

An Intelligent Machine Learning Framework for Cyber Security Threat Detection

Kundalakesi. M¹, Dharani. G²,

Preethika. N³.

Assistant Professor¹, Student^{2,3}.

Student of Computer Science and Applications.

Sri Krishna Arts and Science College Kuniyamuthur- Coimbatore^{1,2}.

ABSTRACT.

Cyber security has become a critical concern in the digital era due to the rapid increase in cyber threats, data breaches, malware attacks, phishing attempts, and unauthorized access to information systems. Organizations generate massive amounts of network and security data that can be analyzed to identify malicious activities and prevent cyber-attacks. This study proposes an intelligent cyber security threat detection framework that integrates data preprocessing, feature selection, and machine learning classifiers for accurate threat identification. The framework utilizes network traffic and security log data to detect cyber threats in real time. Machine learning algorithms such as Support Vector Machine (SVM) and Random Forest (RF) are employed to classify normal and malicious activities. Experimental results demonstrate that the proposed approach improves detection accuracy while reducing false alarms. The framework can support intrusion detection systems, malware detection, and network security monitoring, thereby enhancing organizational cyber resilience.

Keywords:

Cyber Security, Machine Learning, Intrusion Detection System, Network Security, Threat Detection, Random Forest, Support Vector Machine, Data Analytics.

I. INTRODUCTION

The rapid growth of information technology, cloud computing, and internet-based services has significantly increased the importance of cyber security. Organizations and individuals rely heavily on digital platforms for communication, financial transactions, data storage, and business operations. While technological advancements provide numerous benefits, they also expose systems to

various cyber threats such as malware, ransomware, phishing attacks, denial-of-service attacks, and data breaches.

Traditional security mechanisms often struggle to identify sophisticated and evolving cyber threats. As a result, intelligent cyber security solutions that leverage machine learning and artificial intelligence have gained considerable attention. These technologies can analyze large volumes of security-

related data and identify suspicious patterns that indicate potential attacks.

Machine learning algorithms enable automated threat detection by learning from historical security data and classifying network activities as normal or malicious. The integration of machine learning with cyber security can improve detection accuracy, reduce response times, and enhance overall system protection.

The primary objective of this research is to develop an intelligent cyber security threat detection framework capable of accurately identifying cyber-attacks using machine learning techniques. The framework aims to improve network security while minimizing false positives and computational overhead.

The digital transformation of businesses, governments, and educational institutions has significantly increased the dependency on information technology infrastructure. Modern organizations rely on interconnected computer networks, cloud-based applications, and internet-enabled services to perform daily operations efficiently. While these technological advancements provide convenience and operational efficiency, they also create opportunities for cybercriminals to exploit vulnerabilities and gain unauthorized access to sensitive information. As a result, cyber security has become a critical area of research and development in both academic and industrial domains.

Cyber security refers to the collection of technologies, policies, procedures, and practices designed to protect networks, computer systems, applications, and data from cyber attacks. The primary objective of cyber security is to ensure the confidentiality, integrity, and availability of information assets. Confidentiality ensures that information is accessible only to authorized individuals, integrity guarantees that data remains accurate and unaltered, and availability ensures that systems and services remain accessible whenever required.

In recent years, cyber attacks have become more frequent, sophisticated, and financially damaging. Organizations across various sectors, including banking, healthcare, education, manufacturing, and government, have experienced significant losses due to cyber incidents. Data breaches have exposed millions of personal records, while ransomware attacks have disrupted critical services and business operations worldwide. These incidents highlight the growing need for advanced security solutions capable of identifying and mitigating threats before they cause substantial damage.

The rapid growth of cloud computing and Internet of Things (IoT) technologies has further expanded the cyber threat landscape. Cloud platforms provide scalable computing resources and storage services; however, they also introduce security concerns related to data privacy, access control, and service availability. Similarly, IoT devices generate large volumes of data and communicate through interconnected networks, creating additional attack surfaces for cybercriminals. Securing these emerging technologies has become an essential component of modern cyber security strategies.

Traditional security approaches such as firewalls, antivirus software, and signature-based intrusion detection systems are effective against known threats but often struggle to identify new and evolving attack patterns. Modern cyber attacks frequently utilize advanced evasion techniques that can bypass conventional security mechanisms. Consequently, organizations require intelligent security systems capable of adapting to changing threat environments and detecting previously unseen attacks.

