learning algorithms may be used to anticipate and stop crime is this study's primary focus. Also, an exploration of methods wherein ML integrated strategies have been deployed in the criminal justice system, especially by law enforcement and security agencies.

Various research on different crime data will be investigated for this study. Focus will be on factors like crime reports, criminal histories, and other pertinent data attributes. The project will also involve the testing of a variety of machine learning techniques, including supervised and unsupervised learning algorithms for crime prevention. This research will not include implementation of these ML models in actual environments as this has limitations like time constraint, computing power for large datasets and ethical considerations. Instead, it will focus on evaluating the feasibility of machine models as crime forecasting tools in analytical research works and police settings. Also, the analysis of the Vancouver crime dataset to evaluate prediction accuracies of different ML models and methods to improve the accuracies, will be included in this project.

This study will also assess if applying machine learning models to actual crime prevention and law enforcement situations is practical factoring personal data usage, computational time and costs of integrating into useable formats. Evaluation of the constraints and potential biases of using machine learning models for the prediction of crime will also be a key component. This will entail analysing the moral ramifications of employing these methods as well as any potential effects on vulnerable populations.

A quantitative approach will be employed to conduct this research and results will be presented through exploratory data analysis, statistical models with evaluation of prediction accuracies.

## 1.6. Ethical Issues

This project is my original work, and all the materials belonging to others have been rightly cited. The tools to be used includes Python, which is open source and the dataset analysed is a publicly available one. This dissertation followed the University of Gloucestershire guidelines for ethical approval.

## 1.7. Conclusion

The application of ML models for crime prediction and prevention can significantly improve public safety when the occurrence of criminal activities is reduced. This will impact economic growth, as businesses or individuals will thrive in safety. Analysing historical crime data and identifying trends and patterns using ML models, can produce accurate predictions or forecasts of crime incidences and hotspots. This knowledge could be useful to law enforcement agencies in allocating resources effectively to prevent crime activities.

However, there are challenges linked to the deployment of ML models for forecasting crime which includes bias and fairness in algorithmic decision-making, as well as civil liberties and privacy concerns. Therefore, it is important to carefully put into consideration the ethical implications of ML algorithm-based technologies for crime predictions and the need to develop appropriate safeguards to protect people’s rights and liberties when collating and analysing crime data.

The investigation of ML models for predicting crime is an important research area that has the power to yield notable benefits for a safer society. Leveraging the power of data and analytics can help with insights into criminal behaviours and crime patterns, which can be useful in fighting crime.

# Chapter 2-Literature Review

## 2.1. Overview Of Machine Learning Models for Forecasting and Preventing Crimes

A research work was proposed to develop a crime prevention and safety measure system which can be deployed by security officers. This entailed analysing and predicting crimes like Assault, Homicide and Carjacking, using Multiclass Logistic Regression and Multiclass Neural Networks to analyse crime data from Baltimore, United States. The prediction accuracy results were between 71% to 90% (Kim and Jeong, 2021). However, the creation of a system which can predict the probability of criminals recommitting crimes based on demography and crime type, could be more effective in crime prevention as this has a level of specificity that can be highly useful to security forces. Serial crimes account for a large percentage of crimes committed in any demography. Using behavioural similarity with victim demographics for serial crime detection can be useful in policing prevention strategies (Ali, Alvi and Rehman, 2019). Analysing crime evidence and victim demographic data from the National-Incident-Based-Reporting-System (NIBBS) and applying Linear Regression, KNN, Logistic Regression, RF and Gradient Boosting algorithms, the resulting prediction accuracies were highest for Logistic Regression and RF at around 70%-80% (Rahman and Islam, 2022).

Deep Neural Network (DNN), an efficient decision-making model, was found to give good accuracy prediction in analysing a combination of data from the city of Chicago data portal, Weather Underground and Google Street View. Specifically, online databases of crime statistics, images, demographic and meteorological data was used to generate the dataset for this research (Kang and Kang, 2017). The result was high prediction accuracy of 84.25%, result of which was intended to assist crime prevention when integrated with police patrols. However, a more thorough study was on forecasting similar crimes, with focus on incident patterns. A study using Naïve Bayes (NB) to model a dataset of crimes in Cheltenham, United Kingdom, with attributes like location, crime type, criminal identification and acquaintances of criminals. The solution proposed was for mitigating crime incidents by data learning and analysing the results of crime prediction which indicated that property crime was most prevalent with a prediction accuracy of 74.4% (Kumar and Nagpal, 2019).

A deeper investigation in this field of study used three fundamental Cathodoluminescence (CL) architecture configurations: spatial and then temporal patterns, temporal and spatial patterns, temporal and spatial patterns in parallel, for forecasting crimes. Datasets used for this research work was a selection from law enforcement agencies in some cities including Minneapolis, Seattle, Philadelphia and Philadelphia. Analysis included attributes like location, event type and time the crime events took place (Stalidis, Semertzidis and Daras, 2021). The aim of the research was to create a setting where emerging hotspots are forecasted in advance by evaluating past crime events in those areas. The metric of accuracy was not used, instead the F1 score was preferred as it aided the accurate prediction of hotspots with scores of 93% for DT and 94% for SVM. Overall, RF performed better than the baseline methods using a binary classification approach. Similar research which had better crime prediction results was based on enhancing the Generalised Linear Model from Crime Site Selection, to analyse crime events in northern cities of India using modified Auto Regressive Integrated Moving Average (ARIMA). This provided a significant insight into the scope and complexity of crime prediction with ML models using the crime events obtained from the National Crime Record Bureau (NCRB) (Yadav and Kumari Sheoran, 2018). The results showed that the predicted values were close to real values, confirming the superiority of the model over other existing models.

The role of Machine Learning in predicting crime patterns and trends in Bangladesh was the direction of a recent research using a dataset obtained from the police’s website. The forecasted results and the actual results were compared along with an evaluation of the metrics for the different regression models which included LR, Polynomial Regression, Random Forest Regression (Biswas and Basak, 2019). The Polynomial Regression was found to outperform the other two regression models.

The use of reinforced ML and unsupervised learning to predict suspicious financial patterns can be a key tool to reduce financial crimes (Canhoto, 2021). Reconciling lack of expertise in application of ML models with the risks of leveraging on the algorithms for detecting and preventing money laundering, was the aim of this research. The research objective was achieved by using the ‘theory of affordances’ (an individual’s perception as related to the environment) to highlight technical attributes of anti-money laundering aided by ML models.

Forecasting crimes accurately can be instrumental in the creation of preventive policing strategies and a proactive approach to deploy police resources exactly where needed (Birks, Townsley and Hart, 2023). Another study used a hybrid model of Long Short-Term Memory (LSTM) and Space-Time Autoregressive Moving Average (STARMA) to analyse the San Francisco crime dataset, for prediction. The STARMA model dealt with non-stationary space-time data aspect with low performance, while the LSTM was solely for dealing with the trend and seasonal components (Liu and Lu, 2019). The result was compared to the outcome of using the traditional STARMA model. The hybrid model handled instability of the criminal data successfully and had improved precision of model fitting. Another study proposed that data mining techniques, which can find patterns and trends, could be a useful solution to data instability (Gowtham, 2018). Using data collected from Tripoli, Benghazi and Al-Jafara, the pre-processed data was successfully analysed using association rule mining and clustering to classify crime trends and behaviours, in relation to the age of the criminals in the 3 cities.

A concerning scourge of modern society is Gender-based crime, especially against women. Despite various governments’ economic investments to fight this crime, occurrence of this incidence has been on the rise globally (Amusa, Bengesai and Khan, 2022). Therefore, development of effective methods, particularly ML model tools for crime prediction will help in preventing such incidences. This was the basis of a study (González-Prieto *et al.*, 2021) which showed the high effectiveness of the ML method, Nearest Centroid (NC), for accurate and fair gender-crime forecasts.

## 2.2. Theoretical Foundations Supporting Machine Learning Models in Forecasting Crime

Determination of the best crime forecasting or predictive method, specifically by combining classification methods, has great value in crime deterrence. However, the theoretical foundations which support ML models in predicting crime basically stems from the idea that criminal behaviour is influenced by environmental, social and human behavioural factors (Chun *et al.*, 2019). Use of architected datasets created from original criminal records which included geographical and environmental factors, was obtained from some US cities, for analysis, with affirmative prediction called ‘hotspot’(Yu *et al.*, 2011). Using DT, SVM, One Nearest Neighbour, NB, with focus on residential burglary predictions, accuracies ranged between 82% and 88%. This was useful in achieving the aim to create a deployable ML model framework for a specific police department in a North-eastern US city. However, the data used for prediction was obtained from various city agencies, who had no training on data collation, therefore possibility of bias and errors in the data was high. Research using a San Francisco crime dataset was a better study for predicting crimes as the data was collected by the police force (Abdulrahman and Abedalkhader, 2017). KNN and NB were used for prediction, but their accuracies were low.

Future crime incidences can be well-managed with the aid of three types of information: environment data, location and time. These 3 form the basis of spatio-temporal crime patterns (Kounadi *et al.*, 2020). To forecast when crimes are most likely to occur, a genetic-fuzzy system was developed, encompassing crime patterns in data with features including location (longitude and latitude) and time (Farjami and Abdi, 2021). The proposed system used a simulated and real-world criminal dataset from Tehran, Iran. Since Mean Square Error (MSE) was not appropriate for evaluating spatio-temporal systems, Predictive Accuracy Index (PAI) was used for evaluating hotspots in this study.

The prediction of the likelihood of an offender to commit new crime was the objective of a research on how risk assessment tools or ML models can reduce crimes (Travaini *et al.*, 2022). Security professionals can routinely exploit predictions of recidivism risk based on ML techniques used to analyse past crimes of the offenders. However, the use of historical data for crime prediction can be plagued with issues of bias and privacy concerns. A gradient boosting model framework of a selection of appropriate ML algorithms including deep learning and linear models is another study for the prediction of spatial crime occurrences but without the use of past crime as predictor (Lamari *et al.*, 2020). Real world dataset of crimes reported in 11 cities in the US was used for predicting violent crimes and property crimes, with accuracies of 73% and 77%. However, an improved research of ML model framework for crime predictions, proposed an assembled-stacking based method (SBCMP) based on SVM algorithms and application of MATrix LABoratory (MATLAB) to identify appropriate forecasts of crimes. RF, NB and J48 models were also applied to the dataset which was a collation of crime reports provided by the National Crime Record Bureau in India (Kshatri *et al.*, 2021). The outcome showed that the SBCMP model gave the highest accuracy of 99.5%.

Distinguishing and analysing epistemologies like ML, Mathematical Social Science (MSS) and Social Physics (SP) of predictive policing, can be useful in generating crime forecasts. The ways future crimes are made knowledgeable for appropriate actions, directs law enforcement officers to certain crimes and offenders, which will satisfy the usefulness of the predictions (Hälterlein, 2021). This study posited that some ML approaches to crime predictions have the capability to go beyond the boundaries of existing knowledge of criminality with a disregard for the value of understanding criminal behaviours, which limits accurate forecasting. Focus was also on the MSS and SP perspectives of predictive modelling, with the former pointing to the need for an understanding of the human agency behind the mechanism and the later, for treating the humans behind the mechanism as separate entities. Hence, the need arises for including the guidance of the various epistemologies for accurate and valid crime predictions (Castro-Toledo, Miró-Llinares and Aguerri, 2023).

## 2.3. The Historical Evolution of Machine Learning Models to Forecast Crimes

Police departments in the United States have been leveraging machine learning models in predicting crimes over the years. Before ML, one of the first methods Police forces used was extrapolation of historical data. This worked well but the challenge was deviations from the crime trends could not be predicted (William *et al.*, 2023). As early as 1998, there have been attempts to use machine learning techniques to solve crime prediction problem, but the limitation then was the required computational power. Multivariate regression improved crime predictions but the drawback was it relied solely on historical data (Stec and Klabjan, 2018). With ML methods and the development of high-performance computers, there has been impressive results with various types of classification problems. CNN can successfully classify images which is important for spatial aspects of crime prediction while RNN can address temporal aspects of crime prediction particularly useful when the inputs and outputs are not independent of each other (Stalidis, Semertzidis and Daras, 2021).

The temporal relationship between crime incidences and weather, holidays and paydays using linear regression modelling indicated a dependence of aggressive crimes on the hour of the day and whether the incidences occurred indoors or outdoors. Hot or cold temperatures were also found to be influential in the upward or downward trend of crimes of aggression, anytime of the year (Towers *et al.*, 2018). This could improve short term predictions of crimes. Using a dataset from Chicago, USA covering crimes for 14 years and using bootstrap model selection method, the result of modelling showed that features like holidays, temperature, trends, has good predictive ability for varying crime types. In contrast, a more in-depth research based on similar variables in relation to crime predictions was about the critical role of data features, especially their individual contributions to crime predictions. The results of this analysis overcame the limitation of lack of interpretability of ML models (Zhang *et al.*, 2022). LR and Shapely additive explanation (SHAP) were applied to determine the contribution of each variable in a theft public theft dataset obtained from a coastal city in China. The variables with high SHAP values, had interpretable contributions to the ML model for crime prediction. A novel technique for crime spatiotemporal prediction, involving crime spots, criminal’s biography and geographic data of criminal events, was the aim of another study on serial crimes. Time delay Neural network with a fusion mechanism of activation function, specifically a Nonlinear Auto-Regressive with Exogenous Input (NARX) technique with fused activation functions (hyperbolic tangent and radial basis function), was utilised to analyse two Malaysian crime spatial datasets (Ghazvini *et al.*, 2020). A Back Propagation Through Time algorithm (BPTT) and an enhanced NARX (eNARX) were also employed for the analysis with eNARX presenting a higher accuracy of 82.15% compared to the NARX model’s 74.21% and BPTT model’s 49.71%.

