International Journal of Engineering Research and Science & Technology

ENHANCED INTRUSION DETECTION SYSTEM IN COMMUNICATION SYSTEM USING WEIGHTED FEATURE SELECTION AND A VOTING CLASSIFIER

GUIDE

[1]Name : Mrs. V. PRAGATHI - pragathi@tkrcet.com

Assistant Professor

Department of Computer Science and Engineering

TKRCET Autonomous ,

Hyderabad, India

[2] V. Sravanthi

sravanthivaggu@gmail.com

Student, Department of Computer Science and Engineering TKRCET Autonomous , Hyderabad, India

[3] V.Akhilesh

akhileshvr00@gmail.com

Student, Department of Computer Science and Engineering TKRCET Autonomous, Hyderabad, India

[4] V.Pranavi

pranavivishwanatham29@gmail.com

Student, Department of Computer Science and Engineering TKRCET Autonomous , Hyderabad, India

[5] T.Sainath

sainathchowdary0246@gmail.com

Student, Department of Computer Science and Engineering TKRCET Autonomous , Hyderabad, India

DEPARTMENT OF COMPUTER SCIENCE &ENGINEERING

TKR COLLEGE OF ENGINEERING & TECHNOLOGY

(AUTONOMOUS)

(Accredited by NBA and NAAC with 'A+' Grade)

Medbowli, Meerpet, Saroornagar, Hyderabad-500097

ABSTRACT

The rapid growth of communication networks and internet-based applications has significantly increased the risk of cyber-attacks, making Intrusion Detection Systems (IDS) a critical component of network security. However, modern IDS face challenges in handling high-dimensional network traffic data, which often contains redundant and irrelevant features that degrade detection performance and increase computational cost.

This project proposes an enhanced intrusion detection system that integrates **CHI-REV weighted feature selection** with a **Voting Classifier-based ensemble learning**

approach to improve detection accuracy and efficiency. The system preprocesses the CICIDS-2017 dataset by cleaning, normalizing, and transforming network traffic data into suitable formats for analysis. The CHI-REV method is applied to identify and select the most informative features, thereby reducing dimensionality and eliminating noise.

To overcome the limitations of single-model classifiers, the proposed system employs a voting ensemble that combines multiple machine learning algorithms, enabling more robust and reliable classification of network traffic. The model is evaluated across binary, multi-class, and all-class classification tasks using metrics such as accuracy, precision, recall, and F1-score. Experimental results demonstrate that the ensemble approach outperforms traditional single classifiers, particularly in detecting rare and complex attack types while maintaining high overall accuracy.

The study concludes that combining weighted feature selection with ensemble learning significantly enhances IDS performance, providing a scalable, efficient, and reliable solution for securing modern communication networks.

Keywords

Intrusion Detection System (IDS), CHI-REV Feature Selection, Voting Classifier, Ensemble Learning, CICIDS-2017 Dataset, Network Security, Machine Learning, Feature Selection, Cyber Attack Detection, Classification Algorithms.

I. INTRODUCTION

In the modern digital era, communication networks form the backbone of critical infrastructure, enabling services such as online banking, cloud computing, e-commerce, and real-time communication. As reliance on interconnected systems increases, so does the exposure to cyber threats including Denial of Service (DoS), Distributed Denial of Service (DDoS), brute-force attacks, and advanced persistent threats. These attacks compromise the confidentiality, integrity, and availability of data, making network security a major concern for organizations and individuals.

Intrusion Detection Systems (IDS) play a vital role in safeguarding communication networks by monitoring traffic and identifying suspicious activities. Traditional IDS are primarily signature-based, meaning they detect attacks by comparing network activity against known attack patterns. While effective for known threats, these systems fail to detect new or

evolving attacks, often referred to as zero-day attacks. To address this limitation, anomaly-based IDS using Machine Learning (ML) techniques have been introduced, which learn normal network behavior and identify deviations that indicate potential intrusions.

Despite their advantages, ML-based IDS face significant challenges when dealing with high-dimensional network datasets such as CICIDS-2017. These datasets contain numerous features, many of which are redundant or irrelevant. The presence of such features leads to increased computational complexity, longer training times, and reduced detection accuracy due to noise and overfitting. Therefore, effective feature selection becomes essential to improve system performance and efficiency.

