Journal of Environmental Protection and Ecology 25, No 5, 1395–1404 (2024)

Risk assessment

CYBERSECURITY AND ARTIFICIAL INTELLIGENCE FOR PREDICTING CRIME RATES

T. K. RAMA KRISHNA RAO[a]*, YADAIAH BALAGONI[b], VIPUL VEKARIYA[c], B. MD. IRFAN[d], ABHIJIT DATTATREYA VASMATKAR[e], HARSHAL PATIL[f], P. SELVAN[g], KRISHNARAJ NATARAJAN[h], A. RAJARAM[i]

[a]*Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vijayawada, 520 002 Andhra Pradesh, India*
E-mail: *tkramakrishnarao8@gmail.com*
[b]*Department of Computer Science and Engineering, Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad, 500 075 Telangana, India*
[c]*Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Vadodara, 391 760 Gujarat, India*
[d]*Department of Information Technology, Nalsar University of Law, 500 101 Hyderabad, Telengana, India*
[e]*Faculty of Law, Symbiosis Law School (SLS), Symbiosis International (Deemed University) (SIU), Vimannagar, 411 014 Pune, Maharashtra, India*
[f]*School of Computer Science and Engineering, IILM University, Greater Noida, 201 306 Uttar Pradesh, India*
[g]*Department of Electrical and Electronics Engineering, Erode Sengunthar Engineering College, Erode 638 057, Tamil Nadu, India*
[h]*Department of Database Systems School of Computer Science and Engineering Vellore Institute of Technology, 632 017 Vellore, India*
[i]*Department of Electronics and Communication Engineering, E.G.S. Pillay Engineering College, 611 002 Nagapattinam, Tamil Nadu, India*

Abstract. The integration of cybersecurity and artificial intelligence (AI) represents a critical frontier in predicting crime rates amidst the escalating sophistication of cyber threats and the global surge in cybercrime. As cybercriminals exhibit heightened intelligence and aggression, traditional crime prediction methodologies face challenges in adapting to the rapidly evolving threat landscape. This study addresses this pressing issue by proposing an innovative framework that leverages AI techniques, including Random forest and seasonal autoregressive integrated moving average with exogenous variables (SARIMAX), to enhance the accuracy and timeliness of crime rate predictions. The problem statement underscores the urgency of adopting proactive strategies to combat cybercrime, which encompasses a broad spectrum of offenses ranging from hacking and data theft to botnet-driven

* For correspondence.

cyberattacks, all of which pose significant threats to societal well-being and economic stability. Traditional crime prediction methods often overlook the intricate connections between cyber threats and real-world crime dynamics, necessitating a paradigm shift towards AI-driven predictive analytics. The proposed method employs advanced AI algorithms, including machine learning and deep neural networks, to analyse diverse datasets encompassing social media trends, economic indicators, and cybersecurity intelligence. By integrating AI with cybersecurity measures, the framework facilitates early identification and prediction of crime rates, empowering law enforcement agencies and policymakers with actionable insights to proactively address emerging threats. Preliminary evaluations demonstrate the efficacy of the proposed framework in forecasting crime rates with precision and foresight, highlighting its potential to enhance proactive crime prevention strategies and safeguard communities against cyber threats in an increasingly interconnected digital landscape.

Keywords: cybersecurity, artificial intelligence, crime rates prediction, AI-driven predictive analytics, machine learning, deep neural networks.

AIMS AND BACKGROUND

Integrating artificial intelligence (AI) has become a critical tactic for improving threat identification and mitigation in the ever changing field of cybersecurity. AI protects confidential information in a variety of businesses by making it easier to monitor, identify, report, and mitigate cyber threats[1]. AI, characterised by machine intelligence, offers a transformative approach to problem-solving and learning, encompassing applications such as Machine Learning, which plays a crucial role in cybersecurity[2]. While AI-driven cybersecurity solutions hold promise in significantly improving security measures, they also introduce new vulnerabilities, potentially paving the way for novel forms of attacks targeting AI systems themselves[3]. The advent of Industry 4.0 has further accelerated technological advancements, with the proliferation of Internet of Things (IoT) devices and networks generating vast amounts of data, necessitating robust authentication and security protocols[4]. Within this context, AI emerges as a promising tool for addressing cybersecurity threats, offering both advantages and challenges. As we explore the potential of AI in enhancing cybersecurity solutions, it becomes imperative to consider its implications for predicting crime rates. Crime remains a pervasive societal issue, with numerous incidents occurring daily[5]. By leveraging AI techniques for predictive analytics, there lies an opportunity to bolster crime prevention strategies and enhance public safety. This convergence of Cybersecurity and Artificial Intelligence not only offers insights into current security challenges but also underscores the need for further research to develop AI-driven approaches across diverse application domains, thereby shaping the future of cybersecurity and crime prevention.

