JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY
https://e-journal.uum.edu.my/index.php/jict

How to cite this article:
Khairuddin, A. R., Alwee, R., & Haron, H. (2023). Hybrid neighbourhood component analysis with gradient tree boosting for feature selection in forecasting crime rate. *Journal of Information and Communication Technology, 22*(2), 207-229. https://doi.org/10.32890/jict2023.22.2.3

Hybrid Neighbourhood Component Analysis with Gradient Tree Boosting for Feature Selection in Forecasting Crime Rate

*[1]Alif Ridzuan Khairuddin, [2]Razana Alwee & [3]Habibollah Haron
Applied Industrial Analytics Research Group (ALIAS),
Faculty of Computing,
Universiti Teknologi Malaysia, Malaysia

*[1]alifridzuan@utm.my
[2]razana@utm.my
[3]habib@utm.my
*Corresponding author

Received: 21/9/2022 Revised: 5/12/2022 Accepted: 2/1/2023 P ublished: 3/4/2023

ABSTRACT

Crime forecasting is beneficial as it provides valuable information to the government and authorities in planning an efficient crime prevention measure. Most criminology studies found that influence from several factors, such as social, demographic, and economic factors, significantly affects crime occurrence. Therefore, most criminology experts and researchers study and observe the effect of factors on criminal activities as it provides relevant insight into possible future crime trends. Based on the literature review, the applications of proper analysis in identifying significant factors that

influence crime are scarce and limited. Therefore, this study proposed a hybrid model that integrates Neighbourhood Component Analysis (NCA) with Gradient Tree Boosting (GTB) in modelling the United States (US) crime rate data. NCA is a feature selection technique used in this study to identify the significant factors influencing crime rate. Once the significant factors were identified, an artificial intelligence technique, i.e., GTB, was implemented in modelling the crime data, where the crime rate value was predicted. The performance of the proposed model was compared with other existing models using quantitative measurement error analysis. Based on the result, the proposed NCA-GTB model outperformed other crime models in predicting the crime rate. As proven by the experimental result, the proposed model produced the smallest quantitative measurement error in the case study.

Keywords: Feature Selection, Artificial Intelligence, Neighbourhood Component Analysis, Gradient Tree Boosting, Crime Forecasting.

INTRODUCTION

In the real world, crime is a part of society that cannot be predicted by the police (Ghazvini et al., 2015). The crime rate itself represents the degree of public safety in a country. It is known that the change in crime rates is used as an indicator of macroeconomic development. The purpose of the crime rate is for strategic decision-making in formulating crime prevention strategies. Therefore, analysing crime data helps to understand future crime patterns through forecasting (Shrivastav & Ekata, 2012).

There are two types of crime forecasting models proposed by different researchers, namely statistical and artificial intelligence (AI) models. The statistical model adapts several statistical techniques, such as linear regression, moving average, exponential smoothing, and autoregressive integrated moving average (ARIMA), in analysing past or present crime data trends to estimate crime patterns in the future. Meanwhile, the AI model adopts machine learning techniques in evaluating the possible outcome of crime. Artificial Neural Network (ANN) and Support Vector Regression (SVR) are among the popular applied AI models in crime forecasting. In past years, it has been observed that researchers have shifted their research interest

from statistical models to AI models in crime forecasting. One of the reasons is that the statistical model is incapable of handling abrupt changes in any type of environment or system (Baliyan et al., 2015).

A past study showed that crime is influenced by various factors (Hanslmaiere et al., 2015). Previous researchers have conducted studies to observe the influence of several factors on crime, such as economic factor (Habibullah & Baharom, 2009; Alwee, 2014; Osborn, 2018; Wang & Hu, 2022), social factor (Hipp & Yates, 2011; Hanslmaiere et al., 2015; Mills et al., 2017; Anser et al., 2020), and demographic factor (Brown & Males, 2011; Ranson, 2014; Kim, 2018; Blakeslee et al., 2021). These studies have provided relevant insight into possible future crime trends based on recent issues. In assessing this type of analysis, multivariate crime forecasting analysis is considered. In multivariate crime analysis, extensive studies have been conducted to observe the relationships between factors and their impact on crime (Gorr & Thompson, 2003; Li et al., 2010; Alwee, 2014; Vineeth et al., 2016; Quick et al., 2018; Chen, 2022). Studies on the influence of several factors in crime analysis are highly beneficial because crime occurrence patterns are not heavily dependent on past crime trends but are affected by various factors, such as social mistreatment, population densities, and economic disadvantages.

This study aims to propose an artificial intelligence-based crime model to forecast the United States (US) crime rate data. The proposed model is also hybridised with a feature selection technique to evaluate the nine influential factors that have potentially affect the US crime rate data. The hypothesis is that by selecting significant factors that influence crime rate, a better prediction accuracy can be achieved compared to the model that uses all available factors. This study implemented Gradient Tree Boosting (GTB) as the selected artificial intelligence model in developing the proposed model. For the feature selection technique in identifying significant factors that influence crime, Neighbourhood Component Analysis (NCA) was considered.

