iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44577

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

Deep Learning Process in Analyzing Crimes

Bathula Rakesh[1], Ruka Sairam Yadhav[2], M Rakesh[3], Dr. Y. Sreenivasulu[4]
[1, 2]*UG scholar, ECE department, SNIST, Hyderabad*
[3, 4]*Associate Professor, ECE department, SNIST, Hyderabad*

***Abstract:** This research explores its use of machine learning to predict crime. The data from the last 15 years of Vancouver crime is studied using two alternative data-processing methods in this study approaches. A crime is detected through machine-learning predictive techniques such as K-nearest Neighbour and boosted decision tree. When the forecast accuracy is between 39 percent and 44 percent, Vancouver crime forecasting*
***Keywords:** data analysis, machine learning, and crime prediction.*

I. PRESENTATION

Crime is a socioeconomic issue that has a negative impact on life quality and economic progress. Depending on the type of civilization and community, the intricacies of how crime is committed differ. Previous crime prediction studies have discovered that characteristics such as education, poverty, employment, and climate have an impact on crime rates. The most populated, ethnically diverse, and multi-cultural metropolis in Canada is Vancouver. Although Vancouver's general crime rate decreased by 1.5 percent in 2017, significant vehicle break-ins and theft remain a problem. The Vancouver Police Department (VPD) recently deployed a crime predictive model to predict crimes linked to property break-ins, and the city of Vancouver saw a 27 percent reduction in home break- ins after it was installed. Crime prediction is a tactic used by law enforcement. Technique for identifying the most likely crimes based on data and statistical analysis. This area in many regions of the world, study is still ongoing in the world. Machine learning is the science of allowing computers to make decisions. Without the need for human involvement Machines have recently learning has been used in self-driving cars and voice recognition. Web search, recognition, and a better grasp of the human genome It has also made crime prediction based on data easier. Data with references possible the process of classification is supervised. Nominal class labels are allowed in this prediction approach. Weather classification has been applied to a wide range of applications. Forecasts, medical care, financial and banking services, and homeland security business intelligence and security. Data collection, categorization, pattern identification, prediction visualization, and association analysis are all part of machine-learning-based crime investigation. Traditional data mining techniques such as classification and prediction, cluster analysis, and outlier analysis find patterns in structured data, whereas newer techniques find patterns in both structured and unstructured data. The project's main purpose is to create a model that can accurately anticipate crime. The VPD crime dataset was examined using two classification algorithms: K-Nearest Neighbour (KNN) and enhanced decision tree, which was generated between 2003 and There were around 560,000 records in 2018. The data was processed using two different methodologies:

1) When a specific crime occurred in a specific neighborhood, the first approach assigned a unique number to each neighborhood and crime category.
2) In the second method, a binary number was assigned to the neighborhood and the day of the week on which the crime occurred, with 1 indicating that the crime occurred on that day in that neighborhood and 0 indicating that it did not the remainder of this document is structured as follows: The second section of this study gives a summary of previous research on machine-learning-based crime prediction. The data-analysis and machine-learning methodologies utilized in this work are outlined in sections III and IV, and the results are presented and compared. Section V contains the conclusions.

II. CONNECTED WORK

Various scholars have looked at the issues of crime control and suggested various crime-prediction systems. The attributes employed and the dataset used as a reference determine prediction accuracy. To estimate crime hotspots in London, UK, human behavioral data from mobile network activity was combined with demographic data generated from real crime data.

WEKA, an open-source data mining software, and 10-fold cross-validation were used to compare two classification methods, Decision Tree and Nave Bayesian, in the socioeconomic, law enforcement, and crime datasets for this study were created using data from the 1990 US Census, 1 US LEMAS survey from 1990 and FBI UCR from 1995, respectively, investigated Ethiopian traffic accident patterns, taking into account a variety of characteristics such as the driver, weather, vehicle, and road conditions. Three different classification algorithms, KNN, KNN, and KNN, were used on a dataset of 18,288 accidents.

Nave Bayesian, and Decision tree, were applied. The accuracy of the percentage of people who used three algorithms ranged from 79 to 81 percent. The precise and efficient analysis of big crime datasets is a major difficulty in crime prediction. Data mining is used to swiftly and efficiently identify hidden trends in big criminal datasets. The accuracy of crime prediction improves as crime data-mining algorithms become more efficient and error-free. Based on the success of the Cop Link experiment at the University of Arizona, a general data-mining framework was built in the majority of crime prediction research focuses on finding crime hotspots, or areas with greater crime rates than the national average.

Kernel Density Estimation (KDE) and Risk Terrain Modeling were compared by the authors (RTM) methods for constructing hotspot maps and developed area-specific predictive models based on sparse data. For crime hotspot prediction, researchers used a spatial-temporal model based on histogram-based statistical approaches, Linear Discriminant Analysis (LDA), and KNN. In Bangladesh, the Gamma test was used to train an Artificial Neural Network (ANN) to predict crime hotspots. analyzed drug-related crime data in Taiwan and predicted rising hotspots using a data- driven machine-learning method based on broken-window theory, geographical analysis, and visualization approaches.

The authors used a reverse-geocoding technique and a density-based clustering algorithm to develop a machine-learning model for crime prediction utilizing Open Street Map (OSM) and geospatial data in the province of Nova Scotia, Canada. suggested a feature-level data- fusion strategy for forecasting crimes in the City of Chicago based on a Deep Neural Network (DNN) trained by spatial-, temporal-, environmental-, and joint-feature representation layers. Knowledge Discovery in Databases (KDD) approaches were used to examine numerous crime-prediction strategies.

