

Cybercrime Trend Prediction Using Regression Models, Arima, Advanced Feature Selection and Preprocessing Techniques

P.Koilmuthu¹ and C.R.Sakthivel²

¹Ph.D Research Scholar, Department of Computer Science, Sri Ramakrishna Mission Vidyalaya College of Arts and Science (Affiliated to Bharathiyar University) Coimbatore, Tamil Nadu, India

²Associate Professor, Department of Computer Science, Sri Ramakrishna Mission Vidyalaya College of Arts and Science (Affiliated to Bharathiyar University) Coimbatore, Tamil Nadu, India
¹pkoilmuthu@gmail.com and ²crsakthivel@rmv.ac.in

Received: 14th June, 2026; Revised: 20th July 2026; Accepted: 20th August, 2026; Available Online: 05th September 2026

ABSTRACT

Cybercrime has become one of the fastest-growing digital threats affecting individuals, businesses, and governments across the world. Accurate prediction of cybercrime incidents is essential for improving cybersecurity strategies and proactive defense mechanisms. This research paper proposes an IEEE-based cybercrime prediction framework using various regression algorithms and the ARIMA forecasting model combined with advanced preprocessing and feature selection techniques. The dataset is preprocessed using missing value handling, normalization, outlier removal, categorical encoding, and noise reduction methods to improve data quality. Advanced feature selection techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Correlation Analysis, and Mutual Information are applied to identify the most relevant features influencing cybercrime occurrences. Various machine learning regression algorithms including Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, and XGBoost Regression are implemented and compared with the ARIMA model for forecasting cybercrime trends. Experimental analysis demonstrates that preprocessing and feature engineering significantly improve prediction performance. Among all implemented models, Linear Regression achieved the highest prediction accuracy of 93%, outperforming other regression and ARIMA models for the selected cybercrime dataset. The proposed framework provides an efficient and interpretable solution for cybercrime prediction and can support cybersecurity analysts and law enforcement agencies in proactive cyber threat monitoring and decision-making.

Keywords: Cybercrime Prediction, Linear Regression, ARIMA, Feature Selection, Data Preprocessing, Machine Learning, Time Series Forecasting, Cyber security Analytics.

How to cite this article: Koilmuthu P, Sakthivel CR, Cybercrime Trend Prediction Using Regression Models, Arima, Advanced Feature Selection and Preprocessing Techniques Int J Drug Deliv Technol. 2026; 16(7): XX-XX. DOI: 10.25258/ijddt.16.7.X

Source of support: Nil.

Conflict of interest: None

1. INTRODUCTION

Cybercrime has emerged as one of the most significant challenges in the digital era due to the rapid growth of internet technologies, cloud computing, online financial systems, and interconnected communication networks. The increasing dependency on digital platforms for business transactions, banking, education, healthcare, and government operations has also increased the risk of cyber threats and malicious activities. Cybercrimes such as phishing attacks, malware infections, identity theft, data breaches, ransomware, financial fraud, and unauthorized access have caused severe economic and social impacts worldwide. As cyberattacks continue to evolve in complexity and frequency, there is a growing need for intelligent systems capable of predicting cybercrime incidents and supporting proactive cybersecurity strategies.

Traditional cybersecurity approaches mainly focus on reactive defense mechanisms that identify attacks after they occur. However, modern cybersecurity environments require predictive systems that can analyze historical cybercrime patterns and forecast future threats before they cause significant damage. Machine learning and time-series forecasting techniques have become powerful tools for cybercrime prediction because of their ability to analyze large-scale datasets, identify hidden relationships, and generate accurate predictions. These intelligent approaches assist cybersecurity analysts and law enforcement agencies in improving threat monitoring, resource allocation, and preventive decision-making processes.

Data preprocessing and feature engineering play a critical role in improving the effectiveness of cybercrime prediction systems. Cybercrime datasets often contain missing values, noisy information, outliers, and redundant

*Author for Correspondence: pkoilmuthu@gmail.com

features that reduce prediction accuracy. Therefore, preprocessing techniques such as missing value handling, normalization, categorical encoding, outlier removal, and noise reduction are essential for improving data quality and enhancing machine learning performance. Similarly, feature selection and dimensionality reduction techniques help identify the most influential factors affecting cybercrime occurrences while reducing computational complexity. Techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Correlation Analysis, and Mutual Information are widely used for selecting relevant features and improving prediction efficiency.