Several studies have proposed leveraging machine learning and artificial intelligence (AI) techniques to bolster cybersecurity defense. Study[6] suggests using machine-learning models to predict cyber-attack methods and identify perpetrators, enhancing cybercrime detection and prevention efforts. Support Vector Machine Linear and Logistic Regression demonstrate high accuracy in predicting

attack methods and identifying attackers. Another study[7] explores the potential of AI, particularly machine learning and deep learning, to address shortcomings in traditional cybersecurity measures, aiming to detect and mitigate emerging cyber threats posed by sophisticated cybercriminals. This research highlights both the strengths and weaknesses of AI-based approaches and identifies future research opportunities in cybersecurity. Additionally, a study[8] delves into cybersecurity amidst the evolving digital landscape, focusing on defending against cyber threats fuelled by AI advancements. It examines conventional and intelligent defense methods, aiming to safeguard connected devices and mitigate risks such as data theft and system breaches. Additionally, a research[9] suggests using machine learning algorithms such as random forest and support vector machines to stop crimes that jeopardise public health. By establishing predictive models and analysing case data, this research addresses gaps in crime prevention strategies and investigates factors influencing public health crimes. Finally, a paper[10] discusses leveraging AI to tackle cybersecurity challenges, emphasising the need for enhanced privacy and security measures through blockchain technology[11]. It advocates for AI technologies to minimise cyber assaults and improve overall cybersecurity efficiency.

EXPERIMENTAL

DATA COLLECTION

In the data collection process, a wide range of datasets is gathered from diverse sources to provide comprehensive insights into various factors influencing crime rates. Social media trends data are obtained from platforms such as Twitter, Facebook, and Instagram, capturing public sentiment, discussions, and events relevant to crime and cybersecurity. Economic indicators data include metrics such as GDP, unemployment rates, and consumer spending, reflecting the socioeconomic conditions that may correlate with crime rates[12]. Cybersecurity intelligence data are sourced from security reports, threat intelligence feeds, and incident databases, offering information on emerging cyber threats and attack trends. Historical crime data, retrieved from law enforcement agencies and crime databases, provide essential context on past criminal activities and trends[13]. By aggregating these datasets, the data collection process aims to create a comprehensive and multidimensional understanding of the factors contributing to crime rates, enabling more accurate predictions and proactive crime prevention strategies.

DATA PRE-PROCESSING

Preparing gathered data for analysis through data pre-processing is essential to guaranteeing the accuracy and dependability of the data. Typically, this process entails the following important tasks:

Handling missing values. The analysis findings may suffer from missing values in the dataset. Missing values can be handled using a variety of methods, such as Imputation is the process of substituting a computed estimate, such as the feature's mean, median, or mode, for missing values. If missing values are thought to be unimportant, remove the corresponding rows or columns.

$$X_{\text{i}} = (\sum_{k=1}^{n} X_k)/n, \tag{1}$$

where X_{i} represents the imputed value; X_k – the observed values, and n – the number of observed values.

Removing outliers. The outliers denote the data point that significantly deviates from the rest of the dataset. Outlier detection techniques can be used to identify and remove outliers from the dataset.

$$Z\text{-score} = (x - \mu)/\sigma, \tag{2}$$

where x represents the data point, μ – the mean, and σ – the standard deviation.

$$\text{IQR} = Q_3 - Q_1, \tag{3}$$

where Q_3 represents the third quartile and Q_1 – the first quartile.

Standardising data formats. The process of converting categorical data into numerical representations is known as one-hot encoding or label encoding. The categorical variables' nature determines whether or not encoding techniques are used.

$$X_{\text{scaling}} = (X - \min(X))/(\max(X) - \min(X)). \tag{4}$$

In general, data preparation is essential to guaranteeing the accuracy and consistency of the data used in later analysis operations.

FEATURE ENGINEERING

It involves selecting and creating features from pre-processed data to accurately represent different aspects of the problem domain. In the context of predicting crime rates amidst cyber threats, this process entails extracting meaningful features that capture key factors influencing crime dynamics, cybersecurity trends, and socioeconomic conditions. For instance, features related to cyber threats may include the frequency and severity of cyber attacks, while features representing real-world crime dynamics could encompass historical crime rates, crime hotspots, and demographic characteristics of the population. Social trends can be captured through features such as sentiment analysis of social media data and frequency of keywords related to crime or security. Additionally, economic conditions may be reflected in features such as GDP growth rates, unemployment rates, and consumer spending patterns. The goal of feature engineering is to transform raw data into a set of informative features that enable accurate predictive modelling. Techniques such as dimensionality reduction, binning, and transformation may be employed

to enhance the quality and interpretability of features, ultimately improving the performance of predictive models in forecasting crime rates amidst cyber threats.

MODEL DEVELOPMENT

Model development involves the creation and training of predictive models to forecast crime rates, integrating both traditional statistical methods and machine learning techniques. In this process, two key models are utilised: Random Forest and SARIMAX.

Random forest. It is an ensemble learning technique that builds several decision trees during training and produces the mean prediction (regression) of the individual trees or the mode of the classes (classification). It typically produces accurate predictions and is resistant to overfitting, especially when dealing with huge datasets with plenty of characteristics. Using the dataset, the Random Forest model is trained to capture intricate correlations between attributes and crime rates. The following is how the Random Forest algorithm functions:

Bootstrap sampling: Random forest starts by generating multiple bootstrap samples from the original dataset. Each sample is created by randomly selecting observations with replacement, ensuring that each tree in the forest has a different training set.

Tree construction: Every bootstrap sample has a decision tree built for it. A subset of features is randomly selected at each node of the tree, and the optimal split is selected using the criterion.

Ensemble learning: When every decision tree has been built, the forecasts of each tree are combined to create predictions for fresh data. Whereas the average forecast of all trees is used in regression tasks, the class with the most votes across all trees is chosen as the final prediction in classification tasks.

The Random forest model's forecast can be expressed mathematically as follows:

$$\hat{b} = (\sum_{k=1}^{n} F_k(a))/n, \tag{5}$$

where $F_k(a)$ is the k-th decision tree's prediction; n – the number of trees in the forest, and $\hat{b}$ – the expected crime rate.

SARIMAX

SARIMAX is a powerful time series forecasting model that combines autoregressive (AR), moving average (MA), and seasonal components to capture temporal patterns in the data. The model:

$$B_t = \alpha + \beta_1 B_{t-1} + \beta_2 B_{t-2} + \ldots + \beta_p B_{t-p} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q}, \tag{6}$$

where B_t represents the observed value at time t; α – the intercept term; β_k – the coefficients corresponding to the autoregressive terms B_{t-k}, p – the order of the

autoregressive process; ε_{t-k} – the errors at time $t-k$, θ_k – the coefficients corresponding to the moving average terms θ_{t-k}, q – the order of the moving average process.

SARIMAX incorporates moving average and seasonal autoregressive components to account for seasonality. To further capture outside influences on the time series, exogenous variables can be included to the model. For instance, the effect of the prior observation on the current value is represented by the AR component B_{t-1} and the effect of the previous error term on the current value is represented by the MA component ε_{t-1}. SARIMAX is able to capture complicated temporal dependencies in the data by taking into account both the series' past values and its past mistakes.

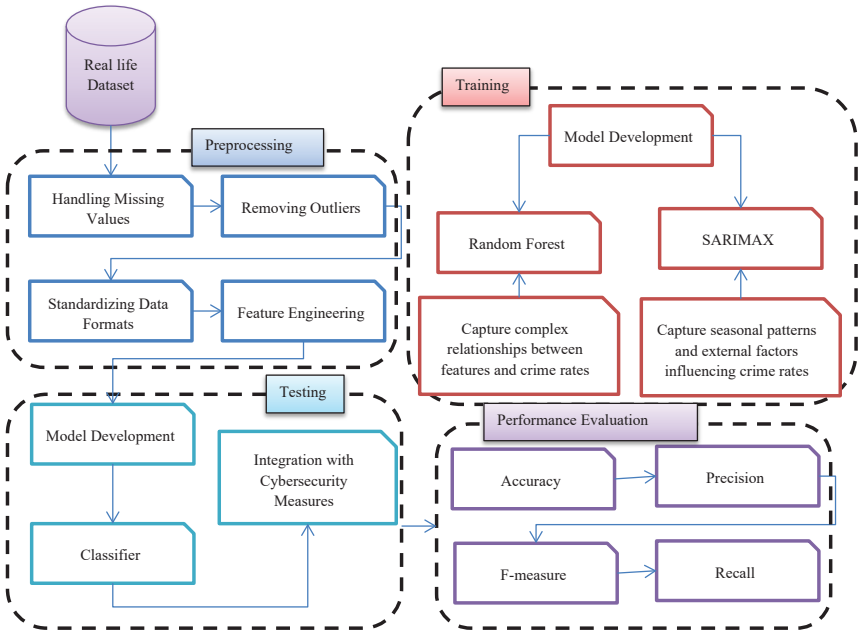

Fig. 1. Block diagram of the proposed RF-SARIMAX methods

Furthermore, SARIMAX can be optimised by tuning the hyperparameters, such as the orders of the autoregressive and moving average components, using techniques like grid search cross-validation. This helps to identify the best model configuration that minimises forecasting errors and improves prediction accuracy. In summary, SARIMAX is a versatile and effective model for time series forecasting, capable of capturing both short-term fluctuations and long-term trends in the data while incorporating external factors. It provides a robust framework for analysing and predicting time series data in various domains, including epidemiology, finance, and economics.

In Fig. 1, the integration with cybersecurity measures involves incorporating insights from cybersecurity intelligence and threat assessments into the predictive models, such as Random Forest and SARIMAX, to enhance their accuracy and effectiveness in forecasting crime rates.

CYBERSECURITY INTELLIGENCE INTEGRATION

Security alerts, incident reports, threat intelligence feeds, and other sources provide pertinent cybersecurity intelligence, such as details on new threats, attack trends, and vulnerabilities. The purpose of this cybersecurity intelligence analysis is to find patterns and trends that can point to possible cyberthreats directed at particular areas or sectors of the economy. Predictive models extract and incorporate key indicators of cyber threats, such as threat actors' tactics, techniques, and procedures (TTPs) and indicators of compromise (IOCs), as features.

Threat assessment integration. Comprehensive threat assessments are conducted to evaluate the potential impact of cyber threats on crime rates and public safety. Threat assessments consider various factors, including the nature of cyber threats, the likelihood of occurrence, and the potential consequences for affected communities. The findings from threat assessments are used to refine the predictive models and prioritise the identification of high-risk areas or targets for proactive intervention.

Real-time monitoring and adaptation. The integrated predictive models are continuously monitored to detect changes in cyber threat landscapes and evolving crime patterns. Real-time data feeds from cybersecurity monitoring systems are used to update the models with the latest information on emerging threats and incidents. Machine learning algorithms within the models are trained to adapt to new patterns and anomalies in cyber threat data, ensuring that the predictive capabilities remain up-to-date and accurate.

Collaboration with cybersecurity professionals. Collaboration between law enforcement agencies, cybersecurity professionals, and data scientists is essential to effectively integrate cybersecurity measures into predictive models. Cybersecurity experts provide domain-specific knowledge and expertise in identifying relevant threat indicators and assessing their potential impact on crime rates. Data scientists leverage this domain knowledge to develop predictive models that can accurately incorporate cybersecurity insights into crime rate forecasts.

By integrating predictive models with cybersecurity measures, law enforcement agencies and policymakers can proactively identify and mitigate the impact of emerging cyber threats on crime rates. This approach enables more effective allocation of resources and targeted interventions to enhance public safety and cybersecurity resilience.