LITERATURE STUDY ON FEATURE SELECTION IN CRIME FORECASTING

Feature selection is an effective solution when handling a multivariate model because it can extract the main features in a dataset and, at

the same time, minimise the model input dimension (Han & Wang, 2009). In forecasting crime using multivariate analysis, using the entire available features (factor data in this case study) to develop the crime model is inefficient. Even though the multivariate model is able to discover more information about the complex system, using insignificant or irrelevant feature data results in the model being prone to overfit and having poor generalisation capabilities (Han & Wang, 2009). Therefore, significant features must be properly identified to avoid the mentioned problems. In addressing such issues, feature selection can be used to find the strong relationship between dependent (crime rate) and independent (factors that influence crime) variables. Implementing a feature selection technique helps to discover a new crime pattern that has never occurred in the past (Alwee, 2014).

Prior studies found that the influence of several factors, such as social, demographic, and economic, significantly impacts crime occurrence (Ranson, 2014; Soundarya et al., 2017; Stansfield et al., 2017). It has been observed that multivariate analysis in crime forecasting is beneficial in improving forecasting performance capabilities. In the literature, several approaches are used by various researchers in selecting the factors that affect crime; these are presented in Table 1.

Table 1

Factor Selection Approaches by Different Researchers in Crime Forecasting

Literature	Factor Selection Approach	Number of Factors
Yearwood and Koinis (2011)	No analysis of factors.	12
Iqbal et al. (2013)	• Selection was based on human understanding and intellect. • Selected factors did not contain any missing value.	128
Alwee (2014)	Grey Relational Analysis (GRA)	4
Babakura et al. (2014)	Selection was based on human understanding and intellect.	128
Bogomolov et al. (2014)	• Pearson Correlation Analysis • Principal Component Analysis (PCA) • Gini Coefficient of Inequality	68

(continued)

Literature	Factor Selection Approach	Number of Factors
McClendon and Meghanathan (2015)	Selection was based on plausible connections to potential crime goals of the study.	128
Castelli et al. (2017)	Selection was based on plausible connections to potential crimes of the study suggested from other works.	128
Nguyen et al. (2017)	No analysis of factors.	21
Liu et al. (2019)	Fuzzy Rough Set-Based	4
Shi (2020)	Random Forest	30

An analysis shows that in most cases, researchers choose significant features using manual selection based on their logical understanding and knowledge. Such approaches are impractical as they may lead to selection bias and misinterpretation of the relevant features under certain conditions (Aldehim & Wang, 2017). Furthermore, it can be observed that some researchers use all the available collected features. They did not apply any statistical approach in selecting the significant features. This approach might lead to another problem; some of the selected features might not have helpful information and are considered irrelevant. These irrelevant features are a burden and serve as pure noise that negatively affects the overall model performances (Chandrashekar & Sahin, 2014).

The identified problem provides a strong argument that applying an appropriate feature selection technique in developing a crime forecasting model is essential in determining the relevant and significant features. It also provides a scientific justification to determine whether the selected features are statistically significant. In addition, an appropriate feature selection technique helps to reduce the dimensionality of features (Sainin et al., 2021). Therefore, this study addressed such issues by introducing an efficient feature selection technique for determining significant external factors that influence crime rates. Motivated by this intent, this study proposed a proper factor selection analysis by implementing a non-parametric embedded feature selection technique called Neighbourhood Component Analysis (NCA) in identifying the significant external factors (features) that influence the patterns of crime.

Neighbourhood Component Analysis (NCA)

Neighbourhood Component Analysis (NCA) is a non-parametric feature selection technique that applies an embedded method to select relevant features that can improve prediction and classification accuracy. The application of NCA for feature selection in various domains and case studies has recently been studied and proposed by researchers (Zheng et al., 2016; Tonkal et al., 2021; Malan & Sharma, 2022). It was introduced by Yang and Wang (2012), where the work motivation was to improve the K-Nearest Neighbour (KNN) algorithm. NCA is a feature weighting method based on the nearest neighbour approach that applies the gradient ascent technique to maximise the expected leave-one-out accuracy with a regularisation term (Yang & Wang, 2012). NCA's main objective is to discover a weighting vector, *w*, which is used to determine the relevant feature by optimising the nearest neighbour to solve a classification or regression problem. The advantages of NCA are that it can minimise overfitting during data training and is insensitive to the number of features (Yang & Wang, 2012).

ARTIFICIAL INTELLIGENCE IN CRIME FORECASTING

In past decades, the application of AI techniques, such as Support Vector Regression (SVR), Genetic Programming, and Artificial Neural Network (ANN), in forecasting crime data has been favoured by researchers (Han & Wang, 2009; Alwee, 2014; Castelli et al., 2017; Liu et al., 2019; Shi, 2020). The reason is that AI techniques possess an ability to identify nonlinear patterns in data that statistical techniques lack (Rather et al., 2017). This ability has led to a new discovery of crime patterns that did not happen in the past (Alwee, 2014). As a result, more accurate crime forecasts can be achieved. Inspired by this, the current study selected the AI technique of Gradient Tree Boosting (GTB) to develop the proposed crime forecasting model.

Gradient Tree Boosting (GTB)

Gradient Tree Boosting (GTB) is an ensemble learning model developed by Friedman (2001). GTB has been implemented in different research areas to solve classification and regression problems (Liu et al., 2017; Sun et al., 2020; Guo et al., 2022). It integrates two base learners, i.e., Decision Tree and Boosting techniques in data learning

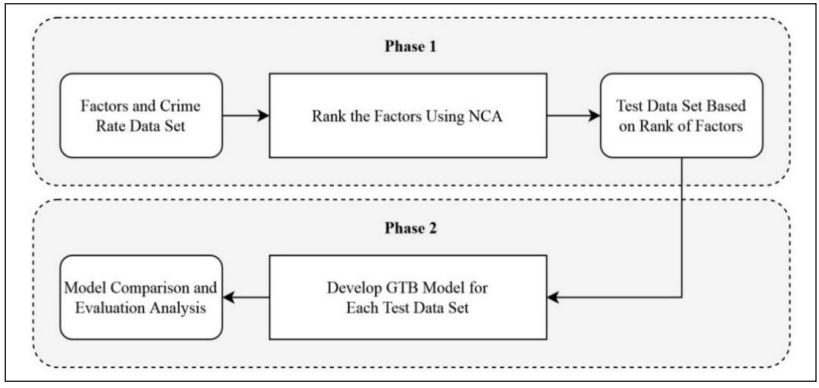
and prediction. GTB's development was inspired by a previously introduced statistical framework, namely Adaptive Reweighting and Combining (ARC) algorithm by Brieman (1997). GTB implements a numerical optimisation approach to reduce the loss function of the predictive model to improve the overall predictive capabilities. GTB can produce robust and interpretable solutions for both classification and regression problems (Friedman, 2001). Moreover, implementing the boosting technique in GTB can reduce the risk of overfitting when adding a new set of data (Friedman, 2001).

PROPOSED NCA-GTB CRIME MODEL

The proposed NCA-GTB model analysed and identified the significant factors to improve the accuracy in forecasting crime rate. It comprised two main phases: ranking the factors using NCA and developing the GTB model for each test data set. In the first phase, ranking the factors using NCA was conducted to evaluate and rank each factor based on its importance. In the final phase, the resulting set of ranked factors test data from the first phase was used to model the crime rate using GTB. This phase allowed observing the impact of applied NCA in identifying significant factors that influence crime, which improved the accuracy of GTB. The observation was made by analysing the calculated quantitative error measurement of the developed crime model. Figure 1 shows the implementation of the proposed NCA-GTB crime model.

Figure 1

Proposed Hybrid NCA-GTB Crime Model



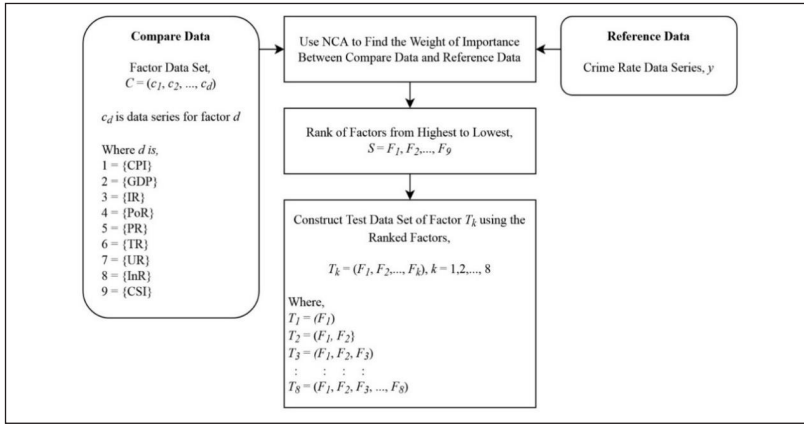
Phase 1: Rank the Factors Using NCA

In the first phase, the ranking of factors using NCA was performed. The implementation of phase 1 is presented in Figure 2. Firstly, NCA analysed the relationship between factors data set *C* and crime rate by assigning a weight of importance to them. The weight of importance defined the level of importance; the higher the value, the more important the factor in the reference crime type. However, if the weight of importance value was zero or negative, there was no significant relationship between factors and crime. Therefore, the factors were considered irrelevant and eliminated. Once the importance weight was assigned to each factor, they were ranked from highest to lowest according to their assigned weight. Rank 1 (F_1) was the factor with the highest weight, rank 2 (F_2) was the factor with the second highest weight and so on until the last rank of factor (F_9) with the lowest weight value.

After the factors were ranked, the test data sets were constructed based on the rank of factors. Note that the last ranked factors were not included in constructing the test data set. If the last ranked factors were also included, this meant all the factors were used in developing the GTB model. The constructed test data set T_k was then used as input data for the next phase, which was developing the GTB model for each test data set.

Figure 2

Implementation of Ranking the Factors Using NCA



Phase 2: Development of GTB Model for Each Test Data Set

In this phase, the output produced in the previous phase, i.e., the constructed test data sets of factors T_k, were used as the input. During the development of each crime model, the prepared factors test data set, T_k, and reference crime rate data, y, were divided into training and test data sets. In developing the GTB model, the model was trained using the specified training data set. GTB was utilised for data fitting (training) in developing the crime model. Once the crime model was developed and trained, it was then used to predict the crime rate using the test data set. Next, quantitative error measurement (root mean square error [RMSE], mean absolute deviation [MAD], and mean absolute percentage error [MAPE]) analyses were conducted to calculate the difference in errors between the model output and the actual value of crime rates. The calculated measurement error result for the GTB model was then analysed and compared with another existing model.

COMPARISON MODEL

In this study, two existing models, Random Forest (RF) by Li et al. (2019) and ReliefF-RF by Zhang et al. (2019), were selected to be compared with the proposed hybrid NCA-GTB crime model. For the first comparison model by Li et al. (2019), the authors proposed an RF-based feature selection method to characterise the importance of multiple factors. The aim was to find the relationship between ridership and crime. Then, the RF model was developed with the selected factors to predict the ridership per capita. The second comparison model developed by Zhang et al. (2019) involved a proposed hybrid of ReliefF and RF for an intrusion detection system. In their work, the ReliefF algorithm was hybridised with RF to calculate the weight of influence factors, and the purpose was to eliminate the redundant information in the original intrusion detection data. They also aimed to overcome RF's slow convergence problem and improve the learning performance.

EXPERIMENTAL SETUP

The experiment was conducted using MATLAB and Python Scikit-learn tools. The implementation of NCA feature selection technique

was conducted in MATLAB. For the development of the GTB crime model, Python Scikit-learn module package tools developed by Pedregosa et al. (2011) were used. The GTB parameters could be configured in these tools to produce reliable forecasting result. Lastly, the quantitative measurement error of the proposed hybrid model was calculated in MATLAB.

Data Collection

The study collected US crime rate data and nine factors data for use in developing the proposed hybrid crime model. Both crime rate and factors data were collected from the period of 1960 to 2015 with 56 samples each. The US crime rate data employed in this study was the annual US total crime rate for all types of crime time series data collected from the United States' Uniform Crime Reporting Statistics (UCRS). For factors data, nine factors were collected from different US government agencies and other related websites. The nine factors data collected for use in this study were unemployment rate (UR), immigration rate (IR), population rate (PR), consumer price index (CPI), gross domestic product (GDP), consumer sentiment index (CSI), poverty rate (PoR), inflation rate (InR), and tax revenue (TR). NCA was implemented to analyse these factors data to identify their significant relationships with crime rate data.

Data Processing and Preparation

The collected raw data sets of crime rate and factors were normalised using a feature scaling method in a scale range between 0 and 1. The normalised data sets were used to develop the proposed hybrid model in forecasting crime rates. Once the forecast of crime rate was done, the forecast output of normalised values was transformed back into actual raw values. Lastly, the transformed actual forecast output values were then used to calculate the quantitative measurement error. During the experiment, both crime rate and selected factors data sets were divided into training (data fitting) and test data (data prediction). In this study, 90 percent (50 samples from 1960 to 2009) of the collected data sets were used for training, while 10 percent (6 samples from 2010 to 2015) were used for testing in forecasting the crime rate.

Parameter Configuration

Before the experiment was conducted, the required input parameters for the proposed and compared crime models needed to be configured.

The parameters in GTB, namely the number of trees, learning rate, and individual size, were set to 100, 0.1, and 3, respectively. For the compared RF and ReliefF-RF crime models, the number of trees parameter was set to 100.

Performance Measurement

In this study, root mean square error (RMSE), mean absolute deviation (MAD), and mean absolute percentage error (MAPE) were used to measure and compare the performance of the developed crime model. Equations 1, 2 and 3 show the calculation of RMSE, MAD, and MAPE, respectively:

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(a_t - b_t\right)^2} \tag{1}$$

$$MAD = \sum_{t=1}^{n}\frac{\left|\left(a_t - b_t\right)^2\right|}{n} \tag{2}$$

$$MAPE = \sum_{t=1}^{n}\left|\frac{a_t - b_t}{z_t}\right| \times \frac{100}{n} \tag{3}$$

where,

n = The total number of test data used during the testing process,

a_t = The actual crime rate raw value,

b_t = The forecast crime rate raw value.

RESULTS AND DISCUSSION

In this study, NCA identified and selected significant factors that influenced the crime rates, and the selected factors were used to model the crime rates using GTB. The quantitative error measurement results produced in the NCA-GTB model were obtained and analysed. Then, the performance of the NCA-GTB model was compared with the RF and ReliefF-RF models. In the first phase, i.e., ranking of factors, the importance weight values for NCA, RF, and ReliefF were obtained and presented in Tables 2, 3, and 4, respectively.

Table 2

Ranking of Factors for NCA

Rank	Factor	Importance Weight Value
1	PR	1.7443
2	GDP	1.5054×10^{-5}
3	CPI	4.9268×10^{-6}
4	CSI	3.5816×10^{-6}
5	InR	2.9296×10^{-6}
6	UR	2.7405×10^{-6}
7	IR	2.6054×10^{-6}
8	TR	1.5104×10^{-6}
9	PoR	1.1918×10^{-7}

Table 3

Ranking of Factors for RF

Rank	Factor	Importance Weight Value
1	CPI	1.5721
2	GDP	1.5584
3	IR	0.7335
4	TR	0.3959
5	UR	0.3771
6	PR	0.2693
7	InR	0.2413
8	CSI	0.0000
9	PoR	0.0000

Table 4

Ranking of Factors for ReliefF

Rank	Factor	Importance Weight Value
1	PR	0.0816
2	PoR	0.0697
3	CPI	0.0369
4	IR	0.0298
5	UR	0.0178
6	InR	0.0169
7	TR	0.0099
8	GDP	0.0058
9	CSI	0.0010

From the assigned importance weight values based on NCA in Table 2, it was observed that the values for factors GDP, CPI, CSI, InR, UR, IR, TR, and PoR were near to zero. Despite these cases, this result did not imply that the factors were considered insignificant since the regularisation parameter influenced the overall NCA weighting calculation. In NCA, the regularisation parameter was calculated as $1/N$, where N is the total number of data samples (Yang et al., 2012). Therefore, the higher the N value, the smaller the regularisation parameter value. In contrast, the observed high importance weight value in factor PR indicated that NCA identified the factor as more significant than others.

Overall, NCA and ReliefF identified factor PR as the most significant since the importance weight value was the highest compared to other factors. Meanwhile, factor PR in RF was ranked sixth. For factor GDP, NCA and RF identified it as the second most important, while in ReliefF, GDP was ranked eighth. Factor CPI was classified as the most important and ranked first in RF, while in NCA and ReliefF, CPI was ranked third. Factor IR was ranked seventh, third, and fourth in NCA, RF, and ReliefF, respectively. Factor UR was ranked fifth in RF and ReliefF, while in NCA, it was ranked sixth.

As for factor TR, it was ranked eighth, fourth, and seventh in NCA, RF, and ReliefF, respectively. Factor InR was ranked fifth in NCA, sixth in ReliefF, and seventh in RF. For factor PoR, ReliefF identified it as the second most important, while in NCA, it was identified as the most insignificant with the weakest relationship with crime rate. Additionally, there was no significant relationship between PoR and crime rate identified by RF as the importance weight value was 0. Therefore, factor PoR was eliminated in RF.

Lastly, NCA identified factor CSI as the fourth most important, while in ReliefF, it was identified as the most insignificant with the weakest relationship. The importance weight value of CSI in RF was 0. This observation indicated that there was no relationship between CSI and crime rate in RF, and thus, it was eliminated. From the obtained ranked factors for each feature selection technique, the factor test data set was then constructed and presented in Table 5.

Table 5

Constructed Factor Test Data Set for Each Feature Selection Technique

Feature Selection	Test Data Set	Factor(s)
NCA	1	PR
	2	PR, GDP
	3	PR, GDP, CPI
	4	PR, GDP, CPI, CSI
	5	PR, GDP, CPI, CSI, InR
	6	PR, GDP, CPI, CSI, InR, UR
	7	PR, GDP, CPI, CSI, InR, UR, IR
	8	PR, GDP, CPI, CSI, InR, UR, IR, TR
RF	1	CPI
	2	CPI, GDP
	3	CPI, GDP, IR
	4	CPI, GDP, IR, TR
	5	CPI, GDP, IR, TR, UR
	6	CPI, GDP, IR, TR, UR, PR
	7	CPI, GDP, IR, TR, UR, PR, InR
ReliefF	1	PR
	2	PR, PoR
	3	PR, PoR, CPI
	4	PR, PoR, CPI, IR
	5	PR, PoR, CPI, IR, UR
	6	PR, PoR, CPI, IR, UR, InR
	7	PR, PoR, CPI, IR, UR, InR, TR
	8	PR, PoR, CPI, IR, UR, InR, TR, GDP

The constructed factor test data set was used as input data in developing the NCA-GTB, RF, and ReliefF-RF models to forecast crime rate. The forecast crime rate values for NCA-GTB, RF, and ReliefF-RF models were then calculated using quantitative error measurement analysis and presented in Tables 6, 7, and 8, respectively.

Table 6

Quantitative Error Measurement Result for NCA-GTB

Test Data Set	Quantitative Error Measurement		
	RMSE	MAD	MAPE
1	1076.7137	1061.7451	33.6159
2	1114.4154	1074.3749	33.8341
3	1223.6001	1198.7157	38.0975

(continued)

Test Data Set	Quantitative Error Measurement		
	RMSE	MAD	MAPE
4	459.0995	305.0448	10.3050
5	**296.6747**	**228.6354**	**7.5064**
6	391.4483	275.5394	9.1572
7	301.7880	250.5024	8.2938
8	299.6174	236.5913	7.8444

Table 7

Quantitative Error Measurement Result for RF

Test Data Set	Quantitative Error Measurement		
	RMSE	MAD	MAPE
1	699.9425	676.6903	21.9697
2	1081.9827	1055.7878	33.6043
3	842.2704	827.8450	26.7553
4	791.5390	771.8187	25.0042
5	749.5713	736.4209	23.7882
6	**620.1495**	**594.0758**	**19.3238**
7	646.2859	626.1496	20.3136

Table 8

Quantitative Error Measurement Result for ReliefF-RF

Test Data Set	Quantitative Error Measurement		
	RMSE	MAD	MAPE
1	674.2355	661.6775	20.9344
2	**633.8008**	**580.5910**	**18.1141**
3	719.0408	696.4265	22.6007
4	666.4235	642.3574	20.8691
5	698.3077	676.0077	21.9404
6	684.7024	661.7300	21.4852
7	712.5178	693.7124	22.4766
8	752.1852	743.3838	23.9513

Table 6 shows that the error increased from test data sets 8 to 6 when factors PoR, TR and IR were removed. From test data sets 6 to 5, the error decreased when an additional factor UR was excluded. From test data sets 5 to 3, an increase in error was observed when additional factors CSI and InR were eliminated. This observation revealed that

factors PR, GDP, CPI, CSI, and InR (test data set 5) greatly influenced the crime rate based on NCA-GTB as these factors produced the smallest error compared to other test data sets.

Based on the observed result in Table 7, from test data sets 7 to 6, the error declined when factor InR was excluded. Then, from test data sets 6 to 2, a reduction in error pattern was observed when factors IR, TR, UR, and PR were eliminated. Lastly, from test data sets 2 to 1, the error declined sharply when factor GDP was excluded. From the analysis, factors CPI, GDP, IR, TR, UR, and PR had a significant influence on the crime rate based on RF as the observed error in test data set 6 was the smallest compared to others.

In Table 8, from test data sets 8 to 6, the error declined when factors TR and GDP were eliminated. From test data sets 6 to 5, the error increased when the InR factor was excluded. However, from test data sets 5 to 4, the error declined when factor UR was eliminated. From test data sets 4 to 3, when factor IR was excluded, the error increased. Then, from test data sets 3 to 2, the error declined sharply when factor CPI was eliminated. Lastly, the error increased from test data set 2 to 1 as factor PoR was excluded. From the analysis, factors PR and PoR significantly influenced the crime rate based on ReliefF-RF as the observed error in test data set 2 was the smallest compared to others. The best results for each crime model were selected and compared based on the analysed quantitative error measurement. A comparison of the proposed hybrid NCA-GTB crime model with other models is presented in Table 9.

Table 9

Comparison of Proposed NCA-GTB Model with Other Models

Model	Quantitative Error Measurement		
	RMSE	MAD	MAPE
NCA-GTB	**296.6747**	**228.6354**	**7.5064**
GTB	500.6431	433.6435	14.3285
RF	620.1495	594.0758	19.3238
Relief-RF	633.8008	580.5910	18.1141
RF Using All Factors	625.8020	607.2424	19.6909

Based on the performance comparison in Table 9, the proposed hybrid NCA-GTB model outperformed the RF and ReliefF-RF models as

shown by its smallest RMSE, MAD, and MAPE values. According to the quantitative error values, the NCA-GTB model was able to predict the crime rate better with 50 percent higher accuracy than the RF and ReliefF-RF models.

A comparison between NCA-GTB and GTB showed a significant improvement as the prediction accuracy improved by up to 45 percent. This result revealed that the evaluation and selection of significant factors were able to eliminate irrelevant factors that could negatively impact the GTB model in forecasting crime rate. The same scenario was also observed when comparing RF and ReliefF-RF with RF using all factors, whereby the forecasting error could be reduced. Even though the error reduction was minimal, with up to an 8 percent decrease, it was still beneficial as it could improve RF's overall forecasting accuracy.

Based on the analysis, the proposed hybrid NCA-GTB model was able to significantly identify significant factors that later improved GTB in forecasting the US crime rate data. Therefore, the proposed NCA-GTB model was more suitable and appropriate for forecasting the US crime rate data with limited samples than other models. In conclusion, the hypothesis made in this study has been achieved.

CONCLUSION

The analysis of crime data helps in understanding future crime patterns through forecasting. There are two types of crime forecasting models proposed by different researchers—statistical and artificial intelligence models. In the last decade, researchers have shifted their research interest from statistical models to artificial intelligence-based models in crime forecasting. Among the introduced artificial intelligence techniques, Gradient Tree Boosting (GTB) is a novel technique in crime forecasting. Inspired by this, GTB was selected as the base model in developing the proposed crime forecasting model.

Most criminologists and researchers have been shown to study and observe the effect of several factors on criminal activities. These studies provided relevant insight into possible future crime trends based on recent issues. A study on the influence of several factors in crime analysis is highly beneficial because crime occurrence

patterns are not heavily dependent on past crime trends. Instead, various factors, such as social mistreatment, population densities, and economic disadvantages, affect crime patterns. Therefore, this study proposed an appropriate factor selection analysis by implementing a feature selection technique called NCA to identify the significant factors that influence crime rate.

From the results of the experiment, the performance of the proposed hybrid NCA-GTB crime model was not affected by the assumption that the forecasting performance could be improved if the factors were reduced significantly. Instead, it was affected by the combinations of several factors based on the constructed test data set. From these arguments, it is recommended to properly identify and analyse the significant relationship between factors and crime rate data. This is an alternative to blindly performing an analysis to reduce factors as much as possible on the basis that this can improve forecasting performances. Overall, the proposed hybrid NCA-GTB crime model outperformed other existing models in terms of quantitative error measurement. This case study found that the proposed NCA-GTB model is suitable for forecasting crime rates using a small data set. Applying the factor selection analysis using NCA in identifying the significant factors yielded promising results.

Although the proposed hybrid model performed well compared to existing models, GTB and NCA share one limitation. The limitation is that both GTB and NCA are sensitive to input parameters. Inappropriate parameter configuration in GTB and NCA leads to overfitting or underfitting problems. Thus, for future research, optimising both NCA and GTB input parameters is suggested to further improve the proposed hybrid model. Furthermore, a hybridisation of NCA with other AI models, such as Random Forest and Artificial Neural Network is recommended for future studies. This approach is to observe and validate the capability of NCA in improving different AI models in forecasting crime rate.

ACKNOWLEDGMENT

This work was supported by the UTM Encouragement Research Grant (Q.J130000.2651.18J46) for School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia (UTM).

REFERENCES

Aldehim, G., & Wang, W. (2017). Determining appropriate approaches for using data in feature selection. *International Journal of Machine Learning and Cybernetics*, *8*, 915–928.

Alwee, R. (2014). *Swarm optimized support vector regression with autoregressive integrated moving average for modeling of crime rate*. [Doctoral thesis Universiti Teknologi Malaysia].

Anser, M. K., Yousaf, Z., Nassani, A. A., Alotaibi, S. M., Kabbani, A., & Zaman, K. (2020). Dynamic linkages between poverty, inequality, crime, and social expenditures in a panel of 16 countries: Two-step GMM estimates. *Journal of Economic Structures*, *9*(1), 1–25.

Babakura, A., Sulaiman, M. N., & Yusuf, M. A. (2014, August). Improved method of classification algorithms for crime prediction. In *2014 IEEE International Symposium on Biometrics and Security Technologies (ISBAST)* (pp. 250–255). IEEE.

Baliyan, A., Gaurav, K., & Mishra, S. K. (2015). A review of short term load forecasting using artificial neural network models. *Procedia Computer Science*, *48*, 121–125.

Blakeslee, D., Chaurey, R., Fishman, R., Malghan, D., & Malik, S. (2021). In the heat of the moment: Economic and non-economic drivers of the weather-crime relationship. *Journal of Economic Behavior & Organization*, *192*, 832–856.

Bogomolov, A., Lepri, B., Staiano, J., Oliver, N., Pianesi, F., & Pentland, A. (2014, November). Once upon a crime: Towards crime prediction from demographics and mobile data. In *Proceedings of 16th International Conference on Multimodal Interaction* (pp. 427–434).

Breiman, L. (1997). Arcing the edge. *Technical Report 486*, Statistics Department, University of California, Berkeley.

Brown, E., & Males, M. (2011). Does age or poverty level best predict criminal arrest and homicide rates? A preliminary investigation. *Justice Policy Journal*, *8*, 1–30.

Castelli, M., Sormani, R., Trujillo, L., & Popovič, A. (2017). Predicting per capita violent crimes in urban areas: An artificial intelligence approach. *Journal of Ambient Intelligence and Humanized Computing*, *8*, 29–36.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, *40*, 16–28.

Chen, T. (2022, April). Multivariate analysis on determining the main influencing factors of police violence in the United States. In *International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS)* (Vol. 12163, pp. 16–21). SPIE.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Ghazvini, A., Nazri, M. Z. B. A., Abdullah, S. N. H. S., Junoh, M. N., & Kasim, Z. A. (2015, November). Biography commercial serial crime analysis using enhanced dynamic neural network. In *IEEE 7th International Conference of Soft Computing and Pattern Recognition (SoCPaR)* (pp. 334–339). IEEE.

Gorr, W., Olligschlaeger, A., & Thompson, Y. (2003). Short-term forecasting of crime. *International Journal of Forecasting*, 19, 579–594.

Guo, R., Fu, D., & Sollazzo, G. (2022). An ensemble learning model for asphalt pavement performance prediction based on gradient boosting decision tree. *International Journal of Pavement Engineering*, *23*(10), 3633–3646.

Han, M., & Wang, Y. (2009). Analysis and odelling of multivariate chaotic time series based on neural network. *Expert Systems with Applications*, *36*, 1280–1290.

Hanslmaier, M., Kemme, S., Stoll, K., & Baier, D. (2015). Forecasting crime in Germany in times of demographic change. *European Journal on Criminal Policy and Research*, *21*, 591–610.

Hipp, J. R., & Yates, D. K. (2011). Ghettos, thresholds, and crime: Does concentrated poverty really have an accelerating increasing effect on crime? *Criminology*, *49*, 955–990.

Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., & Khanahmadliravi, N. (2013). An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology*, *6*, 4219–4225.

Kim, Y. A. (2018). Examining the relationship between the structural characteristics of place and crime by imputing census block data in street segments: Is the pain worth the gain? *Journal of Quantitative Criminology*, *34*(1), 67–110.

Li, Q., Qiao, F., Mao, A., & McCreight, C. (2019). Characterizing the importance of criminal factors affecting bus ridership using random forest ensemble algorithm. *Transportation Research Record*, *2673*(4), 864–876.

Li, S.-T., Kuo, S.-C., & Tsai, F.-C. (2010). An intelligent decision-support model using FSOM and rule extraction for crime prevention. *Expert Systems with Applications*, *37*, 7108–7119.

Liu, X., Shen, C., Wang, W., & Guan, X. (2019). CoEvil: A coevolutionary model for crime inference based on fuzzy rough feature selection. *IEEE Transactions on Fuzzy Systems*, *28*, 806–817.

Liu, Y., Gu, Y., Nguyen, J. C., Li, H., Zhang, J., Gao, Y., & Huang, Y. (2017). Symptom severity classification with gradient tree boosting. *Journal of Biomedical Informatics*, *75*, S105–S111.

Malan, N. S., & Sharma, S. (2022). Motor imagery EEG spectral-spatial feature optimization using dual-tree complex wavelet and neighbourhood component analysis. *IRBM*, *43*(3), 198–209.

McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ)*, *2*(1), 1–12.

Mills, C. E., Freilich, J. D., & Chermak, S. M. (2017). Extreme hatred: Revisiting the hate crime and terrorism relationship to determine whether they are "close cousins" or "distant relatives". *Crime & Delinquency*, *63*(10), 1191–1223.

Nguyen, T. T., Hatua, A., & Sung, A. H. (2017). Building a learning machine classifier with inadequate data for crime prediction. *Journal of Advances in Information Technology*, *8*(2).

Osborn, D. (2018). An investigation into quarterly crime and its relationship to the economy. In *Illicit activity* (pp. 75–101). Routledge.

Pedregosa, F., Gaël, V., Alexandre, G., Vincent, M., Bertrand, T., Olivier, G., Mathieu, B., Peter, P., Ron, W., Vincent, D., Jake, V., Alexandre, P., David, C., Matthieu, B., Matthieu, P., & Édouard, D. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Quick, M., Li, G., & Brunton-Smith, I. (2018). Crime-general and crime-specific spatial patterns: A multivariate spatial analysis of four crime types at the small-area scale. *Journal of Criminal Justice*, *58*, 22–32.

Ranson, M. (2014). Crime, weather, and climate change. *Journal of Environmental Economics and Management*, *67*, 274–302.

Rather, A. M., Sastry, V., & Agarwal, A. (2017). Stock market prediction and portfolio selection models: A survey. *OPSEARCH*, *54*, 558–579.

Sainin, M. S., Alfred, R., & Ahmad, F. (2021). Ensemble meta classifier with sampling and feature selection for data with imbalance multiclass problem. *Journal of Information and Communication Technology*, *20*(2), 103–133.

Shah Habibullah, M., & Baharom, A. H. (2009). Crime and economic conditions in Malaysia. *International Journal of Social Economics*, *36*, 1071–1081.

Shi, T. (2020, July). A method of predicting crime of theft based on bagging ensemble feature selection. In *2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS)* (pp. 140–143). IEEE.

Shrivastav, A. K., & Ekata, D. (2012). Applicability of soft computing technique for crime forecasting: A preliminary investigation. *International Journal of Computer Science and Engineering Technology*, *9*(9), 415–421.

Soundarya, V., Kanimozhi, U., & Manjula, D. (2017). Recommendation system for criminal behavioral analysis on social network using genetic weighted K-means clustering. *Journal of Computers*, *12*, 212–220.

Stansfield, R., Williams, K. R., & Parker, K. F. (2017). Economic disadvantage and homicide: Estimating temporal trends in adolescence and adulthood. *Homicide Studies*, *21*, 59–81.

Sun, R., Wang, G., Zhang, W., Hsu, L. T., & Ochieng, W. Y. (2020). A gradient boosting decision tree based GPS signal reception classification algorithm. *Applied Soft Computing*, *86,* 105942.

Tonkal, Ö., Polat, H., Başaran, E., Cömert, Z., & Kocaoğlu, R. (2021). Machine learning approach equipped with neighbourhood component analysis for DDoS attack detection in software-defined networking. *Electronics*, *10*(11), 1227.

Vineeth, K. S., Pandey, A., & Pradhan, T. (2016, May). A novel approach for intelligent crime pattern discovery and prediction. In *2016 IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)* (pp. 531–538).

Wang, X., & Hu, S. (2022, January). Analysis of the relationship between unemployment and crime rate in China. In *2021 International Conference on Social Development and Media Communication (SDMC)* (pp. 665–670). Atlantis Press.

Yang, W., Wang, K., & Zuo, W. (2012). Neighborhood component feature selection for high-dimensional data. *Journal of Computers*, *7*(1), 161–168.

Yearwood, D. L., & Koinis, G. (2011). Revisiting property crime and economic conditions: An exploratory study to identify predictive indicators beyond unemployment rates. *The Social Science Journal*, *48*, 145–158.

Zhang, Y., Ren, X., & Zhang, J. (2019, July). Intrusion detection method based on information gain and ReliefF feature selection. In *IEEE International Joint Conference on Neural Networks (IJCNN)* (pp. 1–5). IEEE.

Zheng, Y., Liu, Q., Chen, E., Zhao, J. L., He, L., & Lv, G. (2015, May). Convolutional nonlinear neighbourhood components analysis for time series classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 534–546).