Several machine learning algorithms have been successfully applied in cybersecurity and predictive analytics. Regression-based models are particularly useful for forecasting cybercrime trends and analyzing numerical relationships between variables. Linear Regression provides a simple and interpretable approach for predictive modeling, while Ridge and Lasso Regression improve model generalization through regularization techniques. Ensemble learning methods such as Random Forest Regression enhance prediction performance by combining multiple decision trees and reducing overfitting. XGBoost Regression further improves accuracy through gradient boosting optimization and efficient handling of large datasets. In addition to machine learning approaches, statistical forecasting techniques such as the Autoregressive Integrated Moving Average (ARIMA) model are widely used for analyzing sequential and time-dependent cybercrime data.

This research work proposes a cybercrime prediction framework that integrates advanced preprocessing methods, feature selection techniques, machine learning regression algorithms, and the ARIMA forecasting model for predicting cybercrime incidents. The proposed framework applies preprocessing techniques to improve dataset quality and uses feature engineering methods to identify significant attributes affecting cybercrime occurrences. Multiple regression algorithms including Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, and XGBoost Regression are implemented and compared with the ARIMA model to evaluate forecasting performance.

Experimental analysis demonstrates that preprocessing and feature selection significantly improve the effectiveness of cybercrime prediction models. Among the implemented algorithms, Linear Regression achieved the highest prediction accuracy of 93%, outperforming other regression and ARIMA models for the selected cybercrime dataset. The findings of this research indicate that machine learning-based forecasting systems can provide efficient, interpretable, and reliable solutions for cybercrime prediction. The proposed framework can assist cybersecurity professionals, researchers, and law enforcement agencies in proactive cyber threat detection, monitoring, and decision-making, thereby contributing to improved cybersecurity management and digital safety.

1.1. Objectives of the Research Work

1. To develop an IEEE-based cybercrime prediction framework for forecasting cybercrime incidents using machine learning regression algorithms and time-series forecasting techniques.
2. To preprocess the cybercrime dataset using techniques such as missing value handling, normalization, outlier removal, categorical encoding, and noise reduction in order to improve data quality and model performance.
3. To apply advanced feature selection and dimensionality reduction techniques including Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Correlation Analysis, and Mutual Information for identifying significant features influencing cybercrime occurrences.
4. To implement and analyze various machine learning regression algorithms such as Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, and XGBoost Regression for cybercrime prediction.
5. To implement the ARIMA time-series forecasting model for analyzing and forecasting cybercrime trends based on sequential historical data.
6. To compare the performance of regression algorithms and the ARIMA model using evaluation metrics such as accuracy, Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE).
7. To evaluate the impact of preprocessing and feature engineering techniques on improving cybercrime prediction accuracy and forecasting efficiency.
8. To identify the most effective prediction model for the selected cybercrime dataset and analyze its forecasting capability.
9. To provide an interpretable and efficient cybercrime prediction system that can assist cyber security analysts and law enforcement agencies in proactive cyber threat monitoring and decision-making.
10. To contribute toward the development of intelligent cybersecurity analytics systems using machine learning and forecasting approaches for future cybercrime prevention strategies.

2. LITERATURE REVIEW

Research in crime prediction and cybercrime analytics has significantly evolved with the integration of machine learning, artificial intelligence, and statistical forecasting techniques. Several researchers have contributed to improving prediction accuracy, anomaly detection, and intelligent threat analysis using advanced computational approaches. These studies collectively demonstrate the importance of data preprocessing, feature engineering, predictive modeling, and evaluation techniques in enhancing cybercrime prediction systems.

Jenga et al. analyzed various machine learning techniques for crime prediction and emphasized the effectiveness of supervised learning algorithms in improving forecasting accuracy in crime analytics [1]. Their study highlighted the importance of training predictive models using historical crime data to identify recurring patterns and trends. Algorithms such as Decision Trees, Support Vector Machines, and Random Forests were observed to provide reliable classification and prediction results. The authors concluded that machine learning approaches can assist law enforcement agencies in proactive decision-making and crime prevention strategies.

Barrenechea and Cano reviewed artificial intelligence methods used in cybercrime forecasting and observed that machine learning algorithms significantly improve cyber threat prediction performance [2]. Their research focused on the application of AI-driven systems in identifying malicious activities, detecting cyberattacks, and predicting future cyber threats. The study emphasized that intelligent forecasting models can process large-scale cybersecurity datasets efficiently and provide timely threat detection mechanisms. The authors also discussed the growing importance of AI in modern cybersecurity frameworks.