A method for crime prediction that combines statistical modelling, machine learning, database storage, and AI technologies has been proposed and suggested a transfer-learning system for exploiting cross domain urban datasets, Weather data, points of interest, people mobility data, and complaint data that capture temporal-spatial tendencies.

In a fully probabilistic algorithm based on a Bayesian approach was used to model the relationship between crime data and environmental factors such as demographic characteristics and geographic location in the Australian state of New South Wales. WEKA was used to examine the accuracy and effectiveness of linear regression, additive regression, and decision stump algorithms for forecasting crime in Mississippi, according to a study published in. In this survey study on criminal data mining, the authors of presented ANN, decision trees, rule induction, nearest-neighbor technique, and genetic algorithm are all explored. Using a technique based on the Auto-Regressive Integrated Moving Average model, we built a reliable prediction model for forecasting crime trends in urban areas (ARIMA).

Authors suggested a probabilistic model of spatial behaviors for known criminals in the Metro Vancouver area in a based on a random-walk-based approach to model offender activity. In order to investigate the impact of urban characteristics in crime prediction in Brazil, researchers used the random forest method. Prospective method, the multi-kernel approach and the Dempster-Shafer theory of evidence were combined to produce a crime prediction solution for Chilean major towns. In the city of San Francisco, three approaches for anticipating crimes were developed, tested, and evaluated Parzan windows, and Neural Networks. The weighted page-rank method was utilized as an effective way to weaken and dismantle criminal networks in, where the Gradient Boosting Machine (GBM) technology was applied in a machine-learning prediction model to discover hidden relationships in criminal networks.

According to the literature, the classification algorithms KNN and boosted decision tree were utilized to examine the VPD crime dataset in this study.

III. ANALYSIS OF DATA

A. *Information Source*

The open data library of the city of Vancouver provided the source datasets. Two datasets were used in this experiment: crime and neighborhood. The VPD has been collecting crime data since 2003, and it is updated every Sunday morning. It details the sort of crime committed, as well as the date and place of the offence. In the Geographic Information System, the neighborhood dataset comprises the boundaries for the city's 22 local regions (GIS). The crime dataset is used for data analysis, while the neighborhood dataset is used to build maps in this project

B. *Preprocessing*

The original dataset is required to fill empty cells, remove extraneous columns, and add various relevant features.

Figure 1 depicts the original and preprocessed datasets.

TYPE	YEAR	MONTH	DAY	HOUR	MINUTE	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y
Other Theft	2003	8	8	13	58	10XX ROBSON ST	West End	490998.3	5459018
Theft from Vehicle	2003	8	14	9	30	12XX CHESTNUT ST	Kitsilano	489368.8	5458066
Mischief	2003	4	4	19	15	16XX COMMERCIAL DR	Grandview-Woodlan	494928.6	5457524
Theft from Vehicle	2003	6	15	22	0	10XX E BROADWAY AVE	Mount Pleasant	494071.3	5456637
Offence Against a Person	2003	10	12			OFFSET TO PROTECT PRIVACY		0	0
Offence Against a Person	2003	10	26			OFFSET TO PROTECT PRIVACY		0	0
Mischief	2003	4	6	23	0	32XX W 39TH AVE	Dunbar-Southlands	487196.8	5453771
Break and Enter Residenti	2003	1	28	18		8th 18XX VENABLES ST			

(a)

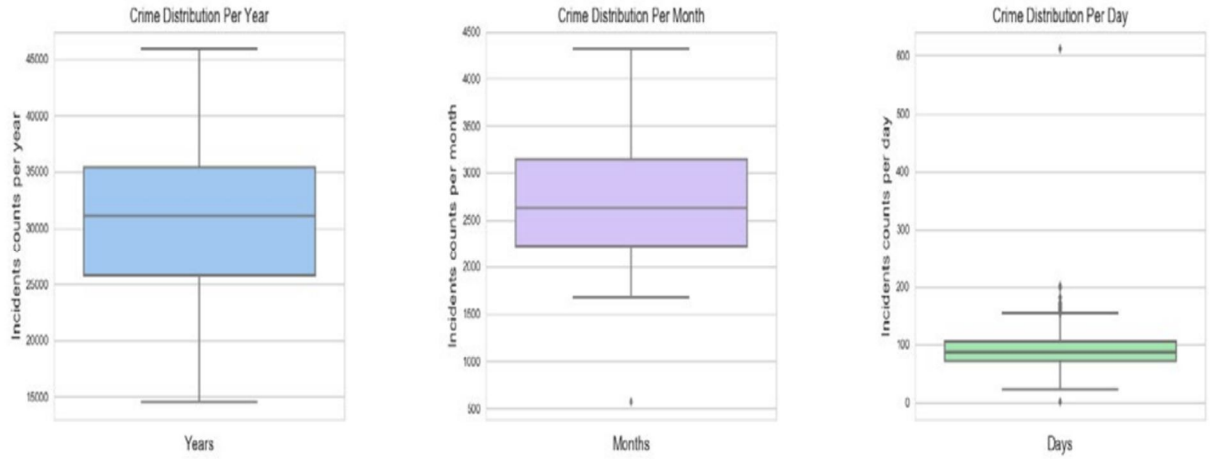
TYPE	YEAR	MONTH	DAY	HOUR	HUNDRED_BLOCK	NEIGHBOURHOOD	X	Y	DATE	WEE
other theft	2003	8	8	13	10XX ROBSON ST	west end	490998.3	5459018	2003-08-08	Frid
theft from vehicle	2003	8	14	9	12XX CHESTNUT ST	kitsilano	489368.8	5458066	2003-08-14	Thu
mischief	2003	4	4	19	16XX COMMERCIAL DR	grandview-woodla	494928.6	5457524	2003-04-04	Frid
theft from vehicle	2003	6	15	22	10XX E BROADWAY AVE	mount pleasant	494071.3	5456637	2003-06-15	Sun
offence against a pe	2003	10	12	0	OFFSET TO PROTECT PRI	n/a	0	0	2003-10-12	Sun
offence against a pe	2003	10	26	0	OFFSET TO PROTECT PRI	n/a	0	0	2003-10-26	Sun
mischief	2003	4	6	23	32XX W 39TH AVE	dunbar-southlands	487196.8	5453771	2003-04-06	Sun
break and enter res	2003	1	28	18	18XX VENABLES ST	grandview-woodla	495047.8	5458317	2003-01-28	Tue