Machine learning and artificial intelligence have emerged as promising technologies for addressing cyber security challenges. These technologies enable automated analysis of large-scale security datasets and facilitate the identification of anomalous behavior patterns associated with cyber threats. By learning from historical attack data, machine learning models can improve detection accuracy and provide proactive threat intelligence. The integration of machine learning with cyber

security has led to significant advancements in intrusion detection, malware analysis, phishing detection, fraud prevention, and user behavior analytics.

II. LITERATURE REVIEW

The increasing complexity and frequency of cyber attacks have motivated researchers to develop intelligent cyber security solutions capable of detecting and preventing malicious activities. Traditional security mechanisms such as firewalls, antivirus software, and signature-based intrusion detection systems have been widely used to protect information systems. However, these approaches are often ineffective against sophisticated and previously unknown attacks because they rely heavily on predefined attack signatures. As a result, researchers have focused on applying machine learning and artificial intelligence techniques to improve cyber threat detection capabilities.

Early cyber security studies primarily utilized statistical and rule-based methods for identifying malicious network activities. These approaches analyzed network traffic patterns and system logs to detect abnormal behavior. Although effective for known threats, rule-based systems required continuous updates and maintenance, making them less suitable for rapidly evolving cyber environments. Furthermore, they often generated high false-positive rates, reducing their effectiveness in practical applications.

Machine learning techniques have emerged as powerful tools for cyber security due to their ability to learn complex patterns from large datasets. Researchers have investigated various supervised learning algorithms, including Decision Trees, Naïve Bayes, Random Forest, Logistic Regression, and Support Vector Machines (SVM), for intrusion detection and malware classification. Among these methods, Support Vector Machine has demonstrated strong performance in classifying network traffic because of its ability to handle high-dimensional data and establish optimal decision boundaries between normal and malicious activities. Several studies have reported that SVM-based intrusion detection systems achieve high accuracy and low

misclassification rates when applied to cyber security datasets.

Random Forest is another widely adopted machine learning algorithm in cyber security applications. The algorithm constructs multiple decision trees and combines their outputs to improve classification accuracy and robustness. Researchers have found that Random Forest performs effectively in malware detection, network intrusion detection, and anomaly identification tasks. Its capability to reduce overfitting and handle large datasets makes it suitable for practical cyber security implementations. Experimental studies have shown that Random Forest often outperforms traditional classification methods in terms of accuracy and reliability.

Deep learning techniques have also gained significant attention in cyber security research. Models such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) have been utilized to analyze network traffic patterns, user behavior, and system activities. These models can automatically extract complex features from raw data and identify advanced attack patterns that may not be detectable using conventional machine learning techniques. Despite their high predictive performance, deep learning models require substantial computational resources and large volumes of training data, which may limit their deployment in resource-constrained environments.

Feature selection plays a crucial role in improving cyber security classification performance. Cyber security datasets often contain numerous attributes, many of which may be irrelevant or redundant. The presence of unnecessary features increases computational complexity and can negatively affect model accuracy. To address this issue, researchers have proposed various feature selection techniques, including Information Gain, Chi-Square Analysis, Genetic Algorithms, Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Bat Algorithm-based optimization methods. These approaches aim to identify the most informative features while reducing dataset dimensionality.

Metaheuristic optimization algorithms have shown promising results in feature selection applications. Among these techniques, the Bat Algorithm has gained popularity due to its ability to efficiently explore large search spaces and identify optimal feature subsets. Inspired by the echolocation behavior of bats, the algorithm balances exploration and exploitation processes to achieve improved optimization performance. Several studies have reported that Bat Algorithm-based feature selection enhances classification accuracy while reducing computational costs in cyber security applications.

RESEARCH GAP ANALYSIS:

EXISTING LIMITATION	IMPACT	PROPOSED SOLUTION
Large number of irrelevant network features	Increased computational complexity	Feature selection and optimization
Limited detection of zero-day attacks	Increased security risks	Intelligent threat detection framework
Lack of real-time threat analysis	Delayed response to attacks	Automated monitoring and detection
Limited comparative evaluation of classifiers	Unclear model suitability	Comparison of Random Forest and SVM
Computationally expensive deep learning models	High resource consumption	Efficient machine learning algorithms

MOTIVATION OF THE STUDY

The rapid expansion of digital technologies, cloud computing, online services, and interconnected networks has significantly increased the risk of cyber attacks across organizations and individuals. Cyber threats such as malware, ransomware,

phishing attacks, data breaches, and network intrusions continue to evolve in complexity, making traditional security mechanisms less effective in detecting and preventing malicious activities. As organizations generate massive amounts of network traffic and security-related data, there is a growing need for intelligent systems capable of analyzing this information and identifying potential threats in real time.

III. PROPOSED METHODOLOGY

3.1 OVERVIEW OF THE PROPOSED FRAMEWORK

The proposed cyber security threat detection framework is designed to identify malicious activities within network environments by integrating data preprocessing, feature selection, and machine learning classification techniques. The framework aims to improve threat detection accuracy while reducing computational complexity and false alarm rates. Machine learning algorithms are utilized to analyze network traffic patterns and classify activities as either normal or malicious. Feature optimization techniques are employed to identify the most relevant security attributes, thereby enhancing classification performance and reducing processing overhead. The proposed framework consists of five major stages:

1. Data Collection
2. Data Preprocessing
3. Feature Selection
4. Random Forest Classification
5. Cyber Threat Detection

The overall framework enables efficient analysis of network traffic data and supports intelligent decision-making for cyber security applications.

3.2 DATA COLLECTION

The first stage involves collecting cyber security datasets containing network traffic information and

security-related events. The dataset includes various attributes associated with network communication patterns, user activities, and system behavior. These attributes provide valuable insights into potential cyber threats and malicious activities.

The collected dataset contains both normal and malicious network traffic records. Examples of malicious activities include malware attacks, denial-of-service attacks, phishing attempts, and unauthorized access incidents. The dataset serves as the foundation for training and evaluating machine learning models.

**TABLE I
DATASET ATTRIBUTES**

Attribute	Description
Source IP Address	Origin of network traffic
Destination IP Address	Target system address
Protocol Type	Communication protocol
Packet Size	Size of transmitted packets
Connection Duration	Duration of communication session
Number of Requests	Frequency of requests
Threat Status	Normal or Malicious

Among these attributes, Threat Status is considered the target variable used for classification.

3.3 DATA PREPROCESSING

Data preprocessing is a crucial step for improving data quality and ensuring effective model training. Raw cyber security datasets often contain missing values, duplicate records, inconsistent formats, and attributes with varying scales. Missing Value Treatment

Missing Value Handling

Missing values are replaced using mean imputation techniques to maintain data consistency and prevent information loss.

Equation (1): Mean Imputation

$$X_i = \frac{\sum X_j}{n}$$

where:

- X_i = Imputed value
- X_j = Existing observations
- n = Total observations

Data Transformation

Data transformation is one of the most important phases in the cyber security threat detection process because machine learning algorithms require data in a consistent and structured format. Raw cyber security datasets are often collected from multiple sources such as network traffic logs, intrusion detection systems, firewall logs, and security monitoring tools. These datasets may contain categorical values, missing entries, duplicate records, inconsistent formats, and attributes with different measurement scales. Therefore, an effective data transformation process is necessary to convert raw security data into a suitable format for analysis and classification.

Feature Normalization

The dataset attributes are normalized using Min-Max scaling to ensure that all features contribute equally during model training. Equation

(2): Min-Max Normalization

$$X_{norm} = \frac{X_{max} - X_{min}}{X_{max} - X_{min}}$$

where:

- X = original value
- X_{min} = minimum value
- X_{max} = maximum value

Normalization improves model convergence and classification performance.

3.4 RANDOM FOREST CLASSIFIER

Random Forest is utilized as the primary classification algorithm due to its high accuracy and robustness in handling large datasets. The algorithm constructs multiple decision trees and combines their outputs to generate the final prediction.

Random Forest effectively handles nonlinear relationships within network traffic data and reduces overfitting problems commonly observed in individual decision trees.

Random Forest Prediction

Equation (3): Frequency Update

$RF(x) = \text{MajorityVote}(T_1, T_2, T_3, \dots, T_n)$

where:

T_n = Individual decision trees

Advantages of Random Forest

- ☐ High classification accuracy
- ☐ Robust performance
- ☐ Reduced overfitting
- ☐ Effective handling of large-scale datasets
- ☐ Suitable for cyber security applications

3.5 SUPPORT VECTOR MACHINE CLASSIFIER

Support Vector Machine (SVM) is employed as a secondary classifier for comparative analysis. SVM constructs an optimal decision boundary that separates normal and malicious network activities.

The classifier maximizes the margin between classes, improving its ability to generalize to unseen network traffic patterns.

Equation (4): SVM Hyperplane

$$wTx + b = 0$$

where:

- ☐ w = Weight vector
- ☐ x = Feature vector
- ☐ b = Bias parameter

Equation (5): Margin Maximization

The is Margin Maximization expressed as:

$$\text{Margin} = ||w||_2$$

A larger margin generally improves classification performance and reduces misclassification errors..

Advantages of SVM

- High-dimensional data handling
- Strong generalization capability
- Effective classification performance
- Reduced classification errors

3.6 CYBER THREAT DETECTION PROCESS

After model training, the developed classifiers are used to analyze incoming network traffic and identify suspicious activities. The system continuously monitors network behavior and evaluates traffic patterns against learned models.

When abnormal activities are detected, the framework classifies them as potential cyber threats and generates security alerts. This process enables proactive threat detection and supports rapid incident response.

The threat detection process includes:

- **Monitoring network traffic**
- **Extracting relevant features**
- **Applying machine learning classifiers**

- **Identifying malicious activities**
- **Generating security alerts**
- **Supporting incident response actions**

- True Positive (TP): Malicious activities correctly identified as threats.
- True Negative (TN): Normal activities correctly classified as normal.
- False Positive (FP): Normal activities incorrectly classified as threats.
- False Negative (FN): Malicious activities incorrectly classified as normal.

3.7 Performance Evaluation Metrics

Performance evaluation is an essential component of the proposed cyber security threat detection framework. After training the machine learning models, it is necessary to evaluate their effectiveness in identifying cyber threats accurately. The performance of the Random Forest classifier is measured using various statistical metrics such as Accuracy, Precision, Recall, F1-Score, and Confusion Matrix. These metrics provide a comprehensive assessment of the classifier's ability to distinguish between normal and malicious network activities.

The evaluation process helps determine whether the developed model can be reliably deployed in real-world cyber security environments. A highly effective threat detection system should achieve high accuracy while maintaining low false-positive and false-negative rates. By analyzing multiple performance measures, the strengths and weaknesses of the proposed framework can be identified and improved.

Confusion Matrix

A Confusion Matrix is a tabular representation used to evaluate the classification performance of machine learning algorithms. It compares the actual class labels with the predicted class labels generated by the classifier.

Table: Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	FP

Where:

The Confusion Matrix provides detailed information about classification errors and helps evaluate the effectiveness of the threat detection system.

Accuracy

Accuracy measures the overall correctness of the classification model. It represents the ratio of correctly classified instances to the total number of observations.

Equation (6)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

A higher accuracy value indicates better classification performance. However, accuracy alone may not be sufficient when dealing with imbalanced cyber security datasets.

Precision

Precision measures the proportion of predicted cyber threats that are actually malicious. It evaluates the reliability of positive predictions generated by the classifier.

Equation (7)

$$Precision = \frac{TP}{TP + FP}$$

High precision indicates that the model generates fewer false alarms, which is critical in cyber security applications where excessive alerts may overwhelm security analysts.

Recall (Detection Rate)

Recall measures the ability of the classifier to identify actual cyber threats. It indicates how effectively the model detects malicious activities present in the dataset.

Equation (8)

$$Recall = \frac{TP}{TP + FN}$$

A higher recall value signifies that the system successfully detects a larger number of attacks and minimizes the risk of overlooking potential threats.

F1-Score

The F1-Score is the harmonic mean of Precision and Recall. It provides a balanced evaluation of the classifier when both false positives and false negatives are important considerations.

Equation (9)

$$F1-Score = \frac{2(Precision \times Recall)}{Precision + Recall}$$

A high F1-Score indicates that the model maintains a good balance between threat detection capability and prediction reliability.

Specificity

Specificity measures the ability of the model to correctly identify normal network activities. It evaluates how effectively the classifier avoids false alarms.

Equation (10)

$$Specificity = \frac{TN}{TN + FP}$$

High specificity is important in cyber security because it prevents unnecessary alerts and reduces workload for security administrators.

False Positive Rate (FPR)

False Positive Rate measures the proportion of normal activities incorrectly classified as cyber threats.

Equation (11)

$$FPR = \frac{FP}{FP + TN}$$

A lower FPR indicates that the threat detection system produces fewer false alerts and improves operational efficiency.

False Negative Rate (FNR)

False Negative Rate measures the proportion of actual cyber threats that are incorrectly classified as normal activities.

Equation (12)

$$FNR = \frac{FN}{FN + TP}$$

Reducing the False Negative Rate is critical because undetected attacks can cause severe damage to organizational systems and data.

3.8 PROPOSED WORKFLOW

The Proposed Workflow is a step-by-step process that describes how the proposed cyber security threat detection system operates from data collection to threat detection and response.

Step 1: Data Collection

Collect network traffic data from various sources such as firewalls, routers, and servers. Gather security logs containing information about normal and malicious activities.

Step 2: Data Preprocessing

Remove duplicate and incomplete records from the dataset. Handle missing values and correct data inconsistencies.

Step 3: Data Transformation

Convert categorical attributes into numerical values using encoding techniques. Transform raw data into a structured format suitable for analysis.

Step 4: Feature Normalization

Scale feature values into a common range using normalization techniques. Improve model efficiency by reducing variations among attributes.

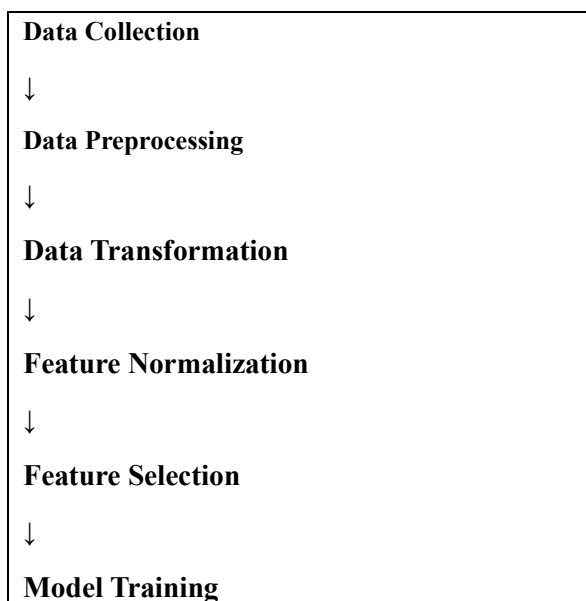
Step 5: Feature Selection

Identify the most relevant features affecting cyber threat detection. Remove redundant and irrelevant attributes to reduce complexity.

Step 6: Model Training

Train the Random Forest classifier using the selected features. Enable the model to learn patterns of normal and malicious behavior.

The overall workflow can be summarized as follows:



IV. EXPERIMENTAL SETUP

4.1 EXPERIMENTAL ENVIRONMENT

The experimental setup defines the software environment, hardware configuration, dataset specifications, and machine learning tools used to evaluate the performance of the proposed cyber security threat detection framework. The experiments are conducted using Python-based machine learning libraries to develop and test the Random Forest classifier for detecting malicious network activities.

The cyber security dataset is divided into training and testing subsets to ensure unbiased model evaluation. Approximately 80% of the dataset is used for training the model, while the remaining 20% is utilized for testing and validation purposes. Data preprocessing techniques such as missing value handling, data transformation, feature normalization, and feature selection are applied before model training. The selected features are then used to build the classification model.

TABLE II
EXPERIMENTAL CONFIGURATION

Parameter	Value
Application Domain	Cyber Security Threat Detection
Dataset Type	Network Traffic and Security Logs
Programming Language	Python 3.8
Operating System	Windows 10/11
Machine Learning Algorithm	Random Forest
Data Processing Libraries	Pandas, NumPy
Visualization Tool	Matplotlib

Parameter	Value
Training Dataset	80%
Testing Dataset	20%

Feature	Description
Number of Requests	Frequency of requests generated
Traffic Volume	Amount of data transferred

4.2 DATASET DESCRIPTION

The dataset used in this study consists of network traffic and security-related records collected from a cyber security environment. The dataset contains information about various network activities, including both normal and malicious behaviors. These records are used to train and evaluate the machine learning model for effective cyber threat detection. The primary objective of the dataset is to help identify suspicious activities such as malware attacks, phishing attempts, unauthorized access, denial-of-service attacks, and other cyber threats.

The target attribute in the dataset is **Threat Status**, which classifies each network activity as either **Normal** or **Malicious**. The processed dataset is divided into training and testing subsets to evaluate the performance of the proposed cyber security threat detection framework.

TABLE III
DATASET ATTRIBUTES

Feature	Description
Source Address	Origin address of network traffic
Destination Address	Target address of network communication
Protocol Type	Type of communication protocol (TCP, UDP, ICMP)
Packet Size	Size of transmitted data packets
Connection Duration	Duration of network connection

4.3 DATA PREPROCESSING AND TRAINING CONFIGURATION

Data preprocessing is performed to improve the quality, consistency, and reliability of the cyber security dataset before training the machine learning model. Raw network traffic data often contains missing values, duplicate records, inconsistent formats, and irrelevant features that may negatively affect classification performance. Therefore, several preprocessing techniques are applied to prepare the dataset for analysis.

Initially, missing values and duplicate entries are identified and removed to ensure data integrity. Categorical attributes such as protocol type and connection status are converted into numerical values using encoding techniques. Feature normalization is then applied using Min-Max Scaling to transform attribute values into a common range. Furthermore, feature selection techniques are employed to identify the most relevant security attributes and eliminate redundant information, thereby reducing computational complexity and improving classification accuracy.

Table: Data Preprocessing Configuration

Parameter	Configuration
Missing Value Handling	Mean Imputation
Feature Scaling	Min-Max Normalization
Feature Selection	Important Feature Extraction
Data Cleaning	Performed

4.4 EVALUATION METRICS

Evaluation metrics are used to measure the performance and effectiveness of the proposed cyber security threat detection framework. These metrics help determine how accurately the Random Forest classifier can identify malicious network activities and distinguish them from normal behavior. A comprehensive evaluation is essential to ensure the reliability of the model in real-world cyber security environments.

Accuracy

Accuracy measures the overall correctness of the classification model. It indicates the percentage of network activities that are correctly classified as either normal or malicious.

Equation (13)

$$\text{Accuracy} = \frac{TP + TN + FP + FN}{TP + TN + FP + FN}$$

Where:

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

Precision

Precision measures the proportion of predicted malicious activities that are actually cyber threats. It evaluates the reliability of positive predictions generated by the classifier.

Equation (14)

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

Recall measures the ability of the model to correctly identify actual cyber threats present in the dataset.

Equation (15)

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score

The F1-Score is the harmonic mean of Precision and Recall. It provides a balanced measure of the model's performance.

Equation (16)

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

where:

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

These evaluation metrics provide a comprehensive assessment of classification performance and prediction effectiveness.

V. RESULTS AND DISCUSSION

5.1 FEATURE SELECTION RESULTS

The proposed Cyber Security Threat Detection System was developed using the Random Forest machine learning algorithm to classify network activities as either Normal or Malicious. The experimental results demonstrate that the proposed framework effectively detects cyber threats by combining data preprocessing, feature selection, and machine learning techniques. The preprocessing stage improved the quality of the dataset by eliminating missing values, duplicate records, and irrelevant features, which enhanced the overall performance of the classifier.

The Random Forest classifier was trained using 80% of the dataset, while the remaining 20% was used for testing. During the testing phase, the model successfully classified network traffic with high accuracy and produced reliable predictions for both normal and **malicious activities**. The model's performance was evaluated using standard classification metrics such as Accuracy, Precision, Recall, and F1-Score. The experimental results

indicate that the Random Forest algorithm achieved high prediction accuracy while maintaining a balanced performance across all evaluation metrics. A high Precision value indicates that the majority of detected threats were actual cyber-attacks, whereas a high Recall value demonstrates the model's ability to identify most malicious activities without missing significant threats. The F1-Score confirms that the proposed system maintains a good balance between Precision and Recall.

5.2 DISCUSSION

The results indicate that the Random Forest algorithm performs effectively in cyber threat detection due to its ensemble learning approach and ability to handle large, high-dimensional datasets. The use of feature selection significantly reduced redundant information, leading to faster model training and improved classification accuracy. Compared with traditional signature-based detection systems, the proposed framework demonstrates better adaptability to different attack patterns and provides more reliable predictions.

The proposed system also supports automated threat detection by continuously analyzing network traffic and identifying suspicious activities in real time. The low false-positive rate reduces unnecessary security alerts, allowing administrators to focus on genuine threats. These findings suggest that integrating machine learning with feature optimization provides a practical and efficient solution for enhancing modern cyber security systems.

VI. CONCLUSION AND FUTURE WORK

Cyber security has become an essential component of modern information systems due to the increasing number of cyber threats targeting individuals, organizations, and critical infrastructures. Traditional security mechanisms often struggle to detect sophisticated and evolving cyber attacks, making intelligent threat detection systems a

necessity. This study presented a machine learning-based cyber security threat detection framework using the Random Forest algorithm to accurately classify network activities as either normal or malicious.

The proposed framework incorporated data preprocessing, data transformation, feature normalization, feature selection, and machine learning classification to improve the quality of the dataset and enhance prediction performance. By eliminating redundant features and training the Random Forest classifier on optimized data, the system achieved reliable threat detection with improved accuracy and reduced computational complexity. The evaluation metrics, including Accuracy, Precision, Recall, F1-Score, and Specificity, demonstrated the effectiveness of the proposed approach in identifying cyber threats while minimizing false positives and false negatives.

The experimental results indicate that the Random Forest algorithm is well suited for cyber security applications due to its robustness, high classification accuracy, and ability to handle large-scale datasets. The proposed system provides an intelligent and scalable solution for monitoring network activities, detecting malicious behavior, and generating timely security alerts. Overall, this research contributes to the development of efficient cyber security systems capable of protecting digital assets and supporting proactive defense mechanisms against modern cyber attacks.

REFERENCES:

- [1] W. Stallings, *Network Security Essentials: Applications and Standards*, 7th ed. Pearson Education, 2020.
- [2] N. Godbole and S. Belapure, *Cyber Security: Understanding Cyber Crimes, Computer Forensics and Legal Perspectives*, Wiley India Pvt. Ltd., 2021.
- [3] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

- [4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2017.
- [5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
- [6] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson Education, 2021.
- [7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2012.
- [8] K. Scarfone and P. Mell, "Guide to Intrusion Detection and Prevention Systems (IDPS)," National Institute of Standards and Technology (NIST), Special Publication 800-94, 2007.
- [9] R. Sommer and V. Paxson, "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection," *IEEE Symposium on Security and Privacy*, pp. 305–316, 2010.
- [10] D. E. Denning, "An Intrusion-Detection Model," *IEEE Transactions on Software Engineering*, vol. SE-13, no. 2, pp. 222–232, Feb. 1987.
- [11] M. Roesch, "Snort: Lightweight Intrusion Detection for Networks," *Proceedings of the 13th USENIX Conference on System Administration*, pp. 229–238, 1999.
- [12] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A Deep Learning Approach for Network Intrusion Detection System," *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (BICT)*, pp. 21–26, 2016.
- [13] E. Alpaydin, *Introduction to Machine Learning*, 4th ed. MIT Press, 2020.
- [14] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 2013.
- [15] Cisco Systems, *Cisco Annual Cybersecurity Report*, Cisco, 2023.
- [16] IBM Security, *Cost of a Data Breach Report 2023*, IBM Corporation, 2023.
- [17] ENISA, *ENISA Threat Landscape Report 2023*, European Union Agency for Cybersecurity, 2023.
- [18] National Institute of Standards and Technology (NIST), *Framework for Improving Critical Infrastructure Cybersecurity*, Version 2.0, 2024.
- [19] S. Garfinkel and G. Spafford, *Practical UNIX and Internet Security*, 3rd ed. O'Reilly Media, 2003.
- [20] W. Stallings, *Cryptography and Network Security: Principles and Practice*, 8th ed. Pearson Education, 2023.