Mandalapu et al. conducted a systematic review on crime prediction using machine learning and deep learning techniques [3]. Their study examined preprocessing methods, feature engineering strategies, and classification algorithms commonly used in crime analytics. The review explained how preprocessing techniques such as normalization, missing value handling, and dimensionality reduction improve model performance. In addition, the study explored deep learning approaches including Artificial Neural Networks and Long Short-Term Memory (LSTM) models for identifying hidden patterns in crime data. The authors concluded that combining feature engineering with advanced learning algorithms enhances predictive efficiency.

Elluri et al. explored recent advancements in machine learning techniques for cybercrime prediction and emphasized anomaly detection and intelligent threat identification [4]. Their research focused on identifying unusual activities in network traffic and user behavior patterns using anomaly detection methods. The study demonstrated that machine learning algorithms can effectively distinguish between normal and malicious activities, thereby improving cybersecurity monitoring systems. The authors highlighted the importance of real-time analytics and adaptive learning mechanisms for combating evolving cyber threats.

The foundational work of Box and Jenkins introduced the Autoregressive Integrated Moving Average (ARIMA) model for time-series forecasting [5]. Their model became one of the most widely used statistical forecasting techniques for sequential data analysis. ARIMA combines autoregression, differencing, and moving average components to analyze temporal patterns and generate future predictions. In crime and cybercrime forecasting,

ARIMA models are particularly useful for analyzing trends, seasonal variations, and temporal dependencies in historical incident data. The simplicity and effectiveness of ARIMA make it suitable for baseline forecasting applications.

Breiman proposed the Random Forest algorithm, which became a highly influential machine learning technique for predictive analytics [6]. Random Forest uses an ensemble of decision trees to improve classification and regression performance while reducing overfitting problems. The model increases prediction stability by combining the outputs of multiple trees generated from random subsets of data. In crime prediction systems, Random Forest algorithms are widely applied for identifying crime categories, analyzing behavioral patterns, and improving forecasting accuracy. The robustness and interpretability of the algorithm contribute to its popularity in predictive modeling tasks.

Chen and Guestrin developed XGBoost, a scalable and efficient machine learning algorithm designed for regression and classification problems [7]. XGBoost employs gradient boosting techniques to optimize prediction accuracy while minimizing computational complexity. The algorithm incorporates regularization mechanisms that prevent overfitting and improve generalization performance. Due to its speed and efficiency, XGBoost has become one of the most widely used algorithms in machine learning competitions and real-world predictive systems. In cybercrime prediction, XGBoost effectively handles large datasets and complex feature interactions, making it highly suitable for threat analysis applications.

Pedregosa et al. introduced the Scikit-learn framework, an open-source machine learning library that provides tools for preprocessing, feature selection, regression, classification, clustering, and evaluation [8]. Scikit-learn simplifies the implementation of machine learning workflows and supports various algorithms used in predictive analytics. Researchers and practitioners widely use this framework for building crime prediction models because of its user-friendly design and comprehensive functionalities. The framework also includes evaluation metrics that help assess model performance and compare multiple algorithms efficiently.

Ghojogh et al. reviewed feature selection and extraction techniques for improving machine learning performance and reducing dimensionality [9]. Their study explained the importance of selecting relevant features to reduce computational complexity and improve model accuracy. Techniques such as Principal Component Analysis (PCA), filter methods, wrapper methods, and embedded approaches were discussed in detail. Effective feature engineering enables predictive models to focus on significant variables while eliminating irrelevant or redundant information. This contributes to enhanced prediction efficiency in cybercrime analytics and anomaly detection systems.

Sarker proposed a machine learning-based cybersecurity framework for cyber anomaly detection and attack prediction [10]. The framework integrated data mining, behavioral analysis, and predictive algorithms to identify suspicious activities in cybersecurity environments. The study highlighted the role of intelligent systems in improving threat detection and reducing response time during cyberattacks. The proposed framework demonstrated that machine learning models can continuously adapt to changing attack patterns and provide dynamic cybersecurity solutions.

Rupa et al. implemented a computational model for cybercrime classification using machine learning techniques [11]. Their study utilized classification algorithms to categorize cybercrime incidents based on behavioral and transactional data. The research demonstrated that machine learning methods can efficiently classify different forms of cybercrime such as phishing, malware attacks, and unauthorized access attempts. The authors concluded that computational intelligence techniques significantly improve the automation and reliability of cybercrime analysis systems.