(b)

Fig 1. The snapshot of the (a)original and (b) preprocessed datasets

C. Analytical Statistics

The crime dataset described in Figure 2 is distributed by year, month, and day Each year, 31624 crimes are reported in Vancouver, with 2720 crimes each month and 90 crimes every day. The dataset tends to reflect a normal distribution as the time intervals increase. The graph of each day, however, shows an unusual maximum value of 650 events, which is thought to be an anomaly - and turns out to be the Stanley Cup riot on June 15, 2011.



(a) (b) (c)

D. Analysis of Trends

From 2003 to 2013, the average number of offences per month fell, as seen in Fig. 3 and climbed in 2016, and then decreased somewhat to around 3000 events per year in 2018.The summer season and the middle of each month are the most dangerous, according to the time-featured heat-map graphs in Fig. 4. There are also more crimes on Fridays, Saturdays, and late at night. The heat map indicated the largest values near zero hours, which should be ignored because all the empty data cells were filled with zero. The category of theft from cars had the highest number of instances, followed by mischief. However, while auto theft has decreased dramatically in recent years, other types of theft have surged. Figure 5 depicts the number and trends of each offence.

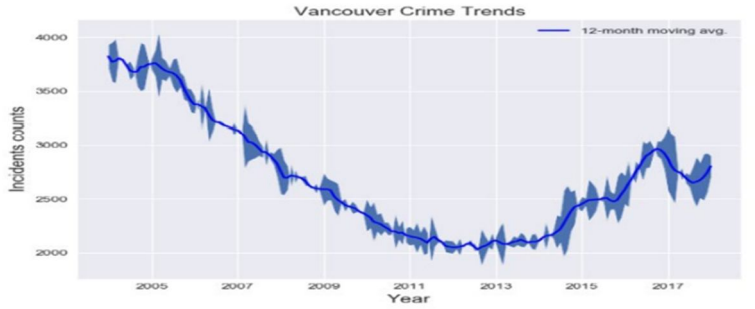
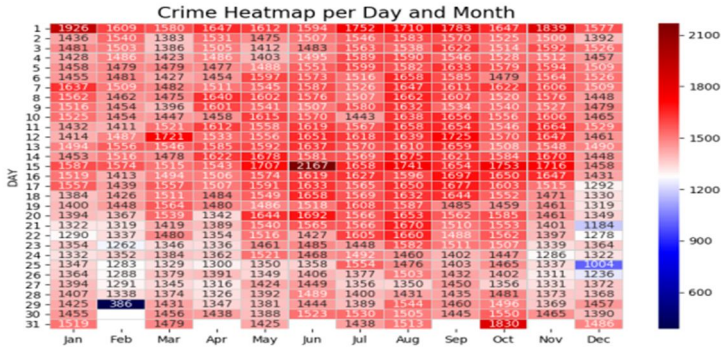
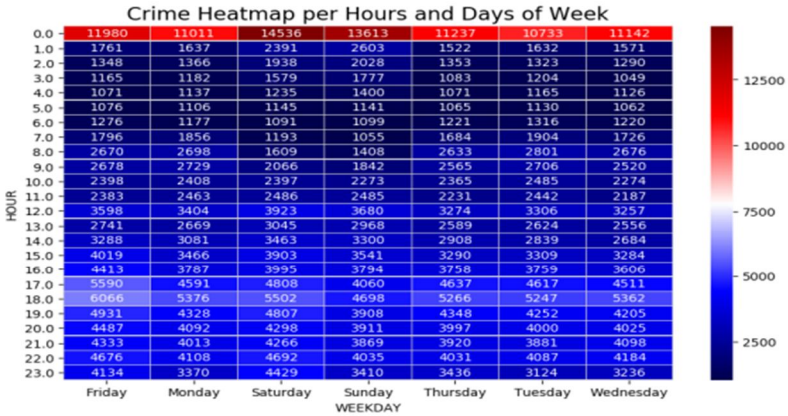