To overcome these challenges, this project introduces an enhanced IDS that combines **CHI-REV weighted feature selection** with an **ensemble-based Voting Classifier**. The CHI-REV technique selects the most relevant features by evaluating their statistical relationship with the target variable, thereby reducing dimensionality and eliminating noise. Furthermore, instead of relying on a single classifier, the proposed system integrates multiple machine learning models through a voting mechanism to produce more reliable and accurate predictions.

The system is developed and evaluated using the CICIDS-2017 dataset, considering multiple classification scenarios such as binary, multi-class, and all-class intrusion detection. By combining efficient feature selection with ensemble learning, the proposed approach aims to improve detection accuracy, enhance robustness against diverse attack types, and reduce false negatives, especially for rare and critical attacks. Overall, this work contributes to the development of a more scalable, accurate, and practical intrusion detection system for modern communication networks.

II. LITERATURE REVIEW

Intrusion Detection Systems (IDS) have become an essential component in securing modern communication networks against cyber threats and unauthorized access. With the rapid growth of network traffic and the increasing sophistication of attacks, traditional security mechanisms are no longer sufficient. Early research in IDS primarily focused on signature-based techniques, which detect known attack patterns effectively but fail to identify new or evolving threats. As a result, researchers have shifted towards anomaly-based and

machine learning-driven approaches to improve detection capabilities.

Several studies have explored the role of IDS in different computing environments such as cloud, fog, and Internet of Things (IoT) networks. These environments generate large volumes of distributed data, making intrusion detection more complex. Researchers have emphasized the need for scalable and intelligent IDS models that can efficiently process high-dimensional data while maintaining accuracy. These works highlight that the effectiveness of IDS largely depends on data representation, feature relevance, and the choice of learning algorithms.

A significant body of research has focused on handling high-dimensional network traffic data. In many datasets, a large number of features are either redundant or irrelevant, which negatively impacts model performance. Studies on dimensionality reduction and feature selection have shown that selecting a subset of meaningful features can significantly improve classification accuracy while reducing computational cost. Techniques such as statistical feature selection, recursive feature elimination, and optimization-based methods have been widely used to address this issue.

Machine learning algorithms have been extensively applied in IDS to improve detection performance. Models such as Support Vector Machines (SVM), Decision Trees, Naive Bayes, and Random Forest have demonstrated strong classification capabilities. However, individual models often suffer from limitations such as overfitting, sensitivity to noise, or poor performance on imbalanced datasets. To overcome these limitations, researchers have explored ensemble learning methods, which combine multiple classifiers to improve overall prediction accuracy and robustness.

Another important research direction involves optimizing IDS performance through hybrid approaches that combine feature selection with advanced classification techniques. Studies have shown that integrating feature selection methods with machine learning models can significantly enhance detection efficiency and reduce execution time. Additionally, handling class imbalance using techniques like resampling and boosting has been found to improve detection rates, especially for rare attack types.

Despite these advancements, existing IDS approaches still face challenges in balancing accuracy, computational efficiency, and robustness. Many methods

focus primarily on improving accuracy without adequately reducing feature redundancy or computational cost. Furthermore, some models struggle to detect rare and complex attacks effectively.

Based on these observations, this project addresses the identified research gap by combining **CHI-REV weighted feature selection** with a **Voting Classifier-based ensemble approach**. This combination aims to reduce irrelevant features while improving classification performance and robustness. By leveraging both efficient feature selection and ensemble learning, the proposed system seeks to provide a more accurate, scalable, and reliable intrusion detection solution for modern communication networks.

III. METHODOLOGY

1. Data Collection and Preprocessing

The system utilizes the CICIDS-2017 dataset, which contains both normal and malicious network traffic. All dataset files are merged into a unified dataset, followed by preprocessing steps such as cleaning column names, handling missing and infinite values, and removing duplicate records. Features with constant values are eliminated to avoid unnecessary computations. Numerical features are normalized using Min-Max scaling to ensure uniformity. Additionally, three types of labels are generated: binary (benign vs attack), multi-class (grouped attack categories), and all-class (individual attack types).

2. CHI-REV Feature Selection

To address the issue of high-dimensional data, the CHI-REV (Chi-Square Reverse) feature selection method is applied. This technique evaluates the statistical relationship between each feature and the target class. Continuous features are discretized, and contingency tables are constructed to compute chi-square contributions. A weighted scoring mechanism is then used to rank features based on their importance. The top-ranked features are selected, significantly reducing dimensionality while preserving relevant information.