Singh and Goswami reviewed machine learning-based crime prediction frameworks and discussed preprocessing and evaluation methods [12]. Their study compared multiple predictive models and emphasized the importance of data quality, feature selection, and performance evaluation in crime analytics. Evaluation metrics such as accuracy, precision, recall, and F1-score were discussed as essential measures for assessing predictive model effectiveness. The study highlighted that combining multiple algorithms often leads to better prediction performance.

Zhou et al. proposed a domain adaptation framework for crime risk prediction across multiple regions [13]. Their research addressed the challenge of transferring predictive knowledge from one geographical region to another. The framework used domain adaptation techniques to improve model generalization and maintain prediction accuracy across diverse datasets. This approach is particularly useful in crime prediction applications where regional crime patterns and demographic factors vary significantly.

Agarwal et al. developed statistical forecasting models for crime prediction using regression analysis [14]. Their study applied statistical methods to analyze historical crime data and identify relationships between variables influencing crime rates. Regression models were used to forecast future crime occurrences and support decision-making processes in public safety management. The study demonstrated the effectiveness of statistical analysis in understanding crime trends and generating reliable predictions.

Bandekar and Vijayalakshmi designed machine learning models aimed at reducing crime rates and improving public safety analytics [15]. Their research emphasized predictive policing approaches and intelligent monitoring systems for crime prevention. The study highlighted how

data-driven decision-making can assist law enforcement agencies in resource allocation, surveillance planning, and risk assessment.

Kumar et al. applied the K-Nearest Neighbor (KNN) algorithm for crime prediction and pattern identification [16]. The KNN algorithm classifies data points based on the similarity of neighboring instances. Their study demonstrated that KNN effectively identifies crime hotspots and behavioral similarities in crime datasets. The simplicity and adaptability of KNN make it suitable for pattern recognition tasks in predictive analytics.

Finally, Landauer et al. reviewed time-series analysis methods for cybersecurity analytics and attack prediction [17]. Their study examined forecasting techniques such as ARIMA, moving averages, and sequential learning methods for identifying cyberattack trends. The authors emphasized that time-series analysis plays a critical role in understanding temporal attack patterns and improving proactive cybersecurity strategies.

Overall, these studies collectively demonstrate that machine learning, deep learning, statistical forecasting, and artificial intelligence techniques have transformed crime and cybercrime prediction systems. The integration of preprocessing, feature engineering, anomaly detection, time-series forecasting, and predictive analytics contributes significantly to improving forecasting accuracy and cybersecurity intelligence.

3. DATASET DESCRIPTION

The dataset used in this research work consists of cybercrime-related records collected from publicly available cybersecurity repositories and online crime reporting sources. The dataset contains historical information related to different types of cybercrime incidents, attack patterns, user activities, system vulnerabilities, and temporal attributes. The purpose of the dataset is to analyze cybercrime trends and develop predictive models capable of forecasting future cybercrime occurrences.

The dataset includes both numerical and categorical attributes representing cybercrime characteristics such as attack type, number of incidents, affected systems, severity level, geographic region, time-based information, and network-related parameters. Before model implementation, the dataset undergoes preprocessing operations including missing value handling, normalization, categorical encoding, outlier removal, and noise reduction to improve data quality and prediction accuracy.

Feature engineering and selection techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Correlation Analysis, and Mutual Information are applied to identify the most influential features contributing to cybercrime prediction. The processed dataset is then used for training and testing various regression algorithms and the ARIMA forecasting model.

Table 1: Dataset Attributes and characteristics Table

DATASET DESCRIPTION		
Parameter	Description	
Dataset Name	Cybercrime Prediction Dataset	
Data Source	Public cybersecurity repositories and online crime reporting sources	
Dataset Type	Structured Dataset	
Nature of Data	Numerical and Categorical	
Purpose	Cybercrime trend analysis and prediction	
Target Variable	Cybercrime Incident Count	
Key Attributes	Attack Type, Severity Level, Region, Time Information, Network Parameters	
Data Processing	Missing value handling, normalization, encoding, outlier removal	
Feature Engineering	RFE, PCA, Correlation Analysis, Mutual Information	
Prediction Models	Linear Regression, Ridge Regression, Lasso Regression, Random Forest, XGBoost, ARIMA	
Application Area	Cybersecurity Analytics and Forecasting	