Fig. 3. Moving average of crimes per month

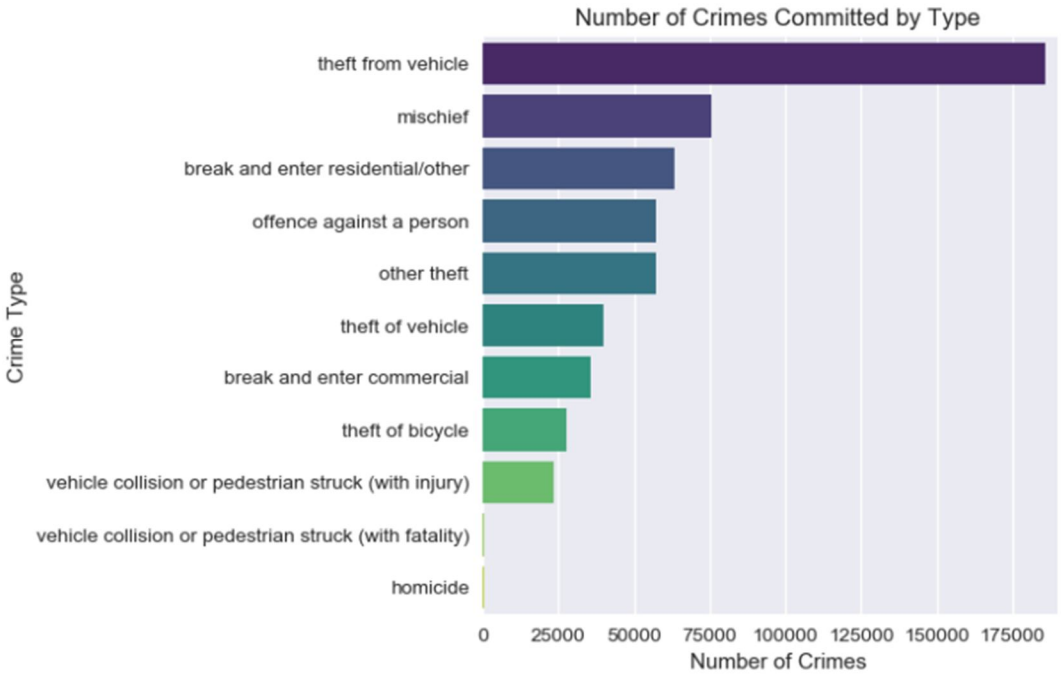


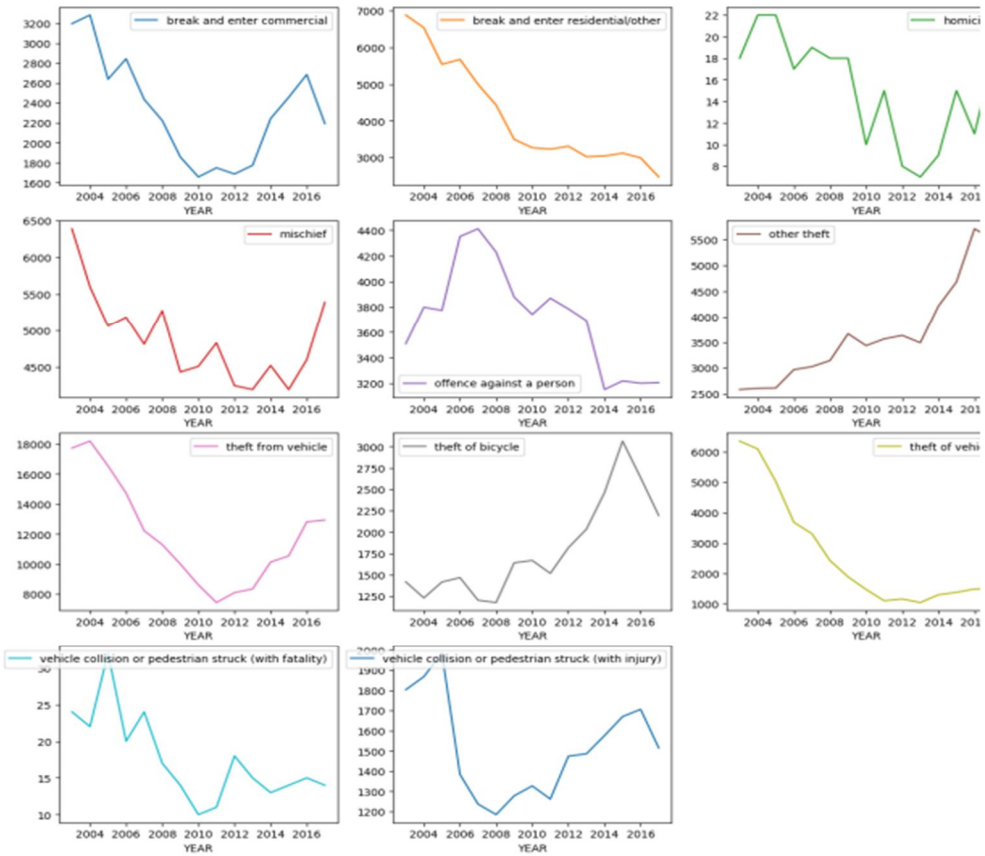
(a)



(b)

Fig. 4. The crime heatmap (a) per days and month, and (b) per hours and days of week





(b)