Research analysing a dataset of crime reports from 2016-2018 from Calauan database in Philippines, was intended as a strategic contribution to lower the rate of crimes. The study deployed multinomial logistic regression (MLR), SVM and Neural Network (Asor, 2020). The best classifier to forecast crimes and predict crime locations was MLR with accuracies between 82.03% and 86.60%. SVM and Neural Network had lower accuracies between 52% and 78.09%.

Police activities follow two pathways: identifying signs of crime preparations followed by crime prevention. When certain conditions for crimes are met, its commission is dependent on certain conditions influenced by time and location (Uzlov *et al.*, 2020). An Adaptive Matrix Model was used for analysis of the dataset involving traditional crime analytical methods such as Crime mapping. The result showed accurate forecasts of crime patterns. This increases the probability of quicker actions by police when the specific hotspots are patrolled. But this study relied on data features like location and time, not considering features like geographical displacement which can influence crime patterns and forecasts. Crime predictions can be impacted by geographical displacement which when included in analysis, can improve forecasts of ML models (Wang *et al.*, 2019). Another research work proposed the significance of crime displacement in forecasting crime incidences by analysing crime records collected by the Public Security Bureau in China. Displacements were identified, aided by the Repeat and Near-Repeat (RNR) victimization phenomena. The quantitative result confirmed the importance of spatial-temporal displacement in improving accuracy of crime prediction, similar to the results of a comparable study on crime prediction using data of past crimes and transitional zones (Yang *et al.*, 2020).

Urbanization has been the source of economic development and on the other hand, cause of societal problems, especially crime. Another contributary study to the field of crime prevention involved testing Linear Regression, Logistic Regression and Gradient Boosting (GB) on criminal dataset of Saint-Petersburg in Russia(Ingilevich and Ivanov, 2018). This research addressed issues of overfitting with feature selection, which is important to machine modelling; with the GB model performing best as crime predictor.

Dynamic features like social diversity and population distribution, based on Criminology research findings, have been found to correlate with different types of crime incidences and can significantly improve the performance of crime predictions (He *et al.*, 2020). Crime incidences records of Queensland, Australia and New York City, USA, were collected and partitioned by census region. With Matrix Factorization modelling, the sparse dynamic features were estimated before modelling with an ensemble based learning network and RF, NN, SVM and Logistic Regression Model (Rumi, Deng and Salim, 2018). The models with dynamic features had better crime prediction accuracies as high as 98.7% when compared with the modelling without dynamic features at 91.2%.

## 2.4. Principal Machine Learning Models Used to Forecast Crimes

As the intensity and complexity of crimes is increasing rapidly, the need for effective ML tools like random forest, SVM, linear regression, K- Means clustering, ANN, NB, KNN, DNN and other different classification methods to analyse crime records, has been on the rise. These tools have been in demand by regulation enforcement corporations and policing agencies (Alparslan *et al.*, 2020). Some of these ML models are successful for achieving specific objectives in the field of crime prediction. SVM has found applications in handwriting identification and computational biology, image processing for crime detection and prediction. But SVM can be slow to train and predict, unlike NB which is easy to deploy and runs fast. NB assumes all features are independent which is not always so for crime datasets; this is how SVM proves to be a better ML model as it handles multi-class classification problems successfully (Suguna, Jeyavathana and Kanimozhi, 2022). SVM and RF take long to run unlike K- Means clustering. Reducing computational time of SVM and RF is achievable using optimized datasets with appropriate feature selection like embedded, wrapper and filter methods. Overall, each of the available ML models have their drawbacks and advantages, but datasets can be engineered to aid the effectiveness of any model to predict crimes (Shinde *et al.*, 2022).

A contributory study using Risk Terrain Modelling (RTM), was to predict crimes related to the homeless in California, USA. (Yoo and Wheeler, 2019). Homeless people have been linked to crime victimisation and criminal offenses, which is a challenging societal problem. Crimes related to homelessness can be reduced when placed-based risk factors (crime generators and crime attractors) are identified. The aim was to analyse the LAPD crime dataset to provide potential policing strategies to address future crimes committed by and against the homeless with prediction result having accuracy index above 17. Comparing capability of RTM in crime forecasting with other approaches, aided identification of crimes associated to homelessness. But using RTM with non-linear models was a more effective analytical strategy for this crime type. This was done with a Spanish crime dataset (Briz-Redón, Mateu and Montes, 2022). The crime counts predicted by the RTM was closer to the observed mean of the crime types, but the non-linear model predicted the crime count better.

Machine learning modelling of criminal histories were posited to provide better predictions and can be instrumental to prioritising incoming calls for service by the police. This was a research objective involving a Bayes classifier, RF and Logistic Regression to analyse a criminal dataset obtained from the Greater Manchester Public Protection Investigation database (Grogger *et al.*, 2021). Results of domestic abuse crime predictions of ML models based on conventional protocol-based technique called DASH (Domestic Abuse, Stalking, Harassment and Honour-Based Violence) risk assessment model was compared to that of ML models based on criminal feature histories. The ML models based on criminal history features performed better as some DASH attributes, collated from lengthy questionnaires filled under challenging circumstances, lead to poor data quality.

## 2.5. Machine Learning Models to Forecast and Prevent Definitive Crimes

## 

### 2.5.1. Terrorism

Terrorism, a global problem, has destructive effects which can be far-reaching, affecting not only people but also economies, therefore methodologies applicable in the prediction of acts of terrorism will be of immense value in preventing future attacks (Singh, Chaudhary and Kaur, 2019). Results of research work which used real time terrorist incidents data collection system for developing a risk model to calculate the risk level of terrorist attacks in different locations, can be helpful in developing counter-terrorism measures (Toure and Gangopadhyay, 2016). A novel algorithm with Latent Dirichlet Allocation was utilized for root cause analysis while the real time data was collected from selected data sources through Really Simple Syndication (RSS) Feeds. The result was a prediction system of terrorist incidents with maximum accuracy of 96.85% and provision of accurate risk values. But another research work with the aim to forecast locations susceptible to terrorist attacks was a better approach in the fight against Terrorism. The prediction accuracies with specificity and sensitivity scores were higher than 82% (Olabanjo *et al.*, 2021). This was achieved using an ensemble machine learning model which combined SVM, KNN and this was applied to data obtained from Global Terrorism Database. Two feature selection techniques, Chi-squared, Information Gain and a hybrid of these two techniques were used. However, it was the hybrid-based selected features which produced the best results with the highest accuracy of 97.81%, specificity score of 99.67% and sensitivity score of 92.2%. One drawback of this research was unavailability of predictions of countries susceptible to terrorist attacks, but the results provided useful predictions of continents like Asia, North America and South America, as possible future targets of terrorists.

Another contributary research effort in the fight against terrorism in the US, utilised news data for predictive analysis. The ML models were trained to learn from news data collected from the Global Database of Events, Language and Tone (GDELT), to predict if a terrorist attack will occur on a specific date and in a specific state (Krieg *et al.*, 2022). RF aided by a novel variable-length moving average method, achieved the best performance with an Achieved Area Under the Receiver Operating Characteristic (AUROC) above 0.677 which was better than Feedforward Neural Network (FFNN) and LSTM. Another paper took a better approach as it posited that democratic institutions could influence transnational terrorist attacks. Utilising Logit model to forecast the terrorist attacks by analysing dyad year cases of 132 targets of transnational terrorist attacks between 1968 and 2007, drawn from the Expected Utility Gene data management software (Gelpi and Avdan, 2018). The result showed that variables for democracy and Gross Domestic Product (GDP) when excluded from the model, hardly had an impact on the crime forecasts. However, distance between nations and time between terrorist attacks, reduced the predictive ability of the ML model.

### 2.5.2. Cybercrime

Currently, priorities in cybersecurity are more focused on the need to identify future threats and their perpetrators. The application of Machine Learning technologies (MLT) specifically automated computer algorithms, are beneficial in discovering valuable crime patterns which manual methods using individual human intelligence, might not find (Ch *et al.*, 2020).

Trained ML models including RF, DT, LR, NB and KNN were deployed to detect sophisticated cyber-attacks which were conducted against a developed Supervisory Control and Data Acquisition (SCADA). This testbed consisted of a water storage tank’s control system and network traffic was obtained with its features extracted to generate a dataset used for cybercrime prediction (Teixeira *et al.*, 2018). A second analysis was performed using online network traffic and the resulting predictions were compared to the results of the analysis of the built, offline dataset. False Alarm Rate (FAR), Un-detection Rate Metric (UND) and accuracy were the metrics used for model evaluation. Accuracy of all the models for the offline data performed excellently, same for the online data which was as high as 99%. However, KNN had lower accuracy of about 72%.

Though ML models can identify future cyber threats, there is need for legal framework as to what actions are allowed by private actors or security agencies which was the objective of another study (Ghappour, 2015). The research proposed how a combination of ML model-driven technologies and socio-legal mechanisms can be used for achieving trustworthy predictive analytics in cyberspace.

Predicting cyberattacks is becoming increasingly important for organisations and has become more important to cybercrime analysts and security administrators. A new method which interprets multi-step attacks from a dataset to provide a universal attack prediction model is a valuable contribution to cybercrime prevention (Fredj, 2022).

### 2.5.3. Financial crime

Crimes like credit card fraud, has been increasing significantly with advanced technology, making detection challenging. However, supervised machine learning models can be effective in detecting and preventing this type of financial crime (Dhankhad, Mohammed and Far, 2018). Another study analysed a real-world dataset with ten ML models including KNN, Gradient Boosted Tree Classifier (GBT), RF, Stacking Classifier (SC) and NB. RF accuracies were as high as 0.94594, while the lowest accuracy was NB at 0.90540. This in-depth research also provided evaluation of the models using F1-Score, Recall, G-mean, Precision and Specificity with RF and SC having the best values. One benefit of Anti-money laundering (AML) solutions integrated with crime detection ML models is that it enables users to investigate financial transactions and flag suspicious activities. Research on the feasibility of ML models like SVM, NB, Neural networks in detecting and preventing financial crimes is beneficial to financial organisations as with their application, there will be reduction of monetary loss (Chen *et al.*, 2018). There are commercial crime detection software like NICE Actimize, Logica ISL and Lexis-Nexis, incorporates ML models which can filter and flag suspicious transactions based on pre-determined thresholds and rules (Sobh, 2020). However, drawbacks include the over-sensitivity of these ML-model based software, racial bias and issues of data privacy.

### 2.5.4. Urban crime

The research involving the analysis of Bahir Dar City Administration Criminal Database in Ethiopia to predict urban crime involved time-consuming data preparation, as the dataset was organized in summary form and only available as hardcopy. The project had 3,000 criminal record samples manually collated for analysis (Alemu, 2020). Instead of using the 15 attributes in the dataset, only 7 were required. The rest were deemed irrelevant for Naïve Bayes, Multi-layer Perceptron (MLP) and KStar classifiers. The prediction accuracies were low at 42.9% and 57.5% which was likely because all the crime groups were used for ML modelling. However, with only 3 crime types: theft, robbery and social crimes, the accuracy of crime prediction improved to 71%. This data selection is a drawback to the validity of accuracy. Research utilising a Bangladesh crime dataset for urban crime forecasting, was not as time-consuming since a soft copy of the data was available, an advantage over the Ethiopian crime dataset analysis. Implementing Linear regression model for crime types: robbery, dacoit, theft (put in a single group), murder, women and child repression in different regions of Bangladesh, resulted in good, predicted values (Awal *et al.*, 2016). Number of murders in 2014 was 792 in Chittagong Range and the predicted value was 757; in the Dhaka Range the actual crime occurrence was 1,395 while the predicted value was 1,243, showing Linear Regression is a better model for urban crime prediction.

## 2.6. Evaluation of Machine Learning Models for Predicting and Preventing Crimes

Top measures of evaluation performance of ML models for prediction includes Prediction Accuracy (PA), Prediction Accuracy Index (PAI) and F1-Score while the validation approach commonly implemented is the Train-Test Split. Others are cross-validation, Leave One Out and the Rolling Horizon (Kounadi *et al.*, 2020). Spatial crime forecasting involves the use of some of these metrics but the top 3 are PA, PA1 and F1-Score with PA implemented mostly by Computer Scientists while Criminologists prefer PA1. For short-term crime predictions, PA and F1-Score have been the metrics for evaluation of ML models. Crime data of incidents in Arizona, USA, involving guns, was the basis of a study to find differences in prediction techniques using metric comparison (Drawve, Moak and Berthelot, 2016). Nearest Neighbour Hierarchical (NNH) hot spot technique and Risk Terrain Modelling (RTM) were specifically used as prediction methods. The research extended the application of other studies to encompass all gun crimes and not only those involved in shootings. The outcome showed the RTM which was constructed from social (spatial risk factors per cent black with drug density) and physical measures, had better precision and reliability over time than the NNH technique.

Increasing changes in crime patterns with limitations of current methods for evaluating ML models, can be highly disadvantageous in crime detection. For local police departments intending to incorporate ML models into their crime prevention strategies, factors like education rate, local time and weather could be considered. They have varying impacts on crime rates and thereby, can limit the accurate evaluation of these models (Albahli *et al.*, 2020). This requirement led to a research involving 1-D convolutional neural network(CNN) and RNN with a proposed fusion technique to analyse online weather information and crime dataset released by the Los Angeles police department in the U.S (Zhou *et al.*, 2021). Training 2 fine-tuned convolutional units and perceptron unit with the dataset, the crime prediction was in real-time and was close to the general trend in the specific period. Root Mean Squares of the plain CNN and RNN models were compared to that of the Mixed Spatio Temporal-Neural Network which performed better. Downside of this research was overfitting which could have been improved by hyperparameter tuning of the 2 networks. Crime datasets often have imbalanced classes with some crimes occurring more than others. This can be an issue for Neural Networks which do not handle imbalanced datasets well (Pavan Naidu Kavala *et al.*, 2022). Therefore, unavailability of high quality, large datasets is another limitation for evaluating supervised ML models employed for crime detection (Canhoto, 2021). But when the size of the crime data for prediction is small, time series forecasting can be highly effective, especially in determining potential increases in criminal activities. Relevant to this, is research work which showed how cybercrimes increased beyond predicted levels using ARIMA models applied to a UK crime dataset culled from crime reports between 2017-2020 (Kemp *et al.*, 2021).