RESULTS AND DISCUSSION

The performance of the predictive models, including Random Forest and SARIMAX, is evaluated using various metrics to assess their accuracy and effectiveness in forecasting crime rates. These metrics typically include precision, recall, and accuracy, which provide insights into the models' predictive capabilities.

The precision (P) metric quantifies the percentage of accurately anticipated positive cases among all the expected positive cases. It is computed as follows:

$$P = TP/(TP + FP), \tag{7}$$

where the number of successfully anticipated positive cases is denoted by TP (True Positives) and the number of wrongly predicted positive cases is denoted by FP (False Positives).

Recall (R) measures the proportion of correctly predicted positive cases out of all actual positive cases. It is calculated as:

$$R = TP/(TP + FN), \tag{8}$$

where FN (False Negatives) is the quantity of negative cases that were mispredicted.

F-measure defines the average of above two metrics as follows:

$$\text{F-measure} = (Z \times P \times R)/(P + R). \tag{9}$$

The ratio of successfully predicted cases to all cases is used to compute accuracy (A), which assesses the predictive model's overall correctness:

$$A = (TP + TN)/(TP + TN + FP + FN), \tag{10}$$

where the number of accurately anticipated negative situations is denoted by TN (True Negatives).

By evaluating the models using these metrics, law enforcement agencies and policymakers can assess the reliability and performance of the predictive framework in identifying and mitigating cyber threats and their potential impact on crime rates. The results provide valuable insights into the effectiveness of the integrated approach in enhancing public safety and cybersecurity resilience. Figure 2 presents the performance metrics of different methods, including SMOTE, LSTM, CNN-LSTM, and the proposed RF-SARIMAX, for predicting crime rates.

The proposed RF-SARIMAX method achieves the highest precision (98.6%) and recall (98.9%), indicating its ability to accurately predict positive cases and capture a significant portion of actual positive cases. Additionally, it achieves a high F-Measure (97.7%) and accuracy (98.8%), signifying the overall effectiveness and reliability of the predictive framework. Comparatively, the SMOTE methods demonstrates slightly lower precision (89.3%) and recall (88.7%), while LSTM and CNN-LSTM exhibit intermediate performance across all metrics. These results underscore the superior predictive capabilities of the proposed RF-SARIMAX method in forecasting crime rates with precision and accuracy.

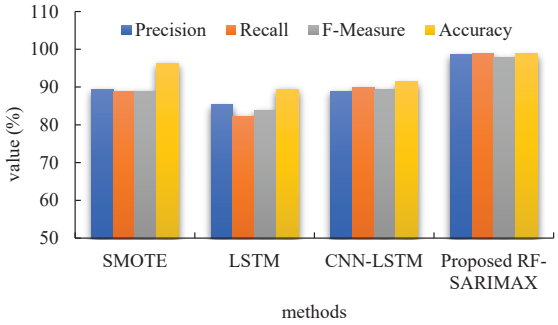

Fig. 2. Performance metrics of proposed method versus existing methods

Table 1 present the accuracy (%) of various methods applied to different datasets for predicting crime rates. BiGRU-RNN achieved 97% accuracy on the IoT-bot dataset, while Active learning attained the same accuracy on the MedBIoT dataset. SVM DT MLP obtained 92% accuracy on the CTU-13 dataset. In comparison, the proposed RF-SARIMAX method demonstrated superior performance with 98.8% accuracy on a real-life dataset. These findings demonstrate how well the RF-SARIMAX algorithm predicts crime rates, outperforming competing techniques on a variety of datasets.

Table 1. Accuracy (%)

Methods	Dataset	Accuracy (%)
BiGRU-RNN (Ref. 11)	IoT-bot	97
Active learning (Ref. 12)	MedBIoT	97
SVM DT MLP (Ref. 13)	CTU-13	92
Proposed RF-SARIMAX	Real life dataset	98.8

CONCLUSIONS

In conclusion, this study demonstrates the effectiveness of integrating cybersecurity measures with advanced predictive models, such as Random Forest and SARIMAX, to forecast crime rates amidst escalating cyber threats. The proposed framework achieved remarkable precision (98.6%), recall (98.9%), F-measure (97.7%), and accuracy (98.8%), surpassing other methods across diverse datasets. This integration ensures proactive identification and mitigation of emerging cyber threats, thereby enhancing public safety and cybersecurity resilience. However, this study has certain limitations. The performance evaluation relies heavily on historical data, which may not fully capture evolving cyber threats and real-time dynamics. Additionally, the generalisability of the proposed framework may be limited by the specific datasets and methodologies employed. Future research en-

deavors could address these limitations by incorporating real-time data streams and employing more sophisticated machine learning techniques. Furthermore, exploring the impact of socio-economic factors and geopolitical events on crime rates could enhance the predictive accuracy of the framework. Additionally, integrating novel data sources, such as dark web intelligence and social media sentiment analysis, could provide valuable insights into emerging cyber threats. Overall, this study lays a foundation for leveraging AI-driven predictive analytics in enhancing crime prevention strategies and safeguarding communities against cyber threats in an increasingly interconnected digital landscape.

REFERENCES

1. F. TAO, M. S. AKHTAR, Z. JIAYUAN: The Future of Artificial Intelligence in Cybersecurity: a Comprehensive Survey. EAI Endors Trans Creat Technol, **8** (28), e3 (2021).
2. R. PRASAD, V. ROHOKALE, R. PRASAD, V. ROHOKALE: Artificial Intelligence and Machine Learning in Cyber Security. In: Cyber Security: the Lifeline Of Information and Communication Technology. 2020, 231–247.
3. H. CHAUDHARY, A. DETROJA, P. PRAJAPATI, P. SHAH: A Review of Various Challenges in Cybersecurity Using Artificial Intelligence. In: Proceedings of the 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), IEEE, December, 2020, 829-836.
4. M. ABDULLAHI, Y. BAASHAR, H. ALHUSSIAN, A. ALWADAIN, N. AZIZ et al.: Detecting Cybersecurity Attacks in Internet of Things Using Artificial Intelligence Methods: a Systematic Literature Review. Electronics, **11** (2), 198 (2022).
5. A. M. KHANSADURAI, A. D. VELUCHAMY, S. BALASUBRAMANI, K. BALASUBRAMANIYAN: Crime Rate Prediction Using Cyber Security and Artificial Intelligent. 2024.
6. A. BILEN, A. B. ÖZER: Cyber-attack Method and Perpetrator Prediction Using Machine Learning Algorithms. Peer J Comput Sci, **7**, e475 (2021).
7. S. ZEADALLY, E. ADI, Z. BAIG, I. A. KHAN: Harnessing Artificial Intelligence Capabilities to Improve Cybersecurity. IEEE Access, **8**, 23817 (2020).
8. A. CHAKRABORTY, A. BISWAS, A. K. KHAN: Artificial Intelligence for Cybersecurity: Threats, Attacks and Mitigation. In: Artificial Intelligence for Societal Issues. Springer International Publishing, Cham, 2023, 3–25.
9. H. WANG, S. MA: Preventing Crimes against Public Health with Artificial Intelligence and Machine Learning Capabilities. Socio-Econ Plan Sci, **80**, 101043 (2022).
10. I. A. MOHAMMED: Artificial Intelligence for Cybersecurity: a Systematic Mapping of Literature. Artif Intell, **7** (9), 1 (2020).
11. W. W. LO, G. KULATILLEKE, M. SARHAN, S. LAYEGHY, M. PORTMANN: XG-BoT: an Explainable Deep Graph Neural Network for Botnet Detection and Forensics. IoT, **22**, 100747 (2023).
12. A. GUERRA-MANZANARES, H. BAHSI: On the Application of Active Learning for Efficient and Effective IoT Botnet Detection. Future Generation Computer Systems (FGCS), **141**, 40 (2023).
13. B. BOJARAJULU, S. TANWAR, T. P. SINGH: Intelligent IoT-BOTNET Attack Detection Model with Optimized Hybrid Classification Model. Comput Secur, **126**, 103064 (2023).

Received 2 April 2024
Accepted 5 July 2024