Fig. 5. (a) Number and (b) trend of crimes committed by type

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

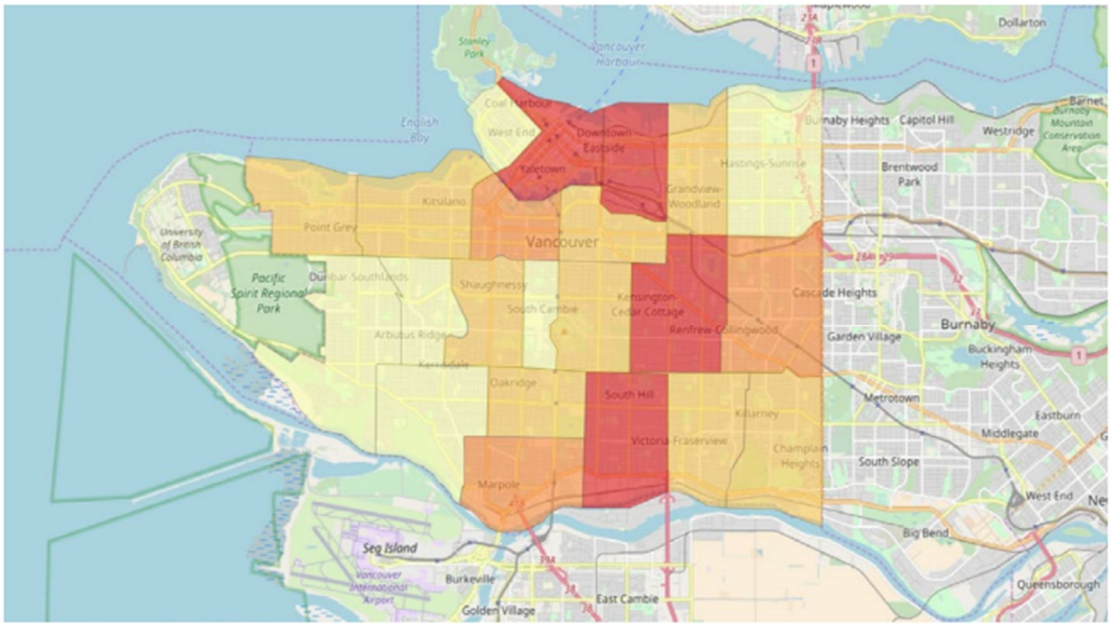
E. Geographical Research

There are other methods for mapping hotspots, however choropleth mapping is one of the most extensively used to convey the geographic information of criminal incidents with shaded colors, A choropleth map illustrates the percentage or density of statistical measurements. This makes it simple to spot areas where crime episodes are concentrated, providing insight into criminal behaviors. A Geographic Information System (GIS) has been used as a powerful analytical tool for crime mapping. It displays the locations of crime series along with relevant geographic data on a single map, assisting police officers in making operational and tactical decisions. As a first step toward geographical analysis, the neighborhood boundary dataset was translated from the Universal Transverse Mercator (UTM) to the World Geodetic System 1984 (WGS84), also known as latitude and longitude. Py Sal, Geo Pandas, Folium, and Shapely are some of the supporting libraries for visualizing geographic data that were utilized to create the map. Crime incidences were tallied for each neighborhood to illustrate crime hotspots on Vancouver's city map.

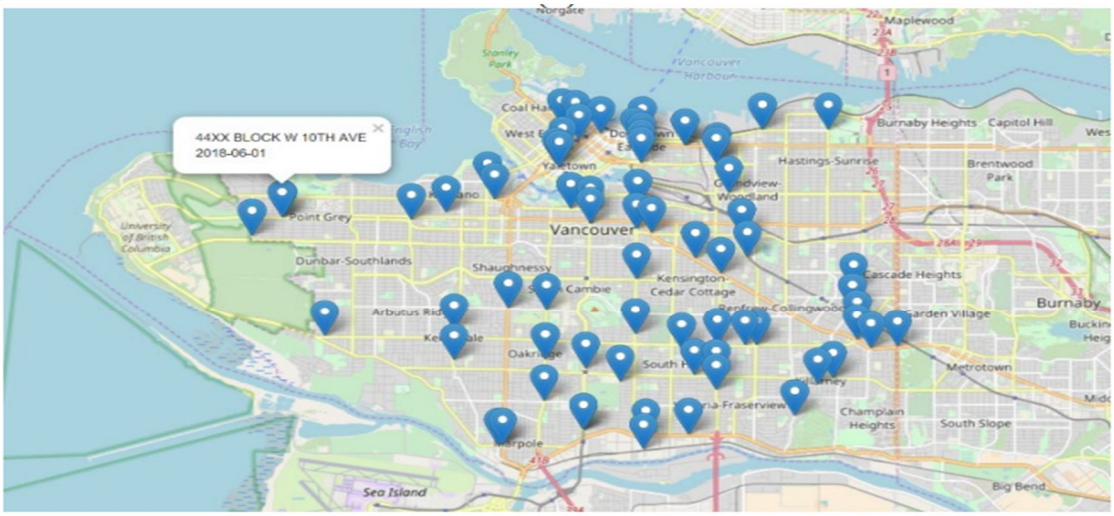
As depicted on the choropleth map, Table 1 displays the top 10 crime-intense neighborhoods over a 30-day period. The hotspot map and point clusters of incidents that happened in the city of Vancouver over a 30-day period are depicted in Figure 6.

Table 1. Top-10 crime-dense neighbourhoods

Map ID	Name	Density (per square miles)
CBD	Downtown	3.938709
SUN	Sunset	3.150967
KC	Kensington-Cedar Cottage	2.363225
STR	Strathcona	2.363225
RC	Renfrew-Collingwood	1.969355
MARP	Marpole	1.575484
FAIR	Fairview	1.575484
MP	Mount Pleasant	1.181613
KITS	Kitsilano	1.181613
OAK	Oakridge	1.181613



(a)



(b)

Fig. 6. The city of Vancouver's (a) hotspot map, and (b) incident point clusters in a 30-day time