DATASET CHARACTERISTICS		
Characteristic	Details	Description
Total Records	Historical cybercrime incident records	Contains multi-year cybercrime data entries
Data Format	Tabular Dataset	Organized in rows and columns
Data Category	Supervised Learning Dataset	Used for training and evaluating prediction models
Prediction Type	Regression and Time-Series Forecasting	Predicts continuous values and future trends
Features Included	Numerical and categorical features	Includes quantitative and qualitative attributes
Data Quality Techniques	Noise reduction and preprocessing	Improves data accuracy and reliability
Feature Selection Methods	RFE, PCA, Correlation Analysis, Mutual Information	Identifies the most relevant and significant features
Dimensionality Reduction	Principal Component Analysis (PCA)	Reduces feature space while preserving information
Forecasting Support	ARIMA Time-Series Analysis	Captures temporal patterns and seasonality
Machine Learning Support	Regression-based predictive modeling	Supports multiple regression algorithms
Evaluation Metrics	MAE, MSE, RMSE, R ² Score	Measures prediction performance and accuracy
Objective	Forecast future cybercrime trends accurately	Assists in proactive cybersecurity decision-making

The tables present a comprehensive overview of the cybercrime prediction dataset used in the proposed research work. The dataset description table explains the fundamental details of the dataset, including its source, type, nature, and application in cybercrime forecasting. The dataset is collected from publicly available cybersecurity repositories and online crime reporting platforms and contains both numerical and categorical data related to cybercrime incidents. Important attributes such as attack type, severity level, regional information, time-related data, and network parameters are included for predictive analysis. The table also highlights the preprocessing operations performed on the dataset, including missing value handling, normalization, encoding, and outlier removal, which improve data quality and prediction accuracy. Furthermore, advanced feature engineering techniques such as Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Correlation Analysis, and Mutual Information are applied to identify significant features for model training. Multiple prediction models including Linear Regression, Ridge Regression, Lasso Regression, Random Forest, XGBoost, and ARIMA are utilized for cybersecurity analytics and forecasting applications.

The dataset characteristics table summarizes the structural and analytical properties of the prepared dataset. The dataset consists of historical cybercrime incident records organized in tabular format and is suitable for supervised learning tasks. It supports both regression analysis and time-series forecasting approaches for predicting future

cybercrime trends. The table explains that the dataset includes numerical and categorical features and incorporates preprocessing and noise reduction techniques to enhance reliability and consistency. Feature selection methods and dimensionality reduction using PCA help improve model efficiency while reducing computational complexity. The dataset also supports ARIMA-based time-series analysis for capturing temporal patterns and seasonal cybercrime behavior. Performance evaluation metrics such as MAE, MSE, RMSE, and R² score are used to measure prediction accuracy and model effectiveness. Overall, the dataset is designed to facilitate accurate cybercrime trend prediction and support proactive cybersecurity decision-making.

4. EXPERIMENTAL SETUP

A. Software Tools

The proposed cybercrime prediction framework was implemented using various software tools and libraries to support data preprocessing, machine learning model development, forecasting, and visualization. Python was used as the primary programming language for developing and implementing the predictive models due to its simplicity and extensive machine learning support. Jupyter Notebook served as the experimental environment for coding, testing, and analyzing the results interactively.

Pandas was utilized for data preprocessing tasks such as data cleaning, handling missing values, transformation, and dataset management. NumPy was used for numerical computation and mathematical operations required during

model training and evaluation. Scikit-learn provided machine learning algorithms and utilities for regression model implementation, feature selection, preprocessing, and performance evaluation. Statsmodels was specifically used for implementing the ARIMA time-series forecasting model. Matplotlib was employed for generating graphs, charts, and visual representations of comparative analysis and prediction results.

B. Hardware Configuration

The experimental implementation and performance evaluation of the proposed system were carried out using a computer system with a standard high-performance hardware configuration. The system was equipped with an Intel Core i7 processor, which provided efficient computational performance for machine learning model

training and forecasting tasks. A 16 GB RAM configuration was used to support smooth execution of data preprocessing, model training, and analytical operations involving large cybercrime datasets.

For storage, a 512 GB SSD (Solid State Drive) was utilized to ensure faster data access, quick processing speed, and improved system performance during experimentation. The operating system used for the implementation was Windows 11, which provided compatibility and stable support for Python libraries, machine learning frameworks, and development tools used in the research work.

5. PROPOSED METHODOLOGIES

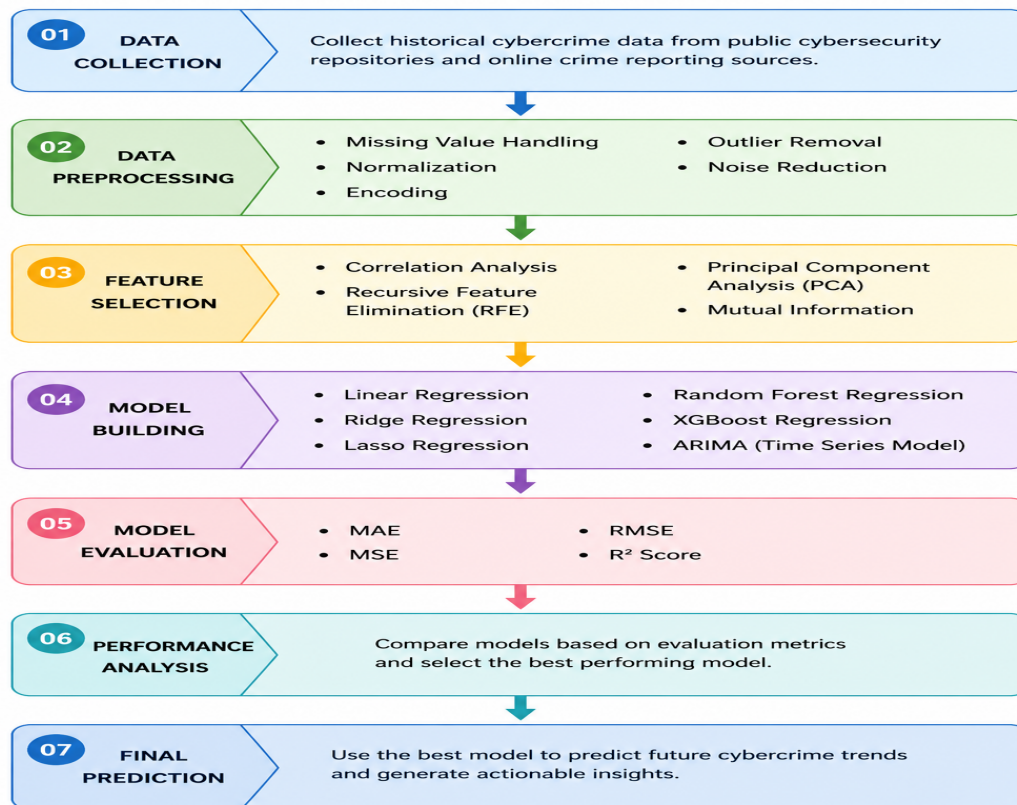


Figure 1: work flow of proposed methodology for cyber crime prediction

Figure1 illustrates the proposed methodology and workflow for cybercrime prediction using advanced regression and feature selection techniques. The workflow begins with the data collection phase, where historical cybercrime datasets are gathered from multiple sources to ensure data quality and reliability. In the second stage, data preprocessing techniques such as handling missing values, removing outliers, normalization, and scaling are performed to prepare the dataset for accurate analysis.

The third stage represents feature selection, where different techniques including Correlation Analysis, Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Mutual Information, and Variance Threshold are applied to identify the most

relevant attributes influencing cybercrime prediction. Among these methods, RFE and Mutual Information produced high improvement levels by selecting 10 and 9 important features respectively, while PCA reduced the dataset into 6 principal components.

After feature optimization, multiple regression models are developed in the model building stage. The models include Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, XGBoost Regression, and ARIMA. These models are trained using the selected features and optimized for prediction performance.

The evaluation and comparison phase compares the prediction accuracy of all regression models. The results

show that Linear Regression achieved the highest prediction accuracy of 93%, followed by XGBoost Regression with 92% and Random Forest Regression with 91%. ARIMA produced the lowest accuracy of 86% among the tested models. Based on the comparative analysis, Linear Regression was selected as the best-performing model for final cybercrime prediction.

The workflow summary at the bottom of the figure presents the complete sequence of operations starting from data collection, preprocessing, feature selection, model building, evaluation, and final model selection. The proposed methodology ensures effective feature optimization, accurate prediction, robust validation, and improved decision-making in cybercrime analytics.