3. Baseline Model Development

After feature selection, multiple machine learning models are trained to establish a performance baseline. These include Naive Bayes, Linear Support Vector Machine (SVM), Decision Tree, and Random Forest. The dataset is split into training and testing sets using stratified sampling to maintain class distribution. Each model is trained and evaluated using metrics such as accuracy, precision, recall, and F1-score. The

Random Forest model is used as a benchmark to compare improvements achieved by the proposed approach.

4. Voting Ensemble Classifier

To improve detection reliability, an ensemble-based Voting Classifier is implemented. Multiple classifiers such as Random Forest, Extra Trees, Histogram Gradient Boosting, and XGBoost are combined using a soft voting mechanism. Each model contributes to the final prediction, and the class with the highest aggregated probability is selected as the output. Different combinations of classifiers are tested to identify the most effective ensemble configuration. This approach enhances robustness and reduces the limitations of individual models.

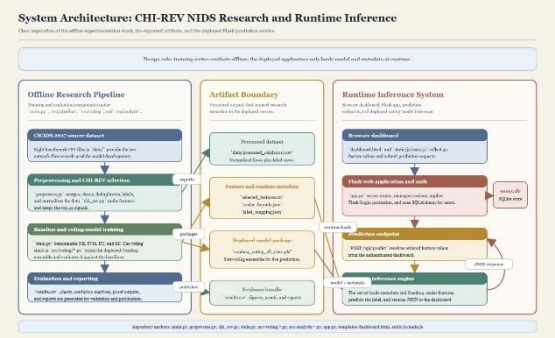
5. Model Evaluation and Validation

The performance of the proposed system is evaluated using multiple metrics including accuracy, macro F1-score, ROC-AUC, precision-recall curves, and false negative rate. Comparative analysis is performed between the baseline Random Forest model and the voting ensemble. Additional validation techniques such as statistical significance testing, robustness analysis, and calibration evaluation are conducted to ensure reliability and consistency of results.

6. Deployment and User Interface

The final trained model is integrated into a web-based application for practical usage. The system accepts user input in the form of network feature values, applies the same preprocessing and scaling techniques, and predicts whether the traffic is benign or malicious. The output is presented in a user-friendly format, enabling real-time decision-making and easy interaction for users.

IV. SYSTEM ARCHITECTURE



V. RESULTS & DISCUSSION

The proposed Intrusion Detection System (IDS) was evaluated using the CICIDS-2017 dataset after applying preprocessing and CHI-REV feature selection. The dataset was cleaned, normalized, and reduced to the most relevant features, enabling efficient model training and improved classification performance. The system was tested on three different tasks: binary classification (benign vs attack), multi-class classification (attack

categories), and all-class classification (individual attack types). This multi-level evaluation ensures that the system is capable of handling both general and detailed intrusion detection scenarios.

The baseline machine learning models, including Naive Bayes, Linear SVM, Decision Tree, and Random Forest, were trained using the selected features. Among these models, Random Forest achieved the highest performance, with accuracy values reaching nearly 99.9% across all classification tasks. Decision Tree also demonstrated strong results, while Linear SVM produced moderate performance. Naive Bayes showed comparatively lower accuracy due to its limitations in handling complex feature distributions and imbalanced data. These results validate the effectiveness of the preprocessing and feature selection stages implemented in the system.

To further enhance detection capability, a Voting Classifier-based ensemble approach was introduced. This method combines multiple machine learning models and aggregates their predictions to produce a final decision. The ensemble model showed consistent improvements over the baseline Random Forest classifier. Although the increase in overall accuracy was relatively small due to the already high baseline, the ensemble provided better stability and reliability in predictions. This demonstrates that combining multiple classifiers can overcome the limitations of individual models and improve overall system robustness.

A key observation from the results is the improvement in macro F1-score, particularly for the all-class classification task. Unlike accuracy, macro F1-score gives equal importance to all classes, including rare attack types. The proposed system achieved higher macro F1-scores, indicating better performance in detecting minority classes. This is especially important in intrusion detection, where rare attacks can be highly dangerous but are often overlooked by traditional models.

The system also demonstrated significant improvement in detecting rare and critical attacks such as Heartbleed. The voting ensemble reduced the false negative rate and improved the F1-score for such attacks compared to the Random Forest baseline. This highlights the ability of the proposed system to capture complex and less frequent attack patterns, which is essential for real-world cybersecurity applications.