IV. MACHINE LEARNING

Machine learning is a subset of artificial intelligence that employs statistical methods to enable computers to learn from their previous experiences. There are three types of machine learning: supervised, unsupervised, and reinforcement learning This study uses supervised learning due to the nature of the required input data and output targets. Two types of supervised learning are classification and regression. Classification and regression are two types of supervised learning. Regression is the challenge of forecasting a continuous quantity, whereas classification is the task of predicting a discrete class label. The goal of this study is to predict the types of crimes that will occur in a specific place. As a result, the study's purpose is to classify crimes. Some of the approaches that can be used for classification are K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Nave Bayesian, Decision Tree, and Ensemble Methods. In terms of complexity, accuracy, and training time, each method has its own advantages and disadvantages, and different results can be obtained from a single dataset. In this study, we used KNN and decision-tree approaches to train our model. KNN is one of the most fundamental categorization algorithms. The sample z is assigned to class A if it has the most values. The sample belongs to class A if it is closer to z; else, it belongs to class B. The KNN is calculated using the formula below. The probability that the test sample will fall into category Ci:

$$P(x \in C_i) = \frac{\sum_{j \in C_i} {}^{n_j}/_{d_j}}{\sum_{l=1}^{k} {}^{n_l}/_{d_l}} \qquad (1)$$

where nj and nl are the number of elements represented by each training dataset sample, and dj and dl are the test and corresponding training sample distances calculated using the Euclidean norm [24]. KNN saves all existing objects and classifies new objects using the similarity measure by looking for the input values' closest Neighbours. Decision trees, on the other hand, are superior at dealing with huge datasets with multiple layers and nodes. While restricting the number of possible decision points, decision-tree classifiers can provide a better combination of flexibility and accuracy The decision-tree-classification approach constructs a tree structure from the dataset by dividing it into smaller pieces. With the use of two functions: Gini impurity and information gain, the decision tree selects a feature that best separates the data at each stage in the process. The Gini impurity assesses the likelihood of erroneously classifying a random sample.

$$I_G(p) = \sum_{i=1}^{k} p_i(1 - p_i) \qquad (2)$$

Gaining knowledge aids in deciding which feature to separate next. Entropy can be used to determine information gain.

$$H(T) = I_E = -\sum_{i=1}^{k} p_i \log_2(p_i) \qquad (3)$$

where pi is the percentage of each feature that remains in the child node after a split.

Before the algorithms can be employed, the data must be translated into a readable format. We tested two distinct data processing procedures and compared the results.

A. First Approach

Before the algorithms can be employed, the data must be translated into a readable format. Every neighborhood and day were turned into a showpiece. Only the correct variable receives a "1," whereas all other variables receive a "0." All variables with the value "0" are dummy variables. This provides additional variables for the algorithm to learn and avoids data from skewing to one side. Experiments with skewed data yield a fake accuracy of 98.9%, which is untrustworthy.

B. Technique 2

In the second approach, categorical variables are converted to numerical variables with unique IDs. Distinct crime types and neighborhoods are assigned different IDs. For example, the crime type ID 10 is attributed to vehicle theft. The work done on Kaggle inspired this technique.

The same algorithms are used in both ways, with the same settings and validation processes. 5-fold cross-validation is used to test the classifier algorithm's validity. The algorithm has been honed to recognize a certain type of crime.

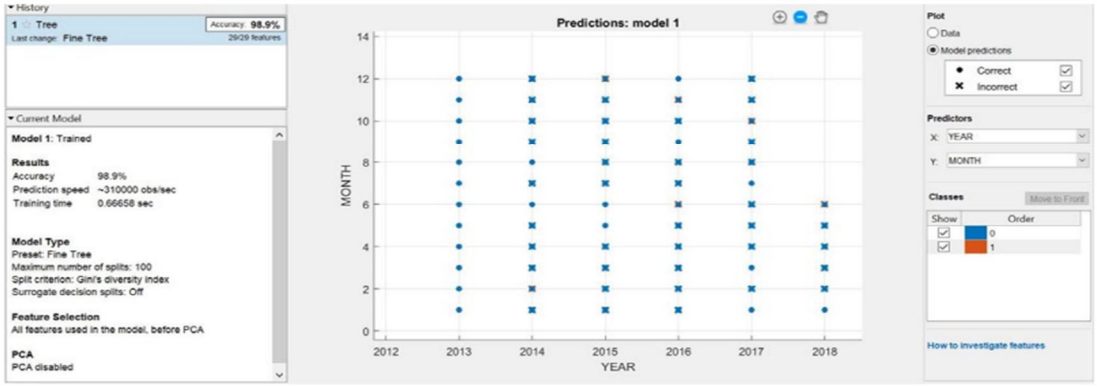
Cross-validation reduces overfitting and guarantees that the prediction model performs well on new data that hasn't been seen before. Figure 7 shows the treated dataset from approaches 1 and 2, as well as the skewed data from approach 1.

	A	B	C	D	E	F	G	H
1	YEAR	CrimeType	MONTH	DAY	HOUR	WEEKDAY_	WEEKDAY_	WEEKDAY_
2	2003	6	8	8	13	1	0	0
3	2003	7	8	14	9	0	0	0
4	2003	4	4	4	19	1	0	0
5	2003	7	6	15	22	0	0	0
6	2003	5	10	12	0	0	0	0
7	2003	5	10	26	0	0	0	0
8	2003	4	4	6	23	0	0	0
9	2003	2	1	28	18	0	0	0
10	2003	2	2	3	17	0	1	0
11	2003	7	3	9	21	0	0	0
12	2003	7	6	13	13	1	0	0
13	2003	7	6	13	16	1	0	0

(a)

	A	B	C	D	E	F
1	YEAR	MONTH	DAY	HOUR	neighbour	crime_type_id
2	2003	8	8	13	24	6
3	2003	8	14	9	10	7
4	2003	4	4	19	5	4
5	2003	6	15	22	12	7
6	2003	10	12	0	14	5
7	2003	10	26	0	14	5
8	2003	4	6	23	3	4
9	2003	1	28	18	5	2
10	2003	2	3	17	5	2
11	2003	3	9	21	24	7
12	2003	6	13	13	4	7
13	2003	6	13	16	24	7
14	2003	3	22	16	24	7
15	2003	2	10	17	5	2
16	2003	5	10	0	14	5
17	2003	4	19	12	5	2
18	2003	2	18	22	3	4

(b)



(c)

C. K-Nearest Neighbour (KNN)

KNN was used with the same parameters in both techniques, and the accuracy and training time were compared.