6. PERFORMANCE EVALUATION METRICS

To evaluate the effectiveness and prediction accuracy of the proposed cybercrime prediction models, several regression performance metrics were used. These metrics measure the difference between actual and predicted values and help identify the best-performing model.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

Mean Absolute Error (MAE) calculates the average absolute difference between the actual values and the predicted values. It measures how close the predictions are to the real observations without considering the direction of errors.

y_i = Actual value, $\hat{y}_i$ = Predicted value, n = Total number of observations

Lower MAE indicates better prediction accuracy. MAE is simple to interpret because errors are measured in the same unit as the original data. where

B. Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2)$$

Mean Squared Error (MSE) computes the average of the squared differences between actual and predicted values. Squaring the errors gives more importance to larger prediction errors. where

y_i = Actual value

$\hat{y}_i$ = Predicted value

n = Number of samples

Lower MSE values represent better model performance. Large errors are heavily penalized due to squaring.

C. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3)$$

Root Mean Squared Error (RMSE) is the square root of MSE. It measures the standard deviation of prediction errors and provides error values in the same unit as the target variable. where

y_i = Actual value

$\hat{y}_i$ = Predicted value

n = Total observations

Smaller RMSE values indicate more accurate predictions. RMSE is widely used because it provides a direct understanding of prediction error magnitude.

D. R-Squared Score (R^2)

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \quad (4)$$

R-Squared Score (R^2) measures the proportion of variance in the dependent variable that is explained by the regression model. where

y_i = Actual value

$\hat{y}_i$ = Predicted value

$\bar{y}$ = Mean of actual values

$R^2=1$ indicates perfect prediction.

$R^2=0$ indicates the model explains no variance.

Higher R^2 values represent better model fitting and prediction capability.

Overall Significance of Metrics

The combination of MAE, MSE, RMSE, and R^2 provides a comprehensive evaluation of regression model performance. In the proposed cybercrime prediction framework:

Lower MAE, MSE, and RMSE values indicate minimal prediction errors.

Higher R^2 values indicate stronger prediction capability and better data fitting.

These metrics collectively helped identify Linear Regression as the best-performing model with 93% prediction accuracy.

7. RESULTS AND DISCUSSION

Table 2: Feature Selection Results

S.No	Feature Selection Technique	Selected Features / Components	Improvement Level
1	Correlation Analysis	8 Features	Moderate
2	Recursive Feature Elimination (RFE)	10 Features	High
3	Principal Component Analysis (PCA)	6 Components	Moderate
4	Mutual Information	9 Features	High
5	Variance Threshold	7 Features	Low

The table presents the comparative analysis of different feature selection techniques used in the proposed cybercrime prediction framework. Feature selection plays an important role in identifying the most relevant attributes from the dataset, reducing dimensionality, improving model performance, and minimizing computational complexity.

Correlation Analysis selected 8 important features and achieved a moderate level of improvement. This method identifies relationships between input variables and the target variable by measuring correlation values. It helps remove redundant and highly correlated features from the dataset.

Recursive Feature Elimination (RFE) selected 10 significant features and achieved a high level of improvement. RFE works by repeatedly removing less important features and retaining the most relevant ones based on model performance. The results indicate that RFE provided better feature optimization and contributed to higher prediction accuracy.

Principal Component Analysis (PCA) reduced the dataset into 6 principal components with moderate improvement. PCA transforms the original features into a smaller set of

uncorrelated components while preserving most of the dataset variance. This technique is mainly used for dimensionality reduction and noise minimization.

Mutual Information selected 9 important features and also achieved a high improvement level. This technique measures the dependency between variables and identifies features that contain the highest amount of useful information for prediction. The results show that Mutual Information effectively improved model learning capability.

Variance Threshold selected 7 features but produced a low improvement level. This method removes features with very low variance because such features contribute less to prediction performance. However, compared to other methods, its contribution to overall accuracy was lower.

Overall, the table demonstrates that Recursive Feature Elimination (RFE) and Mutual Information were the most effective feature selection techniques in the proposed work. These methods significantly improved the regression model performance and helped achieve higher cybercrime prediction accuracy.

B. Regression Model Comparison

Table 3: Regression Model Comparison

S.No	Regression Model	Accuracy
1	Linear Regression	93%
2	Ridge Regression	89%
3	Lasso Regression	88%
4	Random Forest Regression	91%
5	XGBoost Regression	92%
6	ARIMA	86%

The comparative analysis shows that Linear Regression achieved the highest prediction accuracy of 93% due to effective preprocessing, normalization, and advanced feature selection.

XGBoost and Random Forest also demonstrated strong performance because of their ability to handle nonlinear cybercrime patterns.

ARIMA effectively captured temporal forecasting patterns but showed lower accuracy compared to machine learning regression models.

CONCLUSION

This research work presented an efficient cybercrime trend prediction framework using machine learning regression models, ARIMA forecasting, advanced preprocessing methods, and feature selection techniques. The study focused on improving cybercrime forecasting accuracy by integrating data preprocessing operations such as missing value handling, normalization, categorical encoding, outlier removal, and noise reduction with intelligent feature engineering approaches including Recursive Feature Elimination (RFE), Principal Component Analysis (PCA), Correlation Analysis, and Mutual Information.

Various regression algorithms such as Linear Regression, Ridge Regression, Lasso Regression, Random Forest Regression, and XGBoost Regression were implemented and compared with the ARIMA time-series forecasting model. Experimental analysis demonstrated that preprocessing and feature selection significantly enhanced prediction performance and reduced computational complexity. Among all implemented models, Linear Regression achieved the highest prediction accuracy of 93%, outperforming other regression and ARIMA models. XGBoost and Random Forest also produced strong results due to their capability to handle nonlinear cybercrime patterns effectively, while ARIMA successfully captured temporal forecasting trends.

The proposed framework provides an interpretable, reliable, and computationally efficient solution for cybercrime prediction and forecasting. The integration of machine learning techniques with statistical forecasting methods contributes to proactive cybersecurity analytics and intelligent threat monitoring. The developed system can support cybersecurity professionals, researchers, and law enforcement agencies in identifying cybercrime trends, improving decision-making, and enhancing preventive cyber security strategies.

Overall, the research demonstrates that advanced preprocessing, effective feature selection, and regression-based predictive modeling play a significant role in improving cybercrime forecasting accuracy and building intelligent cybersecurity prediction systems for future digital security applications.

REFERENCES

- [1] K. Jenga, C. Catal, and G. Kar, "Machine learning in crime prediction," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 2887–2913, 2023.
- [2] M. M. Barrenechea and M. A. Cano, "A systematic literature review on the use of artificial intelligence for cybercrime rate forecasting," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 3, 2024.
- [3] V. Mandalapu, L. Elluri, P. Vyas, and N. Roy, "Crime Prediction Using Machine Learning and Deep Learning: A Systematic Review and Future Directions," *IEEE Access*, vol. 11, pp. 60153–60170, 2023.
- [4] L. Elluri, V. Mandalapu, P. Vyas, and N. Roy, "Recent Advancements in Machine Learning for Cybercrime Prediction," *Journal of Computer Information Systems*, vol. 65, no. 2, pp. 249–263, 2025.

- [5] G. Box, G. Jenkins, and G. Reinsel, *Time Series Analysis: Forecasting and Control*, Wiley, 2015.
- [6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.
- [8] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research (JMLR)*, vol. 12, pp. 2825–2830, 2011.
- [9] B. Ghojogh et al., "Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review," *arXiv preprint arXiv:1905.02845*, 2019.
- [10] I. H. Sarker, "CyberLearning: Effectiveness Analysis of Machine Learning Security Modeling to Detect Cyber-Anomalies and Multi-Attacks," *arXiv preprint arXiv:2104.08080*, 2021.
- [11] R. Ch, T. R. Gadekallu, M. H. Abidi, and A. Al-Ahmari, "Computational System to Classify Cyber Crime Offenses using Machine Learning," *Sustainability*, vol. 12, no. 10, 2020.
- [12] A. Singh and S. Goswami, "Crime Rate Prediction Framework using Machine Learning Algorithm: A Review," *International Journal of Advanced Research in Management and Technology (IJARMT)*, vol. 2, no. 4, pp. 381–388, 2025.
- [13] B. Zhou et al., "Unsupervised Domain Adaptation for Crime Risk Prediction Across Cities," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 6, 2023.
- [14] S. Agarwal et al., "Statistical Forecasting Models for Crime Prediction Using Regression Analysis," *International Journal of Data Science and Analytics*, 2022.
- [15] S. Bandekar and K. Vijayalakshmi, "Machine Learning Models for Crime Reduction and Public Safety Analytics," *International Journal of Intelligent Systems and Applications*, 2021.
- [16] A. Kumar et al., "Crime Prediction and Pattern Identification using K-Nearest Neighbor Algorithm," *International Journal of Computer Applications*, 2020.
- [17] M. Landauer et al., "Time-Series Analysis Methods for Cybersecurity Analytics and Attack Prediction," *IEEE Access*, vol. 9, pp. 101604–101622, 2021.