## 2.7. Ethical Concerns and Potential Biases in Application of Machine Learning Crime Models

The aim of an in-depth research study on various biases which can occur during the product life cycle of ML algorithms implemented for Policing within England and Wales was to inform The Centre for Data Ethics and Innovation in their review of bias. Also to review its negative impact on crime prediction (Babuta and Oswald, 2020). The findings of this research included systematic bias in decision based on a skewed predictive output when the ML model only used data confirming risk instead of data showing the opposite. Human bias introduced into such datasets will influence the outcomes of the application. In more specific research on Predictive Policing in the United States, the approach to lay out five ethical challenges facing this type of approach was cautious due to the complexity between policing and race (Davis *et al.*, 2022). These challenges include Racial Bias and Civil Liberties, Data purpose, Standard of performance success, Transparency and Accountability with Community oversight. Despite these challenges, in the United Kingdom, mass public surveillance is predominantly enabled by facial recognition-enabled close circuit television (CCTV) cameras. The data obtained from it, is employed for predicting policing. Similar software has been used in about 50 US police departments in 2013 and more since then. With these, they were able to predict when and where a crime will take place based on CCTV data analysed with PredPol (now called Geolitica), a data analytics crime prediction software (Montasari, 2023). The 10-25% accuracy achieved was better than traditional hot spot mapping. But with this progress in crime forecasting, there has been unanimous voting banning the use of the software in several US cities since 2020 due to racial bias. Regardless of these resulting developments, Kent and Greater Manchester police with 12 other police forces in the UK have recently adopted PredPol with significant improvement in crime prediction and reduction recorded (Castets-Renard, 2021). Several of these police forces are not unaware of inherent bias in predictive ML model-driven software. For this reason, they take these products to independent ethics committees and panels before deployment. But one drawback regarding this is the implicit pressure to approve such crime predictive software as oversight committees take considerable time in their assessments (Zilka, Sargeant and Weller, 2022).

Artificial Intelligence (AI) policing driven by ML models, unlike human reasoning, have been found to predict crimes based on the desire to forecast which is served through its self-reinforcing effects, sometimes producing inaccurate results (Kaufmann, Egbert and Leese, 2019). This has generated much criticisms of predictive policing in some communities (Helm and Hagendorff, 2021). Issues like privacy concerns regarding the historical data analysed for crime predictions. Also, demonstrable accusations of bias on the part of the police which can influence the crime data fed to the ML models in use. In addition, these ML algorithms employed by commercial systems are sometimes subject to trade secret protections. Though the argument is that ML models for crime prediction makes policing smarter for Patrol officers, Judges, Parole boards (Young, Bullock and Lecy, 2019) but the lack of responsibility and intrusion into personal rights fuels criticism by some civil rights organisations. Wielding ML algorithms like they are not ethically neutral, could be an effective approach to ensure the awareness of the way these algorithms reproduce existing injustices and contribute to new ones. Predictive policing using ML models relies on historical data to generate forecasts which can produce a self-fulfilling prophecy of repeat criminal activities based on a racialised paranoid rationale. This was the direction of a research on person-based algorithms for predictive policing. One of these is Chicago’s Strategic Subject List (SSL), which is premised on a group of possible actions for pre-empting crimes (Sheehey, 2019).

Technology-based policing in South Korea was the foundation of a project analysing big data-driven crime prevention strategies. Previously accumulated data was deployed in a program called Crime Layout Understanding Engine used for predicting crimes (Lee and Park, 2022). In addition to this, a crime prediction model was developed by the Korean National Police Agency which examined big data collected. This helped in the strategic placement of patrol cars in areas and at times where crime was expected to occur. But from a survey, there were privacy issues regarding the data of citizens used for crime forecasts, especially concerns if they were collected and utilised legitimately. Another insight from this study was the perception of procedural fairness was negative and did not encourage the cooperation of citizens in non-face-to-face crime deterrence situations.

## 2.8. Future Research and Development for Machine Learning Models Utilised in Forecasting Crimes

A recent study was on the application of novel deep learning model-Deep Inception-Residual Network (DIR Net) to forecast theft-related crimes. This involved inception units of asymmetrical convolution layers (Ye, Duan and Peng, 2021). These layers mapped low-level spatiotemporal dependencies hidden in crime incidences in New York City from 2010-2015. The resulting F1 score was 71%. This was better other ML models like NB, DT, SVM used for the same research. Compared to other ML modelling by other researchers using the same dataset, a**deep residual network**based on**information refinement** (DIR) Net was replicated for crime predictions in other cities which lacked various types of auxiliary data. This was useful in investigating the “why” question and not only the “where” and “what” of the criminology of place. However, a more comprehensive development in the deployment of ML models in crime prevention involved a gradient boosting model which gave the most accurate predictions of 73%-77% for violent crimes, property crimes, motor vehicle thefts from real-world crime datasets in 11 US cities. The study involved the implementation of a crime prediction model in three-dimensional (3D) CNN for predicting road traffic offenses as representative for driver behaviour offenses. With an accuracy rate of 96%, the 3D CNN is a development which is instrumental in crime prevention related to road traffic offences (Fan, 2021).

Large datasets are usually required to achieve the best predictive performance of supervised ML models. However, when the dataset is small, these models have low predictive performance, hence a solution was posited by some researchers who trained a small dataset using Deep Reinforcement Learning (DRL), which relies on the parallel computing performance of Graphics Processing Unit (GPU) (Lim, Abdullah and Jhanjhi, 2021). The crime dataset was a cocaine smuggling dataset which contained four criminal groups. It was analysed with RF, SVM and Gradient Boost model. The performance of the GPU-based DRL-Criminal Network Analysis model was better by 7.4% than the next best model which was RF model. But this study did not capture spatio-temporal dynamics which can be important in crime forecasting task of analysing a dataset of cocaine smuggling. This problem was addressed in a research work for analysing the San Francisco crime dataset using Graph Convolutional Neural Network (GCN) and Recurrent Neural Network (RNN) combined in a Temporal Graph Convolutional Neural Network (T GCN) (Jin *et al.*, 2020). The dataset was also trained with XG Boost, RF and Convolutional-LSTM. The prediction performance of TGCN was best when evaluated with RMSE, Jessen Shannon Divergence (JS) and Mean Absolute Percentage Error (MAPE).

Predictive performance of ML models has been used in forecasting pretrial recidivism. However, several common ML methods result in black box models which though successful, are difficult for human understanding (Angelov *et al.*, 2021). The application of interpretable models that produces probabilities instead of binary forecasts, was the objective of a study which generated black box ML models of a Florida and Kentucky criminal recidivism dataset (Wang *et al.*, 2022). Risk-calibrated Super sparse Linear Integer Model (Risk Slim), Classification and Regression Trees (CART) and Explainable Boosting Machine was utilised for analysis. This was compared with the predictive performance of the Arnold Public Safety Assessment (Arnold PSA) and Correctional Offender Management Profiling for Alternative Sanctions (COMPAS). The interpretable ML models performed better than the two existing risk assessments.

## 2.9. Comparison Of Machine Learning Models and Established Techniques for Crime Predictions

This encompasses various City and Community Policing crime prediction strategies.

### 2.9.1. City Policing Plans Driven by Crime Prediction ML Models

Crime data from a large coastal city of China was used to compare the predictive abilities of LSTM, RF, SVM, NB and CNN (Zhang *et al.*, 2020). The dataset for this research was made up of public property crimes. The objective was to predict occurrence of these crimes. LSTM performed best. This is due to its superiority as better pattern and regularity extractor. The LSTM which had built environment covariates also had a better crime prediction accuracy than the original model. However, one challenge of this research was the temporal resolution of the crime prediction. The impact of crime variations captured was within a two-week window, whereas some studies have shown the usefulness of capturing daily changes in crimes or checking variation of risks during daytime. This aids temporal resolution of crime predictions. A research examining a Chicago crime dataset had the aim to not only predict crimes but also to gain insight on how the network of city community areas shapes dynamics of crimes (Niu *et al.*, 2019). The results obtained showed how certain urban factors were likely triggers for the spatial spreading of criminal activities in that specific US city. There has also been evidence of how ML models incorporated with unique demographic characteristics of the inhabitants performed better than models which relied only one type of feature (Wheeler and Steenbeek, 2021). Comparing ML model-driven techniques to traditional methods to improve crime, was the proposal of a recent study on how UK community-based programs like the Neighbourhood Renewal Fund (NRF) had been effective in reducing crimes since the early 2000s (Alonso, Andrews and Jorda, 2019). The outcome of the Differences-in-Differences (DiD) and Regression Discontinuity (RD) analysis indicated reduction in occurrence of domestic crimes, vehicle crimes and burglaries in small, disadvantaged neighbourhoods which received the NRF from the government.

Reducing situational violent crimes with an evidence-based approach informed by Risk Terrain Mapping (RTM) and hotspot analysis, was a policing initiative by the Kanas City (KC), Missouri Police Department. The one-year Risk-based policing (RBP) initiative which involved multiple community stakeholders, successfully decreased crimes by over 22% even without many law enforcement actions against people in the predicted target areas (Caplan *et al.*, 2021). The success of this crime-prevention initiative by the police can be attributed to the training of KC Police Department staff and civilian crime analysts who collaborated on the implementation of the program which relied on an RTM software. This collaboration of the Police and civilian crime analysts could be a solution to the challenge of clear communication between the analysts and the security officers. This was one focus of a study on recent developments in deployment of predictive policing software integrated with ML models (Hardyns and Rummens, 2018).

### 2.9.2. Community Policing Crime Prevention Programs Incorporating ML Models

Traditional methods like Sum-Score screening or dynamic risk score assessment for targeted prevention of youth criminal behaviour problems, did not adequately address the proliferation of delinquency in the society. There were problems with prediction accuracy and bias (Lloyd *et al.*, 2020). Application of ML models like Logistic Regression, SVM and RF, accurately predicted delinquency outcomes in the analysis of data collected from the Pittsburgh Youth Study in the US. This was better than the results of the Sum-Score method (Pelham, Petras and Pardini, 2020). The study was beneficial in identifying youths likely to commit crimes, which is useful to security agencies in crime prevention.

Predictive ML models were integral to the success of a Tulsa Community-Based Crime Prevention program, a collaboration between the police and community stakeholders like business owners, local government and researchers (Corsaro and Engel, 2020). Criminal incident data obtained from the Police, was analysed with time-series technique, based on Generalized Linear Modelling count regression method. The result was that the targeted areas showed notable reductions in crimes when this community-based crime prevention were deployed. Another study contributed to the field of gun violence prevention, based on the collaboration of technology and social media insights (Twitter data), collated during the recent pandemic (Patton *et al.*, 2022). In this transdisciplinary study, data was from a Neighbourhood Navigator project which encompassed focus groups, interviews and social media data. Analysis was with NLP techniques. The aim was to provide recommendations on how gun violence can be predicted by analysing social behaviours of residents of marginalised communities, who regularly engage on social media platforms, contribute to gun violence. Another study using ML models to forecast violent crimes took a different approach, analysing the relationship between urban vacancies (vacant lots) and violent incidences. Models like Multiple regression model and simultaneous regression model were applied to train the dataset obtained from the Michigan Youth Violence Prevention Centre from 2017-2018 (Burt *et al.*, 2022). Findings showed that urban vacancies had a direct impact on crime prevalence in those areas.

## 2.10. Case Studies and Practical Applications of Machine Learning Models for Crime Forecasting

The application of a three-structure approach to enable accurate prediction of crime incorporated the use of a multi-modal fusion approach with a fusion Deep Neural Network (DNN)-based prediction model (Kang and Kang, 2017). The combined datasets were made of Occurrence Reports obtained from online database portal of the City of Chicago, American Factfinder, Weather Underground and Google Street View. However, though the crime prediction from this modelling was highly accurate, one limitation of the DNN model was that it was not effective for regions with limited data. The prediction of a specific crime, at a given time was also an impossibility for this Machine Learning (ML) model. These limitations were addressed by the application of ML models: K-Means clustering and Naïve Bayes, which could predict crimes (in relation to the time of the day), regardless of the data size (Palanivinayagam *et al.*, 2021). This unique research contribution addressed the drawback of the DNN model.

Location, hour, month and coordinates are selected features which can be beneficial for accurate crime forecasts using ML models. This was the direction of a research work modelling an Indian crime dataset with KNN algorithm (Balu, Navya Sri and Bupathi, 2022). The accuracy of the ML model was high but the process of grouping many classes together for easy categorization of the crime dataset was a difficult and time-consuming classification task. Instead of grouping many classes, researchers who analysed a San-Francisco dataset for crime prediction, created new features. They extracted from the feature Day, creating blocks of time and added this to the other features, providing a clearer picture of time with linear data which was better than 24-hour-cycle-time (Hossain *et al.*, 2020).This analysis was uncomplicated and not time-consuming as univariate statistical test was applied to find features strongly related to the target variable. However, the San-Francisco crime dataset was imbalanced, giving poor accuracy score when compared to the Indian crime dataset which had equally distributed classes.

Trained professionals have had issues with the prediction of human behaviours like criminal recidivism in forensic risk assessment as humans do not necessarily make rational decisions. However, combining human abilities of trained professionals with computer-aided tools, specifically ML models, may help improve risk assessment of offenders and thereby prevent crime (Kaur *et al.*, 2020). Correctional Service of Canada had records of more than 72,000 individuals who had been sentenced to prison for varied time periods. (Ghasemi *et al.*, 2021). Analysis of this data with supervised algorithms: DT, RF, Support Vector Machine, was helpful in predicting individuals who were likely recidivists. Classifications were grouped as High Risk and Low Risk. All four ML models had high AUC (Area Under Curve) scores for the two classes of Risk. The accuracies were compared to the results of Level of Service/Case Management Inventory (LS/CMI) tool used by the Correctional institutions to assess criminogenic risk-need individuals. Decision Tree model provided a better accuracy score compared to the LS/CMI. The limitation of this research was the lack of uniformity between the data samples and possibility of sample bias during data collection and analysis. On the other hand, similar research using a Super sparse linear regression model (SLIM) for recidivism predictions produced accurate, fair and interpretable scoring system. This can be helpful for analysts to decide which type of individual input variables will be combined for prediction and determine if they align with ethical values (Zeng, Ustun and Rudin, 2017). Recidivism like Drug, General Violence, Domestic Violence, Sexual Violence and Fatal Violence were crime problems analysed in this research. SLIM’s predictions were compared to that of RF, SVM, Ridge Regression. In terms of accuracy, all the models had similar scores, however, SLIM for recidivism predictions circumvents issues of bias and lack of independent fairness of other ML models.

In a recent project, ML models like DT, RF, SVM, KNN, were applied to investigate a Philadelphia crime dataset. Different features were extracted to predict crime types. Specific crime incidents such as sex offenses, vandalism, thefts, were aggregated over hours, months and years. But instead of centring the data analysis on the city structure and distribution of neighbours, the focus was on angular measure of the crime point like longitude and latitude and dispatch time of crime (Alparslan *et al.*, 2020). All the ML models applied had good performance but low accuracy scores. RF was the best of the models with a log loss of 2.2323 while KNN had the worst log loss of 19.7031. Combining supervised and unsupervised learning techniques, even with no mapping, produced results which could be generalized to other cities. However, another study incorporated geographic mapping of a city into a grid, for crime hot spot forecasting. Police records of auto theft and other crimes provided by Portland, Oregon Police Bureau (PBB), was collated in a Call-for-service (CFS) data (Zhuang *et al.*, 2017). The research involved constructing 3-state-of-the-art RNN architectures: RNN, LSTM and Gated Recurrent Units (GRUs) with Spatial-Temporal Graph Neural Network (STNN). The prediction accuracies were high at 70%-81% which was better result than the research which excluded geographic mapping of the area covered for the crime data.

# **[Chapter 3](#_Toc118461390)** **-** Research Methodology

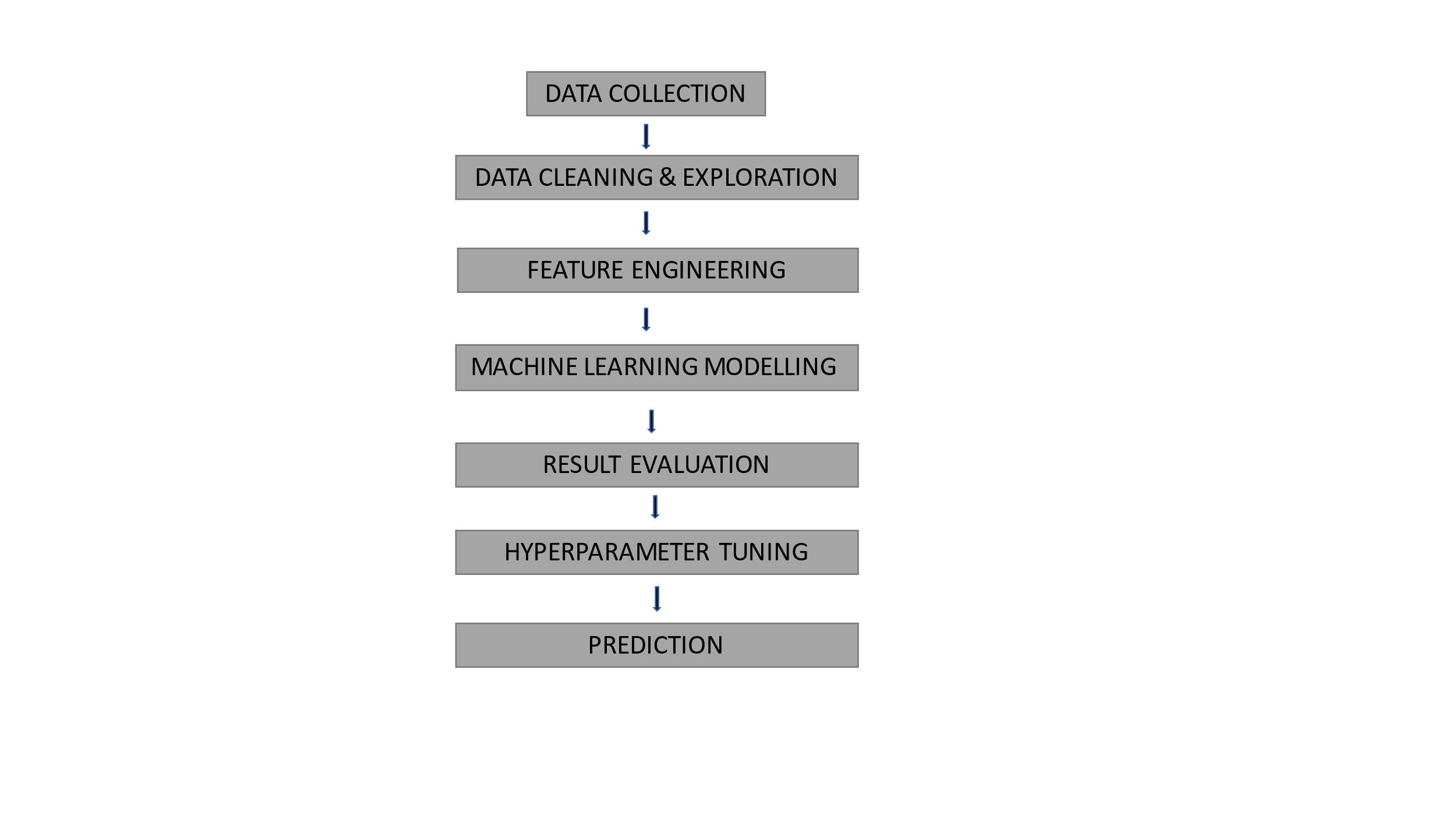


Figure 1. Research methodology workflow.

*Source: The Author. Adapted for the Vancouver crime dataset (2003-2017).*

**Data Collection**

In Figure 1., data collection is the first step. The Vancouver Police Department (VPD) made the crime dataset used for this research work, publicly accessible on its public data portal. It covers reported crimes from 2003 to July 2017 in the city of Vancouver, British Columbia, Canada. The dataset features include crime type, violent or non-violent crimes, neighbourhood, date, time of crimes and location data, which was anonymised to keep the privacy of the individuals involved in the incidents.

**Data Features**

**Crime Types:** Varying crime types such as theft, assault, homicide and robbery.

**Location:** Approximate location (in longitude and latitude) of crime occurrences with the exclusion of private details of individuals involved to protect their privacy.

**Date and Time:** The month, year and time of occurrence for each reported crime.

**Violent and non-violent crimes:** The dataset has classes of violent and non-violent incidences.

The crime data was pre-processed starting with Data Cleaning, Data Exploration, Feature Engineering. Then ML modelling, Result Evaluation and Hyperparameter tuning to obtain the best accuracies of prediction.

**Data Cleaning and Data Exploration**

There are 530,652 rows and 12 columns of crime records from 2003-2017. The data types and their unique attributes includes integers and float values.

Data cleaning involved dropping the ‘Minute’ column as it was not necessary for analysis and columns ‘Hour’, ‘Hundred Block’ and ‘Neighbourhood’ had null values which were replaced with dummy value of ‘100’ and ‘N/A’. Also, as the data was collected only to 13th of July 2017, meaning an incomplete month, July was removed from the dataset.

**Feature Engineering**

Transforming features ‘Year’, ‘Month’ and ‘Day’ which were combined into one column named ‘Date’ for better analysis. Table 1. Below shows ‘Theft from Vehicle’ was the most occurring crime while ‘Homicide’ had the least occurrence.

**Table 1. Occurrence of each crime type**

|  |  |
| --- | --- |
| **CRIME TYPES** | **OCCURRENCE** |
| Break and Enter Commercial | 33787 |
| Break and Enter Residential/Other | 60768 |
| Homicide | 219 |
| Mischief | 70253 |
| Offence Against a Person | 54035 |
| Other Theft | 51995 |
| Theft from Vehicle | 172265 |
| Theft of Bicycle | 25597 |
| Theft of Vehicle | 38359 |
| Vehicle Collision or Pedestrian Struck (Fatality) | 254 |
| Vehicle Collision or Pedestrian Struck (Injury) | 21871 |

**Data Visualisation**

Chart

Description automatically generated

Figure 2.Occurrence of each crime type

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017)*

The number of crimes between 2003 and 2008, is above average with a spike just after 2008. There is a huge spike in June 2011, above average, as seen in Figure 2. Grouping ‘Break and Enter’ crime, ‘Theft’ crime and ‘Vehicle Collision’ crime before analysis of the spike in June 2011 showed how many crimes occurred with crime types.

**Table 2. Occurrence of the grouped crime types**

|  |  |
| --- | --- |
| **CRIME TYPE** | **OCCURRENCE** |
| Mischief | 809 |
| Theft from Vehicle | 521 |
| Offence Against a Person | 314 |
| Break & Enter Commercial | 303 |
| Other Theft | 301 |

Transforming the features of the dataset is important to successful ML modelling for

predictions. The eleven crime types were grouped into five and from Table 3. ‘Mischief’

had the highest occurrence in the dataset.

**Table 3.Occurrence of grouped crime types per neighbourhood**

|  |  |
| --- | --- |
| **DISTRICT** | **OCCURRENCE** |
| Central Business District | 996 |
| N/A | 322 |
| West End | 189 |
| Mount Pleasant | 139 |
| Renfrew-Collingwood | 125 |

The above Table 4. shows the crimes per District/Neighbourhood and their occurrence per day in June 2011. Central Business District (CBD) had the highest number of crimes at 996. Hence further exploration of this result is important to understand the Neighbourhood.

**Table 4.Occurrence of crimes in Central Business District**

|  |  |
| --- | --- |
| **DATE** | **OCCURRENCE** |
| 2011-06-15 | 649 |
| 2011-06-16 | 111 |
| 2011-06-10 | 110 |
| 2011-06-17 | 105 |
| 2011-06-18 | 103 |

The number of crimes in the CBD, in June 2011, is shown in Table 5. The highest number of crimes occurred on the 15th of June in 2011, at 649. This was a huge increase from 111 crimes a day. Also, from the information available online, there was a public riot: Stanley Cup riot in Vancouver, which occurred after a Hockey game. Cars and business were destroyed which explains the high occurrence of crimes of ‘Mischief’. Figure 3., is a pivot table depicting days with highest and lowest average of crime occurrences.

A picture containing text, screenshot, number

Description automatically generated

Figure 3. Average number of crimes daily and monthly.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

The average number of crimes in each day and month of 2011 is represented in Figure 3. Red depicts high average of crime occurrences; Blue depicts low average and White is in between them. Specifically, black depicts very low crime occurrence and dark red, very high crime occurrence. December 25th has low number of crimes occurring while January 1st and October 31st (holidays in Canada), on average has very high occurrence of crimes. April to October, which is around Summer, shows high average of crimes. Months with the highest and lowest average of crimes is seen in Figure 4.

Chart, line chart

Description automatically generated

Figure 4.Average of highest and lowest crimes per month.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

The above plot in Figure 4. shows crimes decreased from around 4000 in 2003 to about 2500 in 2011 and had nearly the same moving average from 2011 to 2014. There was an increase from 2014 to 2017 which then declined afterwards.

Chart, line chart

Description automatically generated

Figure 5. Average of crime categories per month.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

Figure 5. is a line plot with the categories of crimes per month with the Moving Average method. The category ‘Theft’ has a high occurrence around November 2003 at above 2000 which declined to about 1,200 around October 2011.This increased from 2013 to 2016 and reduced from the start of 2017. ‘Vehicle Collision’ had the lowest occurrence at about 200 in November 2003 and declined slightly from the start of 2006 to January 2007. There was no sharp increase from 2007-2017 as the occurrences of crimes in this category remained at around 150. ‘Break and enter’ and the crimes in the Category: ‘Others’ were occurring at about 800 at around the end of 2003 and ‘Break and Enter’ occurrences dropped 2006, rose slightly in 2007 and reduced till ending of 2016. Like ‘Theft’, it fell slightly in beginning of 2017. ‘Others’ occurred within the same range of about 700 to 800 from 2003 to 2017 but unlike the other 3 categories, it increased slightly in occurrence from around 2017.

Figure 6. is a different representation of crime categories each month using a heat map. The crime ‘Break and Enter’ had the highest occurrence in January and the lowest frequency in December. But ‘Theft’ and the crimes under ‘Others’ occurred less in January but more in May and June.

A picture containing text, screenshot, colorfulness, line

Description automatically generated

Figure 6.Average of crime categories per month.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

Crime occurrence per day in Figure 7. using histogram showing a normal distribution with an average of about 92 crimes each day with outlier above 600.

Shape

Description automatically generated

Figure 7.Histogram of crime distribution per day.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

Figure 8. represents line plots of crime types in each year recorded in the dataset. Occurrences of ‘Other Theft’, ‘Theft from Vehicle, ‘Homicide’, ‘Mischief’. ‘Theft of Bicycle’ and ‘Vehicle Collisions’ showed sharp spikes in occurrence between 2014 and 2016. Also of importance is the decrease of all crime occurrences after 2016.

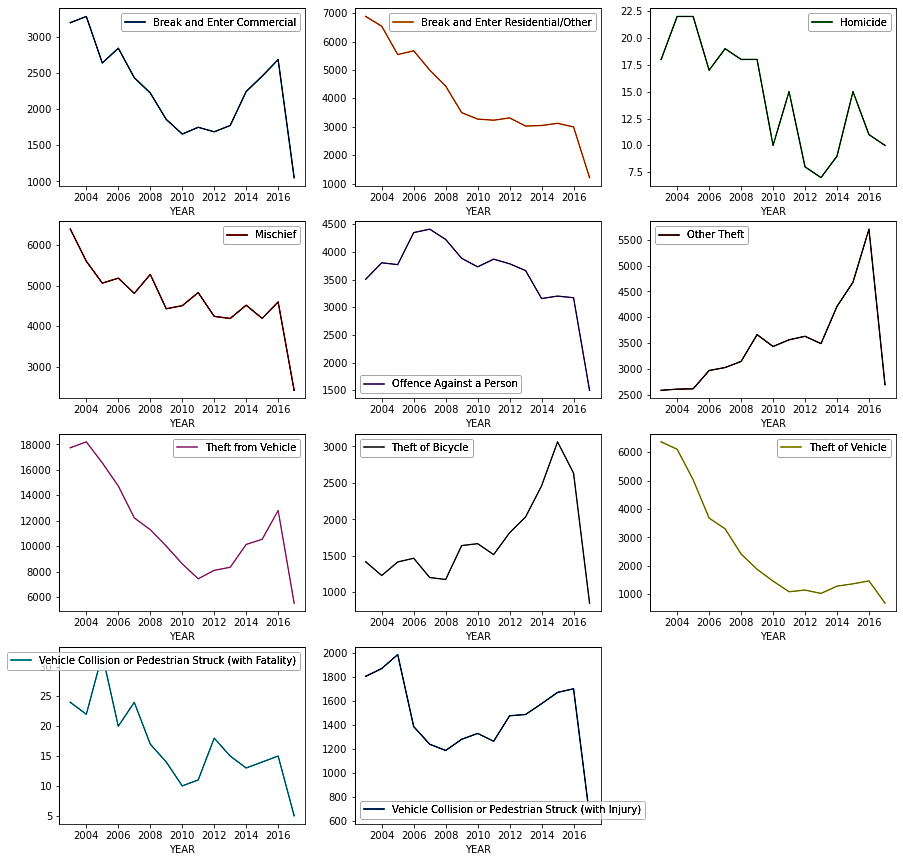


Figure 8.Line plots of each crime per year.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

A bar plot showing number of crimes by day of the week shown in Figure 9. Highest count of crimes occurred on Saturday followed by Friday, at above 8000 while the lowest count of crimes is on Wednesday at about 7,200.

Background pattern

Description automatically generated

Figure 9.Number of crimes per weekday.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

Total number of crimes per day in each year is represented in the Figure 10. This control chart with the Upper Control Limit (UCL) and Lower Control Limit (LCL) is useful in this visualisation to determine the average number of crimes each year. Between 2003 and 2008, this is above average with a spike just after 2008. June 2011, there is a huge spike above average which can be attributed to the Stanley Cup riots that year.

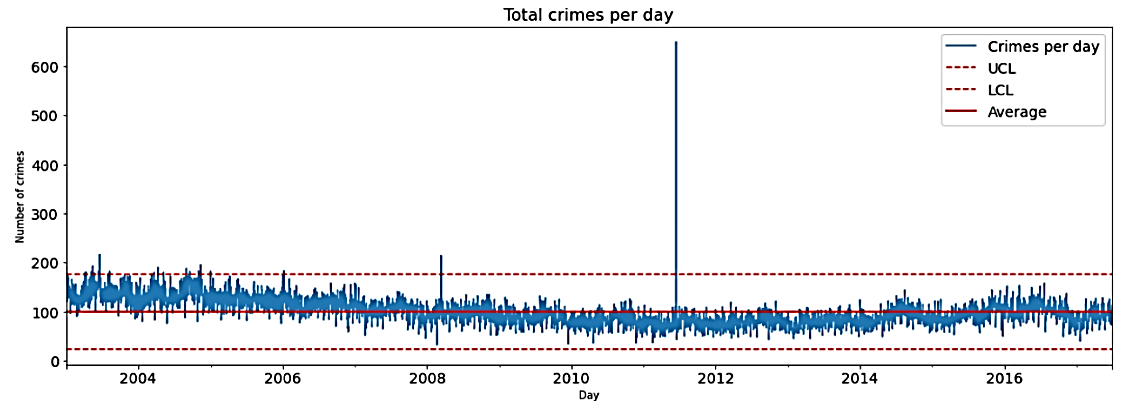


Figure 10.Total Crimes Per Day of Each Year.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

From the bar chart in Figure 11. crime of ‘Theft from Vehicle’ has the highest occurrence at around 160,000 while the lowest occurring crimes were ‘Vehicle collision’ or ‘Pedestrian struck with fatality and Homicide’.

**Fig 11. Count of crime types**

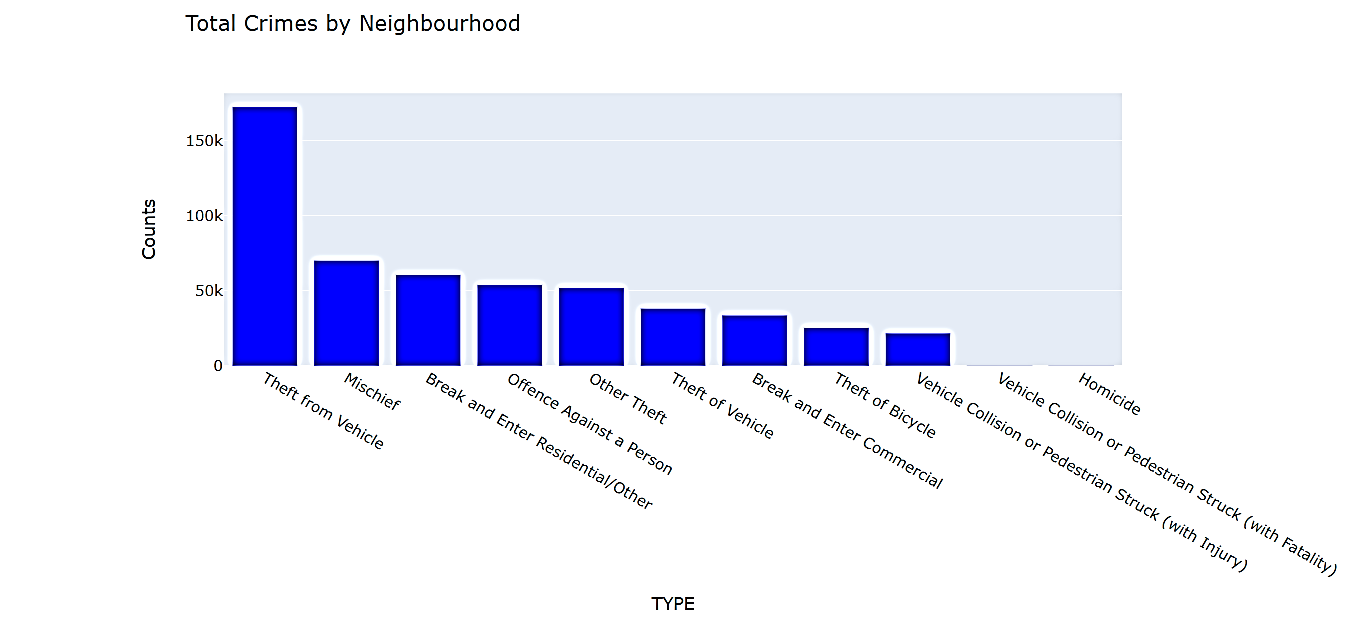


Figure 11.Count of crime types.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

The bar chart in Figure 12. shows the count of the crimes in the Neighbourhoods from 2003 to 2017. Highest count of crimes occurred in Central Business District at above 100,000 while the lowest count of crimes was in Musqueam.

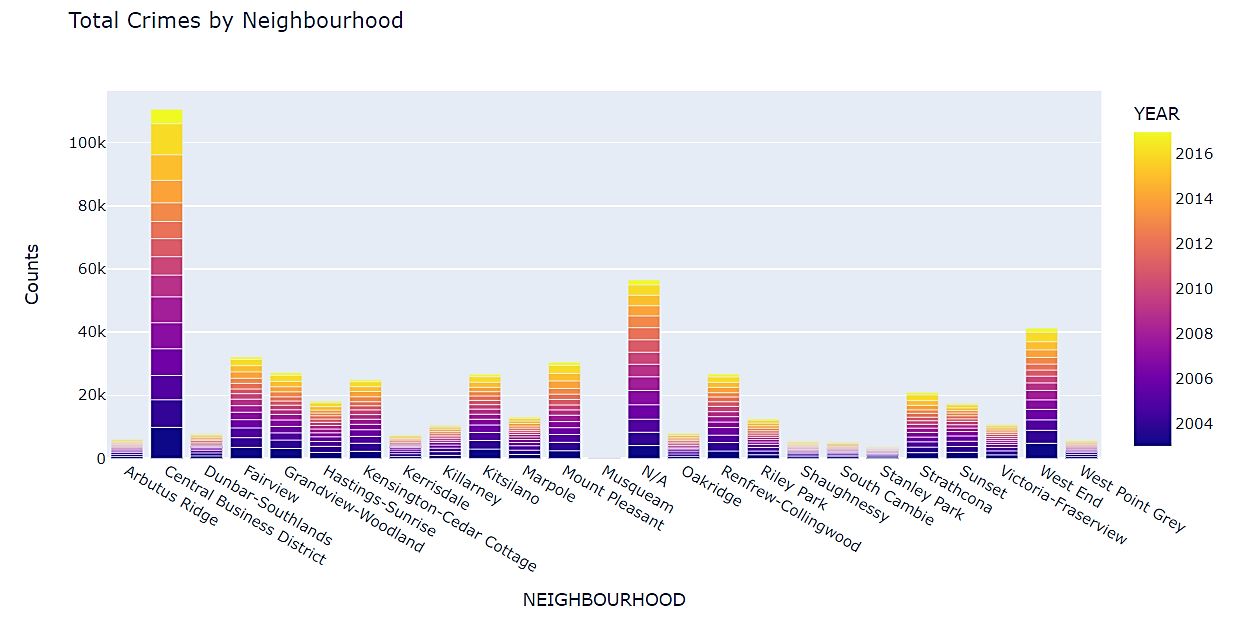


Figure 12.Total Crimes in Each Neighbourhood Per Year.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

## 3.1. Machine Learning Modelling

The dataset was split into test (30%) and train (70%) sets and standardized before modelling. As this analysis will utilise a target variable (‘Crime type’), some supervised and unsupervised learning models will be of primary focus. Logistic Regression (LR), Decision Tree (DT), K-Nearest Neighbour, Gaussian Naïve Bayes (GNB) and K-Means Clustering are some of the Machine Learning models utilised for prediction with Vancouver crime dataset. The prediction accuracies are in Table 6.

**PREDICTION**

Table 5.Machine learning models with prediction accuracy values

|  |  |
| --- | --- |
| **MACHINE LEARNING MODEL** | **PREDICTION ACCURACY (%)** |
| LOGISTIC REGRESSION | 72.76 |
| DECISION TREE | 32.79 |
| K-NEAREST NEIGHBOUR | 34.68 |
| GAUSSIAN NAÏVE BAYES | 28.00 |
| MULTILAYER PERCEPTRON NEURAL NETWORK | 42.89 |

**HYPERPARAMETER TUNING**

Improving the ML model performance can be achieved by tuning hyperparameters like: K value for KNN, random state for MLP, entropy for DT and using methods like Random Search, Grid Search. Table 7. provides the prediction accuracy of each ML model after tuning some of these hyperparameters.

Table 6.Machine learning models’ performance after hyperparameter tuning

|  |  |
| --- | --- |
| **MACHINE LEARNING MODEL** | **PREDICTION ACCURACY (%)** |
| LOGISTIC REGRESSION | 72.96 |
| DECISION TREE | 42.86 |
| K-NEAREST NEIGHBOUR | 40.85 |
| GAUSSIAN NAÏVE BAYES | 42.48 |
| MULTILAYER PERCEPTRON NEURAL NETWORK | 43.01 |

# Chapter 4- Result Evaluation and Discussion

Logistic Regression (LR) had an accuracy of 72.76% which was a good but with Cross Validation, this barely increased. The DT model which ran in two seconds, had a low prediction accuracy of 32.79%. After further feature engineering to improve the accuracy (by dropping some features like ‘Hundred Block’ and ‘Hour’), there was increase to 42.86%, but with a longer run time of nearly one minute. However, with attributes ‘gini’ and ‘entropy’, there was no significant change. KNN performed poorly with accuracy of 34.68% and with hyperparameter tuning, there was little increase to 40.85%. GNB had lowest accuracy at 28% but increased to 42.48% with hyperparameter tuning. MLP had a poor performance at 42.89% and after changing the values of the learning rate, accuracy dropped to about 41%. However, when cross validation was applied, it improved slightly to 43.01% but the run time was more than five minutes compared to the first technique which took about two minutes. With hyperparameter tuning, only GNB showed significant increase in performance of all the models applied. The DT model was also used to determine the five most important data features: 'Year', 'Month', 'Day', 'Hour' and 'Neighbourhood ID'.

Application of ML models to forecast crimes can be impactful in preventing crimes, specifically when some of these tested models, as used in this study, are incorporated into software or processes geared towards detecting crime patterns. Some of the dominant and emerging ML models which have been deployed in different crime prevention strategies by security agencies includes GNB, RF, KNN, DT, Logistic Regression, SVM, ANN, CNN. However, for this study, RF, DT, GNB and KNN have been employed to predict crimes, using the Vancouver crime dataset. The success of these models was found to be dependent on the modification of data features like ‘Location’ (Longitude and Latitude), ‘Time’, ‘Crime type id’, ‘Neighbourhood id’ and ‘Day’. Also, features like ‘Day’ and ‘Year’ were correlated to the target outcome: ‘Crime type’, as shown in data exploration. The in-depth data exploration helps in understanding the relationship between the different features. Knowledge like this, can be indicative of which features are important in further analysis for crime prediction in Vancouver and importantly, can be applicable to similar crime datasets in other regions. In several research studies, other kinds of data features have been shown to have a strong correlation with crime patterns and these include education, race, police per capita but these features were not found in the Vancouver crime data. Therefore, uniqueness of the crime data should be considered in relation to the features of the data collected. Determining features of the crime data which have the strongest influence on prediction, as done in this research, might be useful to crime analysts, researchers and those invested in community security. It can also be a useful aid in choosing the best ML models for application as some of them perform best when the strongest features are applied. Some other research work in relation to ML models for forecasting crimes, did not enunciate the importance of determining the most influential data features, hence the motivation for this research. For this study, some features, like ‘Month’, influenced the crime type as highlighted during data exploration. Also, civil unrest during sporting events like the Stanley Cup, caused an increase of specific crime types (especially ‘Mischief’). Despite the low performances of all the models used in this crime data analysis, the evaluation of their predictive accuracies of ML models shows LR was most effective. This study reinforced the importance of tuning hyperparameters in improving ML models’ performance, but it is important to note that this can be sometimes time-consuming and might not even improve performance as shown in this study. This was so for MLP and LR especially. But the accuracies of the models improved when hyperparameter tuning was combined with feature engineering (aided GNB the most). DT also performed better, but the remaining two models did not show much improvement. However, increasing the value of K, for KNN showed improvement as the accuracy got higher when value of K was 9. These results show that it is possible that some ML models will not have good accuracies for crime datasets with generalised features like the Vancouver crime dataset. One limitation is that important features which could be useful in forecasting crimes, such as social and economic factors (education level, income, unemployment rate), was absent in this dataset. The inclusion of this type of features in crime datasets will not only provide better understanding of the relationship between the data collected and the environment, but it can also improve the effectiveness of ML models in forecasting crimes. However, privacy concerns might be a limitation if these features were included. However, when features like location or time is removed and similar crime types or features are combined into one group, before modelling, the ML models will likely give good performance as crime predictors. Another limitation for this study was time constraint as other research avenues using more ML models, especially other Neural Network models and unsupervised models, could not be explored in the predictive analysis of this crime data. One unsupervised model applied was K-means clustering. It showed some relationship between the data features but was not as interpretable as the supervised ML models shown in the appendix (Appendix B).

Considering the crime dataset for this study is publicly available and was devoid of personal or identifying information of victims or perpetrators of the crimes, ethical and legal concerns related to privacy protection was not an issue. In addition, the credibility of the crime data is incontrovertible as it was collated and released by the Vancouver Police department. This is important in ensuring the prediction results were not tainted with bias or influenced by erroneous crime data. Avoiding issues of data privacy is one approach to reducing the public’s distrust of Predictive Policing using ML algorithms. Application of interpretable (white box) models like DT and LR, can be useful in forecasting crimes as the workings of the models are clear and showed the important features which determined the result of the prediction.

# Chapter 5-Conclusion

This study was conducted to assist analysts and researchers who work on crime prediction techniques, in acquiring a deeper understanding of this field, beneficial to development of efficient ML integrated technology devoid of issues like fairness, privacy and bias. For years, some Police forces have been deploying ML model-based software and techniques to analyse, forecast and prevent crimes, but concerns on ethics, racial discrimination, have not been helpful in the public’s acceptance of this technology. Therefore, there is a need for law enforcement and policy makers to carefully consider these issues with the aim to seek lasting solutions, some of which have been researched and proposed in this study. This academic work examined existing literature to explore the application of dominant ML models like DT, RF, GNB, Neural Networks, crime forecasting. Also, a review of the efficacy, strengths, weaknesses, and potential strategies for improving the prediction accuracies of these ML models was conducted. Some ML models were used to analyse the Vancouver crime dataset. There was an illustration of the relevance of key data features such as ‘Day’ and ‘Hour’ (representing Time), ‘Crime type’ and ‘Neighbourhood ID’ in crime prediction. These also influenced the capability of the models for crime forecasting. Feature engineering was important for each model to perform but while some had high performance without tuning hyperparameters (like LR model), others performed well only after this was combined (like DT model). However, there were issues like run time of the MLP model, which even after further feature engineering, did not improve. Another constraint was dropping features like ‘Location’ which could have been important to predicting crime spots, however this was not essential for this analysis. K-Means clustering was used to model the crime dataset, but was not as explainable as some of the supervised models in predicting crimes.

Decision Tree produced explainable and interpretable results as seen in the visualisation and evaluation with the metrices, included in the appendix (Appendix A). But one limitation lies with its poor ability to generalise data. Hence, further study on the improvement and integration of DT into developing crime predicting software can be worthwhile for future study. However, Neural Networks like MLP, can also be beneficial for crime prediction as they handle large data successfully. But if the interpretability is improved, which due to time constraint could not be performed in this study, it could serve as an effective tool in forecasting crimes.

The successful implementation of carefully crafted system which addresses these issues, while sustaining the use of ML technologies which have worked well in crime forecasting by trained law enforcement, can be helpful to the field of crime prevention. With appropriate and monitored usage, an efficient ML model framework, is expected to significantly curtail crime rate. Data science has been advancing and new developments in ML can positively impact the future of crime prevention, provided more opportunities for innovation is encouraged by analysts, policy makers and law enforcement agencies.

## 5.1 Future Research

Crime forecasting methods which can be useful to security agencies and crime analysts have been thoroughly investigated in this study. The overall impact in the deployment of effective crime prevention techniques with ML models can foster economic development in the society. Therefore, future research for the creation of a well-tested system incorporated with the ML models which have successfully predicted crimes as shown in this project, will be of great benefit. However, as the field of machine learning for forecasting crimes continues to evolve, for the future, crime analysts can combine interpretable, clear guidelines and best practices for deployment of these models which will ensure good ethical standards are maintained. Specifically, the development of universal laws applicable to the use of ML models for crime predictions, with the involvement of data analysts, civil liberty organisations and community leaders. This could be beneficial in mitigating ethical and bias issues plaguing the current use of ML model-integrated software by law enforcement in forecasting crimes. The development of ML models, with fast computational times, that can be fed live data to predict crime trends, can potentially aid quicker and more effective response by law enforcement agencies to emerging criminal threats. An ensemble of multiple data sources like historical crime data, weather reports and live social media feeds can be integrated into crime prevention strategies in the future. Also, due to time constraint, additional ML models could have been utilised to analyse the Vancouver crime dataset and other crime datasets might have been analysed. This would have provided a deeper understanding of the predictive capabilities inherent in machine learning techniques.

# References

Abdul Jalil, M.@ M., Mohd, F. and Mohamad Noor, N.M. (2017) ‘A Comparative Study to Evaluate Filtering Methods for Crime Data Feature Selection’, *Procedia Computer Science*, 116, pp. 113–120. Available at: https://doi.org/10.1016/j.procs.2017.10.018.

Abdulrahman, N. and Abedalkhader, W. (2017) ‘KNN Classifier and Naive Bayse Classifier for Crime Prediction in San Francisco Context’, *International Journal of Database Management Systems*, 9(4), pp. 1–9. Available at: https://doi.org/10.5121/ijdms.2017.9401.

Albahli, S. *et al.* (2020) ‘Predicting the Type of Crime: Intelligence Gathering and Crime Analysis’, *Cmc -Tech Science Press-*, 66, pp. 2318–2341. Available at: https://doi.org/10.32604/cmc.2021.014113.

Ali, S., Alvi, S.A. and Rehman, A.U. (2019) ‘The Usual Suspects: Machine Learning Based Predictive Policing for Criminal Identification’, in *2019 13th International Conference on Open-Source Systems and Technologies (ICOSST)*. *2019 13th International Conference on Open-Source Systems and Technologies (ICOSST)*, pp. 1–6. Available at: https://doi.org/10.1109/ICOSST48232.2019.9043925.

Alonso, J.M., Andrews, R. and Jorda, V. (2019) ‘Do neighbourhood renewal programs reduce crime rates? Evidence from England’, *Journal of Urban Economics*, 110, pp. 51–69. Available at: https://doi.org/10.1016/j.jue.2019.01.004.

Alparslan, Y. *et al.* (2020) ‘Perfecting the Crime Machine’. arXiv. Available at: http://arxiv.org/abs/2001.09764 (Accessed: 21 March 2023).

Amusa, L.B., Bengesai, A.V. and Khan, H.T.A. (2022) ‘Predicting the Vulnerability of Women to Intimate Partner Violence in South Africa: Evidence from Tree-based Machine Learning Techniques’, *Journal of Interpersonal Violence*, 37(7–8), pp. NP5228–NP5245. Available at: https://doi.org/10.1177/0886260520960110.

Angelov, P.P. *et al.* (2021) ‘Explainable artificial intelligence: an analytical review’, *WIREs Data Mining and Knowledge Discovery*, 11(5), p. e1424. Available at: https://doi.org/10.1002/widm.1424.

Asor, J.R. (2020) ‘Implementation of Predictive Crime Analytics in Municipal Crime Management System in Calauan, Laguna, Philippines’, *International Journal of Advanced Trends in Computer Science and Engineering*, 9(1.3), pp. 150–157. Available at: https://doi.org/10.30534/ijatcse/2020/2291.32020.

Awal, Md.A. *et al.* (2016) ‘Using linear regression to forecast future trends in crime of Bangladesh’, in *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*. *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pp. 333–338. Available at: https://doi.org/10.1109/ICIEV.2016.7760021.

Babuta, A. and Oswald, M. (2020) *Data analytics and algorithms in policing in England and Wales: Towards a new policy framework*. (RUSI Occasional Paper).

Balu, V., Navya Sri, T. and Bupathi, M.A. (2022) ‘Crime Prediction and Analysis Using Machine Learning’, *International Journal of Computer Science and Mobile Computing*, 11(3), pp. 95–101. Available at: https://doi.org/10.47760/ijcsmc.2022.v11i03.011.

Birks, D., Townsley, M. and Hart, T. (2023) *Predictive policing in an Australian context: assessing viability and utility*. Australian Institute of Criminology. Available at: https://doi.org/10.52922/ti78870.

Biswas, A.A. and Basak, S. (2019) ‘Forecasting the Trends and Patterns of Crime in Bangladesh using Machine Learning Model’, in *2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*. *2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*, pp. 114–118. Available at: https://doi.org/10.1109/ICCT46177.2019.8969031.

Briz-Redón, Á., Mateu, J. and Montes, F. (2022) ‘Modeling the Influence of Places on Crime Risk Through a Non-Linear Effects Model: a Comparison with Risk Terrain Modeling’, *Applied Spatial Analysis and Policy*, 15(2), pp. 507–527. Available at: https://doi.org/10.1007/s12061-021-09410-6.

Burt, C.J. *et al.* (2022) ‘Community Greening, Fear of Crime, and Mental Health Outcomes’, *American Journal of Community Psychology*, 69(1–2), pp. 46–58. Available at: https://doi.org/10.1002/ajcp.12544.

Canhoto, A.I. (2021) ‘Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective’, *Journal of Business Research*, 131, pp. 441–452. Available at: https://doi.org/10.1016/j.jbusres.2020.10.012.

Caplan, J.M. *et al.* (2021) ‘Data-Informed and Place-Based Violent Crime Prevention: The Kansas City, Missouri Risk-Based Policing Initiative’, *Police Quarterly*, 24(4), pp. 438–464. Available at: https://doi.org/10.1177/10986111211003205.

Castets-Renard, C. (2021) ‘Human Rights and Algorithmic Impact Assessment for Predictive Policing’. Rochester, NY. Available at: https://papers.ssrn.com/abstract=3890283 (Accessed: 25 March 2023).

Castro-Toledo, F.J., Miró-Llinares, F. and Aguerri, J.C. (2023) ‘Data-Driven Criminal Justice in the age of algorithms: epistemic challenges and practical implications’, *Criminal Law Forum* [Preprint]. Available at: https://doi.org/10.1007/s10609-023-09454-y.

Ch, R. *et al.* (2020) ‘Computational System to Classify Cyber Crime Offenses using Machine Learning’, *Sustainability*, 12(10), p. 4087. Available at: https://doi.org/10.3390/su12104087.

Chen, Z. *et al.* (2018) ‘Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review’, *Knowledge and Information Systems*, 57(2), pp. 245–285. Available at: https://doi.org/10.1007/s10115-017-1144-z.

Chun, S.A. *et al.* (2019) ‘Crime Prediction Model using Deep Neural Networks’, in *Proceedings of the 20th Annual International Conference on Digital Government Research*. *dg.o 2019: 20th Annual International Conference on Digital Government Research*, Dubai United Arab Emirates: ACM, pp. 512–514. Available at: https://doi.org/10.1145/3325112.3328221.

Cichosz, P. (2020) ‘Urban Crime Risk Prediction Using Point of Interest Data’, *ISPRS International Journal of Geo-Information*, 9(7), p. 459. Available at: https://doi.org/10.3390/ijgi9070459.

Corsaro, N. and Engel, R.S. (2020) ‘Community-Based Crime Reduction in Tulsa: An Application of Petersilia’s Multi-Stakeholder Collaborative Approach to Improve Policy Relevance and Reduce Crime Risk’, *Justice Quarterly*, 37(7), pp. 1199–1220. Available at: https://doi.org/10.1080/07418825.2020.1825775.

Davis, J. *et al.* (2022) ‘Five ethical challenges facing data-driven policing’, *AI and Ethics*, 2(1), pp. 185–198. Available at: https://doi.org/10.1007/s43681-021-00105-9.

Dhankhad, S., Mohammed, E. and Far, B. (2018) ‘Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study’, in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*. *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, pp. 122–125. Available at: https://doi.org/10.1109/IRI.2018.00025.

Drawve, G., Moak, S.C. and Berthelot, E.R. (2016) ‘Predictability of gun crimes: a comparison of hot spot and risk terrain modelling techniques’, *Policing and Society*, 26(3), pp. 312–331. Available at: https://doi.org/10.1080/10439463.2014.942851.