For approach 1, KNN had a 40.1 percent accuracy and a training time of 2209 seconds, while it was 39.9 percent accurate and training time was 101.73 seconds for approach 2.

D. Decision Tree Boosted

We examined the results of both approaches using the boosted decision tree algorithm. We used the Adaptive Boosting (AdaBoost) ensemble method and a learner-type decision tree for both approaches. AdaBoost is a meta-algorithm that combines many weak learners to improve a poor classifier.

The most splits allowed was 20. Approach 1 was 41.9 percent correct and took 903.63 seconds to train, whereas Approach 2 was 43.2 percent accurate and took 459.26 seconds to train. Figure 8 shows the outcomes of both methods (KNN and boosted decision tree) for both approaches.

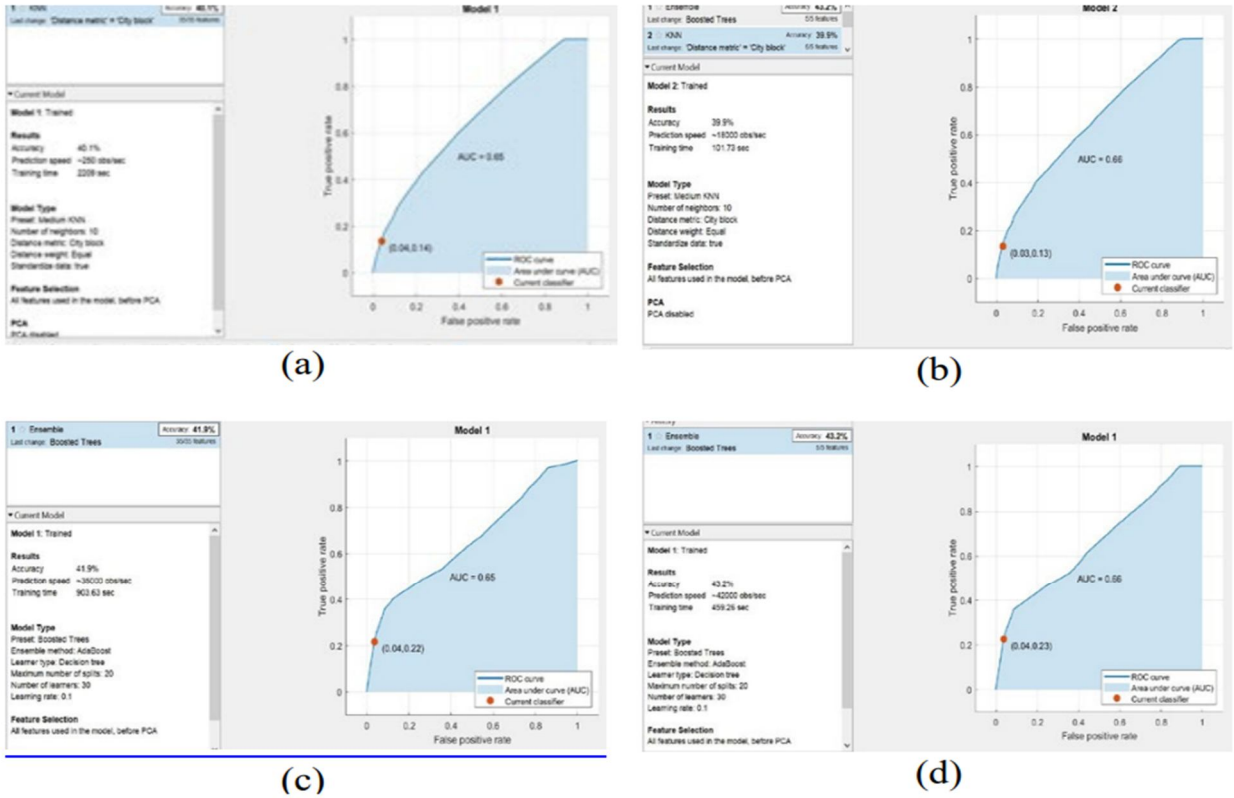


Fig. 8. KNN results for approaches 1 and 2 (a, b) and boosted decision tree results for approaches 1 and 2 (c, d)

V. FINAL REMARKS

Vancouver crime data from the last 15 years was used in two separate dataset techniques in this study. The KNN and boosted decision tree machine learning predictive models were employed to achieve criminal prediction accuracy of 39 to 44 percent. Depending on the method and algorithm, algorithm accuracy, complexity, and training time vary slightly.. Both the algorithm and the data can be fine-tuned for specific applications to increase prediction accuracy. Although this model's prediction accuracy is low, it gives a rough framework for additional research.

VI. ACKNOWLEDGMENT

This article is based on a capstone design project conducted as part of a premaster's degree in product realization at Fraser International College (FIC). Ms. Amy Yeung, Dr. Alberta Seah, Ms. Sandra Kimber, and Ms. Sharla Reid provided invaluable assistance to the authors.