Further evaluation using metrics such as ROC-AUC, Precision-Recall AUC, and calibration error showed that the ensemble model provides more reliable and well-

calibrated predictions. Statistical tests confirmed that the improvements are significant and not due to random variation. Additionally, the system maintained better performance under noisy conditions, indicating strong robustness and stability.

VI. CONCLUSION

This project presents an enhanced Intrusion Detection System (IDS) for communication networks by integrating **CHI-REV weighted feature selection** with a **Voting Classifier-based ensemble learning approach**. The primary objective was to address the challenges of high-dimensional network data and improve intrusion detection accuracy while maintaining computational efficiency. By applying effective preprocessing techniques on the CICIDS-2017 dataset, the system successfully removed noise, handled missing values, and normalized features, creating a strong foundation for model development.

The CHI-REV feature selection method played a crucial role in reducing dataset dimensionality by identifying the most relevant features. This not only improved computational efficiency but also enhanced the learning capability of machine learning models by eliminating redundant and irrelevant data. The baseline models validated the correctness of the implementation, with Random Forest achieving very high accuracy, consistent with the reference study.

A major contribution of this work is the use of a Voting Classifier, which combines multiple machine learning models to produce more reliable predictions. The ensemble approach demonstrated improved performance compared to the single-model baseline, particularly in terms of macro F1-score, rare attack detection, model stability, and prediction reliability. Although the increase in overall accuracy was small due to the already strong baseline performance, the improvements in detecting minority attack classes and reducing false negatives are highly significant for real-world security applications.

The system also proved to be robust and consistent under different evaluation conditions, including noisy data and statistical validation tests. Furthermore, the deployment of the trained model into a web-based application highlights the practical usability of the proposed IDS, enabling real-time prediction and user interaction.

In conclusion, the combination of weighted feature selection and ensemble learning provides an effective and scalable solution for modern intrusion detection. The

proposed system not only improves detection performance but also enhances reliability and robustness, making it suitable for real-world communication network security.

REFERENCES

[1] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, p. 20, Dec. 2019, doi: 10.1186/s42400-019-0038-7.

[2] K. Peng, V. C. M. Leung, L. Zheng, S. Wang, C. Huang, and T. Lin, "Intrusion detection system based on decision tree over big data in fog environment," *Wireless Communications and Mobile Computing*, vol. 2018, pp. 1–10, Mar. 2018, doi: 10.1155/2018/4680867.

[3] P. Pirozmand, M. A. Ghafary, S. Siadat, and J. Ren, "Intrusion detection into cloud-fog-based IoT networks using game theory," *Wireless Communications and Mobile Computing*, vol. 2020, pp. 1–9, Nov. 2020, doi: 10.1155/2020/8819545.

[4] V. K. Singh, B. Nathani, and M. Kumar, "WEED-MC: Wavelet transform for energy efficient data gathering and matrix completion," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 5, pp. 1066–1073, May 2020.

[5] V. K. Singh, M. Kumar, and S. Verma, "Node scheduling and compressed sampling for event reporting in WSNs," *IEEE Transactions on Network Science and Engineering*, vol. 6, no. 3, pp. 418–431, Jul. 2019.

[6] V. K. Singh, M. Kumar, and S. Verma, "Accurate detection of important events in WSNs," *IEEE Systems Journal*, vol. 13, no. 1, pp. 248–257, Mar. 2019.

[7] A. Shivhare, V. K. Singh, and M. Kumar, "Anticomplementary triangles for efficient coverage in sensor network-based IoT," *IEEE Systems Journal*, vol. 14, no. 4, pp. 4854–4863, Dec. 2020.

[8] V. K. Singh, C. Singh, and H. Raza, "Event classification and intensity discrimination for forest fire inference with IoT," *IEEE Sensors Journal*, vol. 22, no. 9, pp. 8869–8880, May 2022.

[9] R. G. Mantovani, A. L. D. Rossi, J. Vanschoren, B. Bischl, and A. C. P. L. F. de Carvalho, "Effectiveness of random search in SVM hyper-parameter tuning," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, Jul. 2015, pp. 1–8.

[10] R. Abdulhammed, H. Musafer, A. Alessa, M. Faezipour, and A. Abuzneid, "Features dimensionality reduction approaches for machine learning based network intrusion detection," *Electronics*, vol. 8, no. 3, p. 322, Mar. 2019, doi: 10.3390/electronics8030322.