Egbert, S. and Leese, M. (2021) *Criminal Futures: Predictive Policing and Everyday Police Work*. Taylor & Francis. Available at: https://doi.org/10.4324/9780429328732.

Fan, Y. (2021) ‘Criminal psychology trend prediction based on deep learning algorithm and three-dimensional convolutional neural network’, *Journal of Psychology in Africa*, 31(3), pp. 292–297. Available at: https://doi.org/10.1080/14330237.2021.1927317.

Farjami, Y. and Abdi, K. (2021) ‘A genetic-fuzzy algorithm for spatio-temporal crime prediction’, *Journal of Ambient Intelligence and Humanized Computing* [Preprint]. Available at: https://doi.org/10.1007/s12652-020-02858-3.

Fredj, O.B. (2022) ‘A NLP-inspired method to predict multi-step cyberattacks’, in *2022 15th International Conference on Security of Information and Networks (SIN)*. *2022 15th International Conference on Security of Information and Networks (SIN)*, pp. 1–6. Available at: https://doi.org/10.1109/SIN56466.2022.9970551.

Gelpi, C. and Avdan, N. (2018) ‘Democracies at risk? A forecasting analysis of regime type and the risk of terrorist attack’, *Conflict Management and Peace Science*, 35(1), pp. 18–42. Available at: https://doi.org/10.1177/0738894215608998.

Ghappour, A. (2015) ‘Machine Generated Culpability: Socio-Legal Agency in Machine Learning for Cybersecurity Enforcement’, *Available at SSRN 2587340* [Preprint].

Ghasemi, M. *et al.* (2021) ‘The Application of Machine Learning to a General Risk–Need Assessment Instrument in the Prediction of Criminal Recidivism’, *Criminal Justice and Behavior*, 48(4), pp. 518–538. Available at: https://doi.org/10.1177/0093854820969753.

Ghazvini, A. *et al.* (2020) ‘Crime Spatiotemporal Prediction With Fused Objective Function in Time Delay Neural Network’, *IEEE Access*, 8, pp. 115167–115183. Available at: https://doi.org/10.1109/ACCESS.2020.3002766.

González-Prieto, Á. *et al.* (2021) ‘Machine learning for risk assessment in gender-based crime’. arXiv. Available at: https://doi.org/10.48550/arXiv.2106.11847.

Gowtham, S. (2018) ‘Crime Data Analysis Using Data Mining Techniques to Improve Crimes Prevention’, *Software Engineering and Technology* [Preprint]. Available at: https://www.semanticscholar.org/paper/Crime-Data-Analysis-Using-Data-Mining-Techniques-to-Gowtham/f78d65eb6529b8f40bfcee6acf95e7c4206a8d71 (Accessed: 15 April 2023).

Grogger, J.T. *et al.* (2021) ‘Comparing Conventional and Machine-Learning Approaches to Risk Assessment in Domestic Abuse Cases’, *SSRN Electronic Journal* [Preprint]. Available at: https://doi.org/10.2139/ssrn.3760094.

Hälterlein, J. (2021) ‘Epistemologies of predictive policing: Mathematical social science, social physics and machine learning’, *Big Data & Society*, 8(1), p. 205395172110031. Available at: https://doi.org/10.1177/20539517211003118.

Hardyns, W. and Rummens, A. (2018) ‘Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges’, *European Journal on Criminal Policy and Research*, 24(3), pp. 201–218. Available at: https://doi.org/10.1007/s10610-017-9361-2.

He, L. *et al.* (2020) ‘Ambient Population and Larceny-Theft: A Spatial Analysis Using Mobile Phone Data’, *ISPRS International Journal of Geo-Information*, 9(6), p. 342. Available at: https://doi.org/10.3390/ijgi9060342.

Helm, P. and Hagendorff, T. (2021) ‘Beyond the Prediction Paradigm: Challenges for AI in the Struggle Against Organized Crime’, *LAW AND CONTEMPORARY PROBLEMS*, 84(3).

Hossain, S. *et al.* (2020) ‘Crime Prediction Using Spatio-Temporal Data’, in N. Chaubey, S. Parikh, and K. Amin (eds) *Computing Science, Communication and Security*. Singapore: Springer (Communications in Computer and Information Science), pp. 277–289. Available at: https://doi.org/10.1007/978-981-15-6648-6\_22.

Ingilevich, V. and Ivanov, S. (2018) ‘Crime rate prediction in the urban environment using social factors’, *Procedia Computer Science*, 136, pp. 472–478. Available at: https://doi.org/10.1016/j.procs.2018.08.261.

Jenga, K., Catal, C. and Kar, G. (2023) ‘Machine learning in crime prediction’, *Journal of Ambient Intelligence and Humanized Computing*, 14(3), pp. 2887–2913. Available at: https://doi.org/10.1007/s12652-023-04530-y.

Jin, G. *et al.* (2020) ‘Addressing Crime Situation Forecasting Task with Temporal Graph Convolutional Neural Network Approach’, in *2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*. *2020 12th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, pp. 474–478. Available at: https://doi.org/10.1109/ICMTMA50254.2020.00108.

Kang, H.-W. and Kang, H.-B. (2017) ‘Prediction of crime occurrence from multi-modal data using deep learning’, *PLOS ONE*, 12(4), p. e0176244. Available at: https://doi.org/10.1371/journal.pone.0176244.

Kaufmann, M., Egbert, S. and Leese, M. (2019) ‘Predictive Policing and the Politics of Patterns’, *The British Journal of Criminology*, 59(3), pp. 674–692. Available at: https://doi.org/10.1093/bjc/azy060.

Kaur, D. *et al.* (2020) ‘Trust-Based Human-Machine Collaboration Mechanism for Predicting Crimes’, in L. Barolli et al. (eds) *Advanced Information Networking and Applications*. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing), pp. 603–616. Available at: https://doi.org/10.1007/978-3-030-44041-1\_54.

Kemp, S. *et al.* (2021) ‘Empty Streets, Busy Internet: A Time-Series Analysis of Cybercrime and Fraud Trends During COVID-19’, *Journal of Contemporary Criminal Justice*, 37(4), pp. 480–501. Available at: https://doi.org/10.1177/10439862211027986.

KIM, K.-S. and JEONG, Y.-H. (2021) ‘A Study on Crime Prediction to Reduce Crime Rate Based on Artificial Intelligence’, *Korea Journal of Artificial Intelligence*, 9(1), pp. 15–20. Available at: https://doi.org/10.24225/KJAI.2021.9.1.15.

Kounadi, O. *et al.* (2020) ‘A systematic review on spatial crime forecasting’, *Crime Science*, 9(1), p. 7. Available at: https://doi.org/10.1186/s40163-020-00116-7.

Krieg, S.J. *et al.* (2022) ‘Predicting terrorist attacks in the United States using localized news data’, *PLOS ONE*. Edited by S. V E, 17(6), p. e0270681. Available at: https://doi.org/10.1371/journal.pone.0270681.

Kshatri, S.S. *et al.* (2021) ‘An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach’, *IEEE Access*, 9, pp. 67488–67500. Available at: https://doi.org/10.1109/ACCESS.2021.3075140.

Kumar, R. and Nagpal, B. (2019) ‘Analysis and prediction of crime patterns using big data’, *International Journal of Information Technology*, 11(4), pp. 799–805. Available at: https://doi.org/10.1007/s41870-018-0260-7.

Lamari, Y. *et al.* (2020) ‘Predicting Spatial Crime Occurrences through an Efficient Ensemble-Learning Model’, *ISPRS International Journal of Geo-Information*, 9(11), p. 645. Available at: https://doi.org/10.3390/ijgi9110645.

Lee, Y. and Park, J. (2022) ‘Using Big Data to Prevent Crime: Legitimacy Matters’, *Asian Journal of Criminology*, 17(1), pp. 61–80. Available at: https://doi.org/10.1007/s11417-021-09353-4.

Lim, M., Abdullah, A. and Jhanjhi, N. (2021) ‘Performance optimization of criminal network hidden link prediction model with deep reinforcement learning’, *Journal of King Saud University - Computer and Information Sciences*, 33(10), pp. 1202–1210. Available at: https://doi.org/10.1016/j.jksuci.2019.07.010.

Lin, Y.-L., Chen, T.-Y. and Yu, L.-C. (2017) ‘Using Machine Learning to Assist Crime Prevention’, in *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pp. 1029–1030. Available at: https://doi.org/10.1109/IIAI-AAI.2017.46.

Liu, M. and Lu, T. (2019) ‘A Hybrid Model of Crime Prediction’, *Journal of Physics: Conference Series*, 1168(3), p. 032031. Available at: https://doi.org/10.1088/1742-6596/1168/3/032031.

Lloyd, C.D. *et al.* (2020) ‘Reassessment improves prediction of criminal recidivism: A prospective study of 3,421 individuals in New Zealand’, *Psychological Assessment*, 32, pp. 568–581. Available at: https://doi.org/10.1037/pas0000813.

Meijer, A. and Wessels, M. (2019) ‘Predictive Policing: Review of Benefits and Drawbacks’, *International Journal of Public Administration*, 42(12), pp. 1031–1039. Available at: https://doi.org/10.1080/01900692.2019.1575664.

Montasari, R. (2023) ‘The Application of Big Data Predictive Analytics and Surveillance Technologies in the Field of Policing’, in R. Montasari (ed.) *Countering Cyberterrorism: The Confluence of Artificial Intelligence, Cyber Forensics and Digital Policing in US and UK National Cybersecurity*. Cham: Springer International Publishing (Advances in Information Security), pp. 81–114. Available at: https://doi.org/10.1007/978-3-031-21920-7\_5.

Mugari, I. and Obioha, E.E. (2021) ‘Predictive Policing and Crime Control in The United States of America and Europe: Trends in a Decade of Research and the Future of Predictive Policing’, *Social Sciences*, 10(6), p. 234. Available at: https://doi.org/10.3390/socsci10060234.

Niu, X. *et al.* (2019) ‘Dynamics of crime activities in the network of city community areas’, *Applied Network Science*, 4(1), p. 127. Available at: https://doi.org/10.1007/s41109-019-0239-8.

Olabanjo, O.A. *et al.* (2021) ‘An ensemble machine learning model for the prediction of danger zones: Towards a global counter-terrorism’, *Soft Computing Letters*, 3, p. 100020. Available at: https://doi.org/10.1016/j.socl.2021.100020.

Palanivinayagam, A. *et al.* (2021) ‘An Optimized Machine Learning and Big Data Approach to Crime Detection’, *Wireless Communications and Mobile Computing*, 2021, p. e5291528. Available at: https://doi.org/10.1155/2021/5291528.

Patton, D.U. *et al.* (2022) ‘Community implications for gun violence prevention during co-occurring pandemics; a qualitative and computational analysis study’, *Preventive Medicine*, 165, p. 107263. Available at: https://doi.org/10.1016/j.ypmed.2022.107263.

Pavan Naidu Kavala, S.A. *et al.* (2022) ‘Crime analysis and Prediction using Neural Networks’, in *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*. *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pp. 2047–2050. Available at: https://doi.org/10.1109/ICACCS54159.2022.9785081.

Pelham, W.E., Petras, H. and Pardini, D.A. (2020) ‘Can Machine Learning Improve Screening for Targeted Delinquency Prevention Programs?’, *Prevention Science*, 21(2), pp. 158–170. Available at: https://doi.org/10.1007/s11121-019-01040-2.

Rahman, Md.A. and Islam, A.B.M.A.A. (2022) ‘Cri-Astrologer: Predicting Demography of Involved Criminals based on Historical Data’, in *2022 9th International Conference on Networking, Systems and Security*. *9th NSysS 2022: 2022 9th International Conference on Networking, Systems and Security*, Cox’s Bazar Bangladesh: ACM, pp. 92–102. Available at: https://doi.org/10.1145/3569551.3569561.

Rumi, S.K., Deng, K. and Salim, F.D. (2018) ‘Crime event prediction with dynamic features’, *EPJ Data Science*, 7(1), p. 43. Available at: https://doi.org/10.1140/epjds/s13688-018-0171-7.

Sheehey, B. (2019) ‘Algorithmic paranoia: the temporal governmentality of predictive policing’, *Ethics and Information Technology*, 21(1), pp. 49–58. Available at: https://doi.org/10.1007/s10676-018-9489-x.

Shinde, P. *et al.* (2022) ‘Study on Crime Examination and Forecasting using Machine Learning’, *Journal of Pharmaceutical Negative Results*, pp. 7586–7594. Available at: https://doi.org/10.47750/pnr.2022.13.S07.915.

Singh, K., Chaudhary, A.S. and Kaur, P. (2019) ‘A Machine Learning Approach for Enhancing Defence Against Global Terrorism’, in *2019 Twelfth International Conference on Contemporary Computing (IC3)*. *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pp. 1–5. Available at: https://doi.org/10.1109/IC3.2019.8844947.

Sobh, T.S. (2020) ‘An Intelligent and Secure Framework for Anti-Money Laundering’, *Journal of Applied Security Research*, 15(4), pp. 517–546. Available at: https://doi.org/10.1080/19361610.2020.1812994.

Stalidis, P., Semertzidis, T. and Daras, P. (2021) ‘Examining Deep Learning Architectures for Crime Classification and Prediction’, *Forecasting*, 3(4), pp. 741–762. Available at: https://doi.org/10.3390/forecast3040046.

Stec, A. and Klabjan, D. (2018) ‘Forecasting Crime with Deep Learning’. arXiv. Available at: http://arxiv.org/abs/1806.01486 (Accessed: 20 March 2023).

Suguna, M.R., Jeyavathana, R.B. and Kanimozhi, K.V. (2022) ‘Comparative analysis of crime predictions using machine learning algorithms with geospatial features’, *AIP Conference Proceedings*, 2393(1), p. 020004. Available at: https://doi.org/10.1063/5.0074141.

Teixeira, M. *et al.* (2018) ‘SCADA System Testbed for Cybersecurity Research Using Machine Learning Approach’, *Future Internet*, 10(8), p. 76. Available at: https://doi.org/10.3390/fi10080076.

ToppiReddy, H.K.R., Saini, B. and Mahajan, G. (2018) ‘Crime Prediction & Monitoring Framework Based on Spatial Analysis’, *Procedia Computer Science*, 132, pp. 696–705. Available at: https://doi.org/10.1016/j.procs.2018.05.075.

Toure, I. and Gangopadhyay, A. (2016) ‘Real time big data analytics for predicting terrorist incidents’, in *2016 IEEE Symposium on Technologies for Homeland Security (HST)*. *2016 IEEE Symposium on Technologies for Homeland Security (HST)*, Waltham, MA, USA: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/THS.2016.7568906.

Towers, S. *et al.* (2018) ‘Factors influencing temporal patterns in crime in a large American city: A predictive analytics perspective’, *PLOS ONE*. Edited by H. Eisenbarth, 13(10), p. e0205151. Available at: https://doi.org/10.1371/journal.pone.0205151.

Travaini, G.V. *et al.* (2022) ‘Machine Learning and Criminal Justice: A Systematic Review of Advanced Methodology for Recidivism Risk Prediction’, *International Journal of Environmental Research and Public Health*, 19(17), p. 10594. Available at: https://doi.org/10.3390/ijerph191710594.

Uzlov, D. *et al.* (2020) ‘Adaptive Matrix Model for a Crime Forecasting Task’, in *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*. *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, pp. 96–101. Available at: https://doi.org/10.1109/DSMP47368.2020.9204127.

Wang, C. *et al.* (2022) ‘In Pursuit of Interpretable, Fair and Accurate Machine Learning for Criminal Recidivism Prediction’, *Journal of Quantitative Criminology* [Preprint]. Available at: https://doi.org/10.1007/s10940-022-09545-w.

Wang, Z. *et al.* (2019) ‘Crime Geographical Displacement: Testing Its Potential Contribution to Crime Prediction’, *ISPRS International Journal of Geo-Information*, 8(9), p. 383. Available at: https://doi.org/10.3390/ijgi8090383.

Wheeler, A.P. and Steenbeek, W. (2021) ‘Mapping the Risk Terrain for Crime Using Machine Learning’, *Journal of Quantitative Criminology*, 37(2), pp. 445–480. Available at: https://doi.org/10.1007/s10940-020-09457-7.

William, P. *et al.* (2023) ‘Crime Analysis Using Computer Vision Approach with Machine Learning’, in N. Marriwala et al. (eds) *Mobile Radio Communications and 5G Networks*. Singapore: Springer Nature (Lecture Notes in Networks and Systems), pp. 297–315. Available at: https://doi.org/10.1007/978-981-19-7982-8\_25.

Yadav, R. and Kumari Sheoran, S. (2018) ‘Modified ARIMA Model for Improving Certainty in Spatio-Temporal Crime Event Prediction’, *2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE)*, pp. 1–4. Available at: https://doi.org/10.1109/ICRAIE.2018.8710398.

Yang, B. *et al.* (2020) ‘A spatio-temporal method for crime prediction using historical crime data and transitional zones identified from nightlight imagery’, *International Journal of Geographical Information Science*, 34(9), pp. 1740–1764. Available at: https://doi.org/10.1080/13658816.2020.1737701.

Ye, X., Duan, L. and Peng, Q. (2021) ‘Spatiotemporal Prediction of Theft Risk with Deep Inception-Residual Networks’, *Smart Cities*, 4(1), pp. 204–216. Available at: https://doi.org/10.3390/smartcities4010013.

Yoo, Y. and Wheeler, A.P. (2019) ‘Using risk terrain modeling to predict homeless related crime in Los Angeles, California’, *Applied Geography*, 109, p. 102039. Available at: https://doi.org/10.1016/j.apgeog.2019.102039.

Young, M.M., Bullock, J.B. and Lecy, J.D. (2019) ‘Artificial Discretion as a Tool of Governance: A Framework for Understanding the Impact of Artificial Intelligence on Public Administration’, *Perspectives on Public Management and Governance*, 2(4), pp. 301–313. Available at: https://doi.org/10.1093/ppmgov/gvz014.

Yu, C.-H. *et al.* (2011) ‘Crime Forecasting Using Data Mining Techniques’, in *2011 IEEE 11th International Conference on Data Mining Workshops*. *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 779–786. Available at: https://doi.org/10.1109/ICDMW.2011.56.

Zeng, J., Ustun, B. and Rudin, C. (2017) ‘Interpretable classification models for recidivism prediction’, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 180(3), pp. 689–722.

Zhang, X. *et al.* (2020) ‘Comparison of Machine Learning Algorithms for Predicting Crime Hotspots’, *IEEE Access*, 8, pp. 181302–181310. Available at: https://doi.org/10.1109/ACCESS.2020.3028420.

Zhang, X. *et al.* (2022) ‘Interpretable machine learning models for crime prediction’, *Computers, Environment and Urban Systems*, 94, p. 101789. Available at: https://doi.org/10.1016/j.compenvurbsys.2022.101789.

Zhou, X. *et al.* (2021) ‘Mixed Spatio-Temporal Neural Networks on Real-time Prediction of Crimes’, in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1749–1754. Available at: https://doi.org/10.1109/ICMLA52953.2021.00277.

Zhuang, Y. *et al.* (2017) ‘Crime Hot Spot Forecasting: A Recurrent Model with Spatial and Temporal Information’, in *2017 IEEE International Conference on Big Knowledge (ICBK)*. *2017 IEEE International Conference on Big Knowledge (ICBK)*, pp. 143–150. Available at: https://doi.org/10.1109/ICBK.2017.3.

Zilka, M., Sargeant, H. and Weller, A. (2022) ‘Transparency, Governance and Regulation of Algorithmic Tools Deployed in the Criminal Justice System: a UK Case Study’, in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 880–889. Available at: https://doi.org/10.1145/3514094.3534200.

# Appendix A

## A.1. Decision Tree Modelling

Fitting the DT model to the Vancouver crime dataset, using criteria: Entropy and Gini Index as seen in Figure 13. and Figure 14.

**A picture containing text, diagram, line, parallel

Description automatically generated**

Figure 13. Decision tree modelling of the Vancouver crime dataset with Entropy.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

**A picture containing text, diagram, line, plot

Description automatically generated**

Figure 14.Decision tree modelling of the Vancouver crime dataset with Gini Index.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

## A.2. Evaluation Metrics for Decision Tree

Figure 12. shows the parent and children nodes of the Decision Tree model, with varying values of entropy and samples. This is illustration of the interpretability of this ML model. Measuring the quality of a DT’s node split can be mathematically calculated with criteria Entropy and Gini.

## A.3. Entropy

Entropy is the metric used for measuring the level of disorder or impurity with the aim to reduce the level of uncertainty. Analysing the Vancouver crime dataset, the accuracy of prediction was 31.71% but this increased to 42.80% with entropy criterion. The value of Entropy lies between 0 and the Log of the number of classes. It is sensitive to the number of classes. The formula is below in Equation 1.:

=

Equation . Mathematical Equation for Entropy

## A.4. Gini index

This evaluation metric which is for Information Gain, calculates the probability of a certain feature or variable being classified erroneously. For the Vancouver crime dataset, the accuracy of prediction improved to 42.85% with Gini index criterion.

Equation . Mathematical Equation for Gini Index

In Equation 2. above, value of Gini Index value is between 0 and 1. Purity of the DT classification is 0 and 1 is for random distribution in the classes. These formulae can be useful in further analysis using DT model for crime forecasts.

Decision tree modelling of the Vancouver dataset, using the above metrices performed better with further feature engineering which included selecting certain features related to time (like ‘Year’) and the neighbourhood, for prediction. With Gini index, accuracy was 42.86% and with Entropy it was 42.99%. However, the precision score and recall score were low at 33.08% and 32.79% respectively which shows this model had a poor performance in predicting crime accurately. However, Table A. and Figure II. shows the degree of importance of the most impactful data features of the crime dataset when modelled with Decision Tree, this was before data transformation which combined four of the time features into ‘Date’:

**Table 7.Feature of importance with score**

|  |
| --- |
| **INDEX** **FEATURE OF IMPORTANCE SCORE** |
| 0 Year 0.113052  1 Month 0.201141  2 Day 0.355761  3 Hour 0.280834  4 Neighbourhood ID 0.049211 |

A picture containing text, screenshot, diagram, number

Description automatically generated

Figure 15.Bar chart of feature of importance with score.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

The most important feature for crime prediction using DT model is ‘Day’, followed by ‘Hour’. The least important feature is ‘Neighbourhood ID’. The details of the period the crimes occurred is essential for crime forecast as it can be instrumental in developing policing strategies for preventing future occurrence. Table 8. is a Classification Report with values of Precision, Recall and F1-Score which evaluates the DT model’s performance in classifying data features. These results show how this model is one of the most explainable for crime forecast analysis as the way it works as a classifier can be understood.

**Table 8. Classification report for Decision Tree**

|  |
| --- |
| **DATA FEATURES PRECISION RECALL F1-SCORE SUPPORT** |
| Type 0.15 0.16 0.15 2512 |
| Month 0.20 0.21 0.20 4484 |
| Year 0.04 0.06 0.05 18 |
| Date 0.18 0.19 0.18 5317 |
| Neighbourhood 1.00 0.99 1.00 4025 |
| Day 0.21 0.22 0.22 3907 |
| Hour 0.41 0.39 0.40 12992 |
| Hundred block 0.10 0.11 0.10 1898 |
| Neighbourhood 0.12 0.12 0.12 2895 |
| Day of week 0.00 0.00 0.00 16 |
| Category 0.06 0.06 0.06 1642 |
|  |
| Accuracy 0.32 39706 |
| Macro average 0.22 0.23 0.23 39706 |
| Weighted average 0.33 0.32 0.33 39706 |

## A.5. Logistic Regression Modelling

The degree of importance of the most valuable data features for crime prediction of the Vancouver crime dataset when modelled with Logistic Regression can be seen in the Table B. and Figure 16.:

**Table 9.Feature of importance with score**

|  |
| --- |
| **INDEX FEATURE OF IMPORTANCE COEFFICIENT** |
| 0 Year 0.00922 |
| 1 Month -0.01539 |
| 2 Day -0.03406 |
| 3 Hour -2.59680 |
| 4 Neighbourhood ID -0.11227 |

A picture containing text, screenshot, rectangle, display

Description automatically generated

Figure 16.Bar chart of feature of importance with score.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

The most important score (or coefficient) for crime prediction using LR model is for ‘Hour’, which means it contributes the most information to the model. This is followed by ‘Neighbourhood ID’. The least important variable is ‘Year’. The details of the time crimes were committed is essential in crime forecast and can be useful in developing crime prevention strategies. This analysis shows how LR is an interpretable ML model for crime prediction.

# APPENDIX B

## B.1. K-MEANS CLUSTERING

Grouping data points based on their similarities can be useful in identifying patterns in the Vancouver crime dataset. This is possible using K-Means Clustering which is an unsupervised ML model. This modelling involved feature engineering by the selection of relevant features: ‘Hour’, ‘Year’, ‘Month’, ‘Day’, ‘Neighbourhood ID’, ‘Crime Type ID’ and scaling the dataset to ensure the features have equal representation in their contributions.

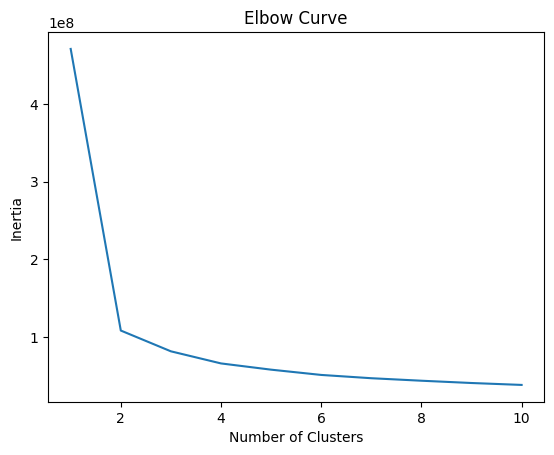


Figure 17. Elbow curve plot using K-Means Clustering

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

The optimal number of clusters was determined with the elbow method which gives this as 2, seen in Figure 17. This was done by the within-cluster sum of squares (WCSS) method. Below is Figure 18. which is a cluster scatter plot and Figure 19., a heat map, with two of the important data features, crime type ID and neighbourhood ID:

A picture containing text, screenshot, line, rectangle

Description automatically generated

Figure 18. Plot representing the distinct clusters.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

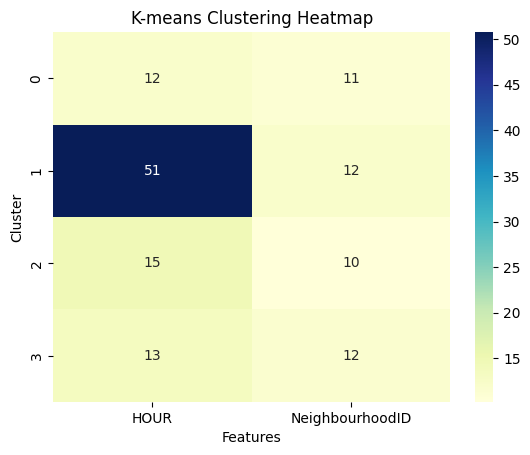
**

Figure 19. K- Means Heatmap.

*Source: The Author. Adapted from Vancouver crime dataset (2003-2017).*

The ‘Crime Type ID’ has a relationship with the ‘Neighbourhood ID’ as seen in the plot above. This is relatable to the data exploration which illustrated how certain Neighbourhoods such as the CBD, experienced certain crimes, at specific times, which can be useful knowledge for law enforcement agencies in preventing crimes with similar data features.