REFERENCES

[1] R. Iqbal, M. A. A. Murad, A. Mustapha, P. H. Shariat Panahi, and N. Khanahmadliravi, "An experimental study of classification algorithms for crime prediction," Indian J. of Sci. and Technol., vol. 6, no. 3, pp. 4219- 4225, Mar. 2013.

[2] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," IEEE Computer, vol. 37, no. 4, pp. 50-56, Apr. 2004.

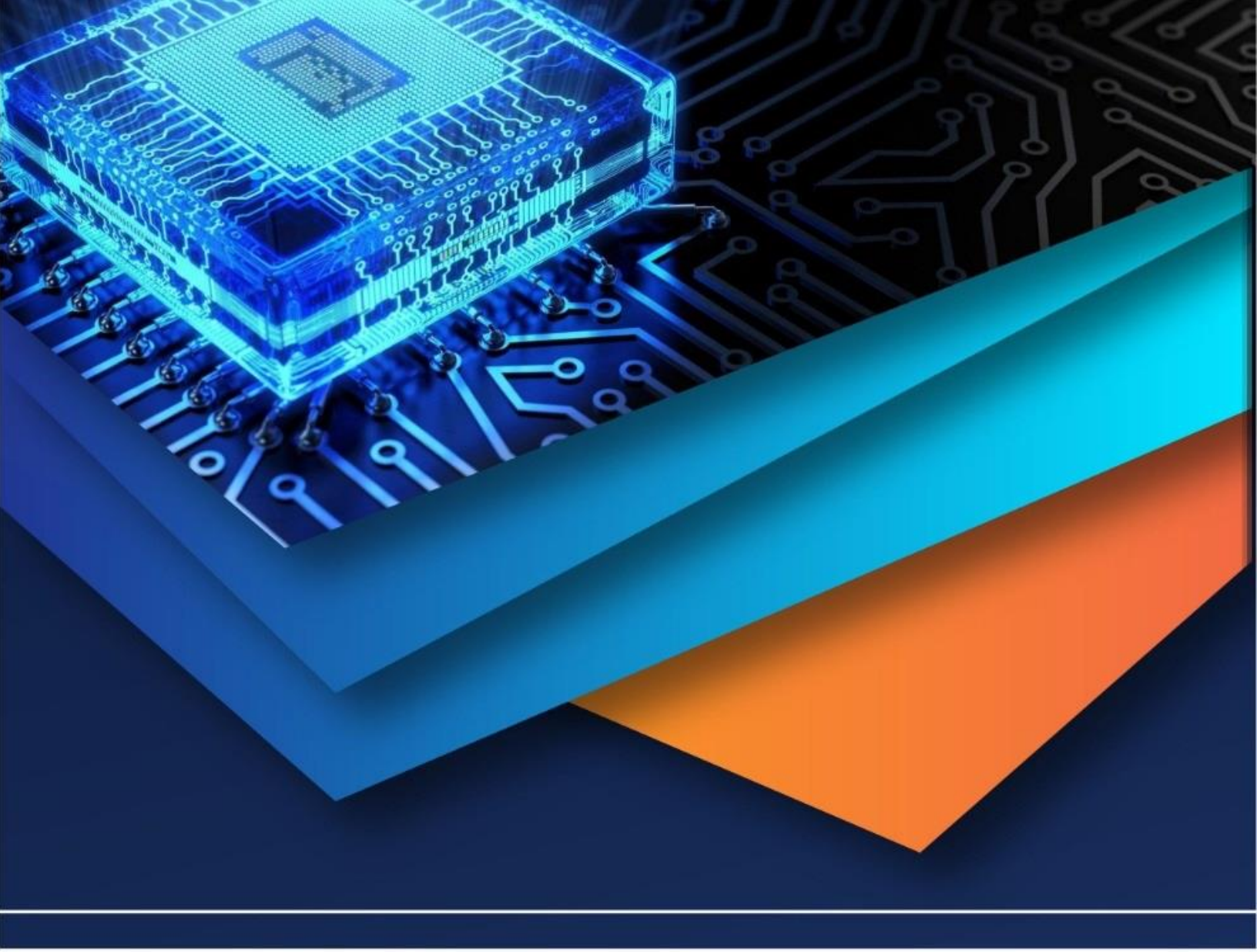
[3] T. Beshah and S. Hill, "Mining Road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia," Proc. of Artificial Intell. for Develop. (AID 2010), pp. 14-19, 2010.

[4] M. Al Boni and M. S. Gerber, "Area-specific crime prediction models," 15th IEEE Intl. Conf. on Mach. Learn. and Appl., Anaheim, CA, USA, Dec. 2016.

[5] Q. Zhang, P. Yuan, Q. Zhou, and Z. Yang, "Mixed spatial-temporal characteristics-based crime hot spots prediction," IEEE 20th Intl. Conf. on Computer. Supported Cooperative Work in Des. (CSCWD), Nanchang, China, May 2016.

[6] N. Mahmud, K. Ibn Zannah, Y. Ar Rahman, and N. Ahmed, "CRIMECAST: a crime prediction and strategy direction service," IEEE 19th Intl. Conf. on Computer. and inform. Technol., Dhaka, Bangladesh, Dec. 2016.

[7] Y. L. Lin, L. C. Yu, and T. Y. Chen, "Using machine learning to assist crime prevention," IEEE 6th Intl. Congar. on Advanced Appl. Inform. (IIAIAAI), Hamamatsu, Japan, Jul. 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY