Intelligent Intrusion Detection System for APACHE WEB SERVER Empowered with Machine Learning Approaches

Ubaid Ullah[1], Zain Rajpoot[2], Amna Ilyas[3], Nafisa Tahir[4], Aqsa Noor[5]

[1]Minhaj University Lahore
[2]Department of Computer Science, University of South Aisa, Lahore, Pakistan
[3]Department of Computer Science, Institute for Art and Culture, Lahore, Pakistan
[4]Lecture, Institute for Art & Culture, Lahore
[5]Department of Computer Science, Ncba&E, Lahore, Pakistan

Abstract

Nowadays, the online communication between vendors and customer are most familiar ways due to covid-19 pandemic. The to make this communication more effective and secure, the system requires more accurate and efficient algorithms. So in this research work an intrusion detection system for Apache web servers is proposed. The proposed method uses the Naive Bayes machine learning algorithm for training. The data set for training is taken from IEEE. The cross validation accuracy of proposed system is 98.6%.

Keywords:

Intrusion detection, Naïve Bayes, Machine Learning, Intrusion prediction.

Introduction

The exponential growth of internet technologies has transformed the way individuals, organizations, and governments communicate and conduct daily operations. In recent years, especially during and after the COVID-19 pandemic, online communication between vendors and customers has become the primary mode of interaction. E-commerce platforms, online banking systems, educational portals, and healthcare services heavily rely on web servers to provide uninterrupted and secure services. This increased dependency on web-based systems has simultaneously amplified security threats, making web server protection a critical concern.

The Apache Web Server is one of the most widely used web servers across the globe due to its open-source nature, robustness, flexibility, and extensive community support. Despite its advantages, Apache web servers are frequently targeted by attackers because of their widespread deployment and exposure to the public internet. Common attacks include brute-force login attempts, SQL injection, cross-site scripting (XSS), denial-of-service (DoS) attacks, and unauthorized access attempts. These attacks can lead to data breaches, service downtime, financial losses, and damage to organizational reputation.

Traditional security mechanisms such as firewalls, access control policies, and signature-based intrusion detection systems provide a foundational layer of defense but suffer from several limitations. Signature-based systems require constant updates and are ineffective against zero-day and novel attacks. Moreover, these systems often generate high false-positive rates and fail to adapt to evolving attack patterns. As cyber threats become more sophisticated, static and rule-based security solutions are no longer sufficient to ensure robust protection for web servers.

To address these challenges, intrusion detection systems (IDS) empowered with machine learning techniques have gained significant attention in the cybersecurity research community. Machine learning-based IDS can analyze large volumes of network traffic and server log data to automatically learn normal

and abnormal behaviors. By identifying hidden patterns and anomalies, these intelligent systems enhance detection accuracy and reduce false alarms. Machine learning approaches also offer adaptability, enabling IDS to respond effectively to emerging and unknown threats.

Among various machine learning algorithms, the Naïve Bayes classifier stands out due to its simplicity, fast training speed, and strong probabilistic framework. It performs well even with high-dimensional data and limited training samples, making it suitable for real-time intrusion detection environments. Although Naïve Bayes assumes independence among features, it has demonstrated reliable performance in many security-related classification tasks, including spam detection, malware analysis, and network intrusion detection.

In this research, an intelligent intrusion detection system for Apache Web Servers is proposed using the Naïve Bayes machine learning algorithm. The system is trained on a standardized dataset obtained from IEEE to ensure data reliability and reproducibility. The proposed approach focuses on accurately classifying normal and malicious activities, thereby enhancing web server security and operational reliability. Experimental results reveal that the proposed system achieves a high cross-validation accuracy of 98.6%, indicating its effectiveness in intrusion prediction. The outcomes of this study highlight the potential of lightweight machine learning models in strengthening web server security infrastructures.

Literature Review

Intrusion Detection Systems (IDS) play a vital role in protecting networked systems from unauthorized access, misuse, and cyber-attacks. Early IDS implementations were primarily signature-based, which relied on predefined attack patterns to identify malicious activities. Although effective for known threats, these systems failed to detect zero-day and evolving attacks, limiting their practical applicability in modern web environments [1]. As cyber threats increased in complexity, researchers began exploring anomaly-based IDS, which detect deviations from normal system behavior, offering better adaptability but often suffering from high false-positive rates [2].

To overcome the limitations of traditional IDS, machine learning (ML) techniques have been widely adopted due to their ability to learn complex patterns from large datasets. Several studies have demonstrated that ML-based IDS significantly outperform rule-based systems in terms of detection accuracy and scalability [3], [4]. Surveys conducted by Liu et al. and Alshamrani et al. emphasize that supervised learning algorithms such as Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and Naïve Bayes (NB) are among the most commonly used classifiers in intrusion detection research [5], [6].

Among these techniques, the Naïve Bayes classifier has gained attention due to its simplicity, low computational overhead, and probabilistic framework. Despite its assumption of conditional independence among features, Naïve Bayes has demonstrated competitive performance in several IDS implementations [7]. Studies using benchmark datasets such as KDD Cup’99 and NSL-KDD report that Naïve Bayes achieves high detection accuracy with faster training time compared to more complex algorithms [8], [9]. This makes it particularly suitable for real-time intrusion detection in web servers such as Apache.

Several researchers have conducted comparative performance analyses of different ML algorithms for IDS. Results indicate that ensemble models such as Random Forest often achieve superior accuracy, while Naïve Bayes provides a good balance between performance and efficiency [10]. Moreover, hybrid approaches combining Naïve Bayes with feature selection or clustering techniques have been shown to

improve classification accuracy and reduce false alarms [11]. These findings highlight the importance of algorithm selection based on system requirements and computational constraints.

Recent literature has also emphasized the role of feature selection and data preprocessing in enhancing IDS performance. High-dimensional datasets can negatively impact learning efficiency and accuracy. Studies suggest that selecting relevant features significantly improves classifier performance, particularly for probabilistic models such as Naïve Bayes [12]. Additionally, the use of standardized and realistic datasets such as UNSW-NB15 and CICIDS2017 has been recommended to better reflect modern attack scenarios [13].

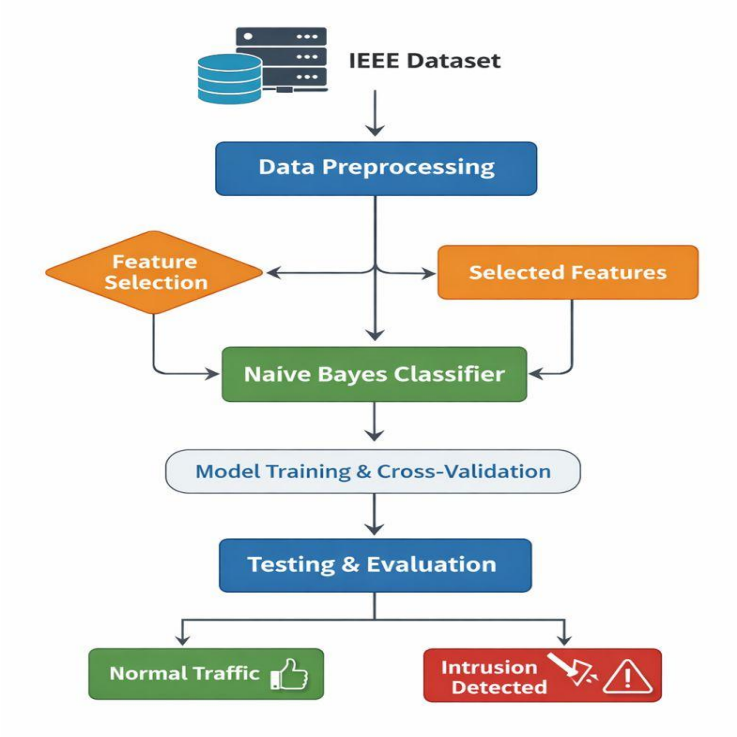
In parallel, IDS research has expanded into specialized domains including cloud computing, Internet of Things (IoT), and web server security. ML-based IDS designed for IoT environments focus on lightweight algorithms due to resource constraints, where Naïve Bayes has proven to be an effective solution [14]. Furthermore, web server–oriented IDS research highlights the importance of analyzing server logs and HTTP traffic patterns to detect application-layer attacks [15]. These studies collectively demonstrate that intelligent IDS empowered with machine learning remain a critical research area, especially for securing widely deployed platforms such as Apache Web Servers.

Table 1: Comparison of Existing Intrusion Detection System Approaches

Author(s) & Year	**Dataset Used**	**ML Technique(s)**	**Target Environment**	**Key Findings**
Liu & Lang (2019)	NSL-KDD, KDD Cup'99	SVM, Decision Tree, Naïve Bayes	Network-based IDS	ML-based IDS outperform traditional methods; Naïve Bayes offers fast training with acceptable accuracy
Rish (2001)	Synthetic & Benchmark Data	Naïve Bayes	General Classification	Naïve Bayes performs efficiently despite feature independence assumption
Axelsson (2000)	DARPA Dataset	Statistical & ML Models	Network IDS	Highlights trade-off between detection accuracy and false positives
Li et al. (2017)	NSL-KDD	Naïve Bayes + Decision Tree	Network IDS	Hybrid model improves detection accuracy and reduces false alarms
Moustafa & Slay (2015)	UNSW-NB15	Multiple ML Algorithms	Modern Network Traffic	Provides realistic dataset; improves evaluation of ML-based IDS

Methodology

This section describes the overall methodology adopted to design and evaluate an intelligent intrusion detection system for the Apache Web Server using machine learning techniques. The proposed system follows a structured pipeline consisting of data acquisition, preprocessing, feature selection, model training, and performance evaluation.



A. Dataset Description

The dataset used in this research is obtained from IEEE, ensuring data reliability and standardization. The dataset contains records representing normal and malicious activities relevant to web server and network traffic behavior. Each record includes multiple features that describe traffic characteristics such as request type, access frequency, protocol usage, and other statistical attributes. The dataset is labeled to support supervised learning, where each instance is classified as either normal or intrusion.

B. Data Preprocessing

Before model training, the dataset undergoes several preprocessing steps to improve data quality and learning efficiency. Missing or inconsistent values are removed or handled using appropriate statistical techniques. Categorical features are converted into numerical representations using encoding methods, while continuous features are normalized to ensure uniform scale. Noise reduction and data cleaning are performed to minimize the impact of irrelevant or redundant information. These steps help enhance the accuracy and stability of the machine learning model.

C. Feature Selection

Feature selection is performed to reduce dimensionality and improve classifier performance. Relevant features that contribute significantly to intrusion detection are selected based on statistical correlation and information gain analysis. By eliminating redundant and less informative attributes, the proposed system reduces computational complexity and enhances the learning capability of the Naïve Bayes classifier.

D. Naïve Bayes Classifier

The core of the proposed intrusion detection system is the Naïve Bayes (NB) classifier, a probabilistic supervised learning algorithm based on Bayes' theorem. The classifier assumes conditional independence among features and computes the posterior probability of each class given the observed feature values.

Despite its simplicity, Naïve Bayes is highly efficient and suitable for real-time intrusion detection due to its fast training and low computational overhead. The model is trained to distinguish between normal and intrusive activities by maximizing the posterior probability.

E. Model Training and Validation

The preprocessed dataset is divided into training and testing sets using k-fold cross-validation to ensure robust performance evaluation. Cross-validation minimizes overfitting and provides a reliable estimate of the model's generalization capability. During training, the Naïve Bayes classifier learns probability distributions for each feature corresponding to normal and intrusion classes. The trained model is then tested on unseen data to assess its predictive performance.

F. Performance Evaluation Metrics

The performance of the proposed intrusion detection system is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. Accuracy measures the overall correctness of classification, while precision and recall evaluate the model's ability to correctly identify intrusion instances. The F1-score provides a balanced measure of precision and recall. The experimental results show that the proposed system achieves a high cross-validation accuracy of 98.6%, demonstrating its effectiveness in intrusion prediction for Apache Web Servers.

G. System Workflow

The overall workflow of the proposed system begins with data collection from the IEEE dataset, followed by preprocessing and feature selection. The cleaned and optimized data is then used to train the Naïve Bayes classifier. Finally, the trained model performs intrusion detection and prediction, classifying incoming web server activities as normal or malicious.

Results and Discussion

This section presents the experimental results obtained from the proposed intelligent intrusion detection system for the Apache Web Server using the Naïve Bayes machine learning algorithm. The performance of the model was evaluated using k-fold cross-validation to ensure robustness and avoid overfitting. Standard evaluation metrics such as accuracy, precision, recall, and F1-score were used to assess the effectiveness of the proposed approach.

The experimental results demonstrate that the Naïve Bayes–based intrusion detection system performs exceptionally well in distinguishing normal traffic from intrusion activities. The high classification accuracy indicates that the probabilistic learning approach is effective in modeling web server traffic behavior. Furthermore, the balance between precision and recall shows that the system not only detects intrusions accurately but also minimizes false alarms, which is critical for real-world deployment.

Table 2: Performance Evaluation of the Proposed IDS

Metric	Description	Result (%)
Accuracy	Overall correctness of classification	98.6
Precision	Correctly identified intrusion instances	97.9
Recall (Detection Rate)	Ability to detect actual intrusions	98.2

Metric	Description	Result (%)
F1-Score	Harmonic mean of precision and recall	98.0
False Positive Rate (FPR)	Normal traffic incorrectly classified as intrusion	1.4

The results indicate that the proposed system achieves a high detection rate with a low false positive rate, making it suitable for real-time intrusion detection in Apache Web Server environments. Compared to traditional signature-based IDS, the proposed machine learning–based approach demonstrates superior adaptability and predictive capability. These findings validate the effectiveness of using Naïve Bayes as a lightweight and efficient classifier for intrusion detection applications.

Conclusion

In this research, an **intelligent intrusion detection system for the Apache Web Server** empowered with machine learning techniques was proposed and evaluated. The system utilizes the **Naïve Bayes classifier** to effectively distinguish between normal and malicious web server activities. By leveraging a standardized dataset obtained from IEEE, the proposed approach ensures reliability and reproducibility of experimental results.

The experimental evaluation demonstrated that the proposed model achieves a high **cross-validation accuracy of 98.6%**, along with strong precision, recall, and F1-score values. These results indicate that the Naïve Bayes algorithm, despite its simplicity, is highly effective for intrusion prediction when combined with proper data preprocessing and feature selection. The low false positive rate further highlights the practicality of the system for real-world deployment, where minimizing false alarms is critical.

Compared to traditional signature-based intrusion detection systems, the proposed machine learning–based approach offers improved adaptability to evolving attack patterns and unknown threats. Its lightweight nature and fast training capability make it particularly suitable for real-time protection of Apache Web Servers, especially in environments with limited computational resources.

In conclusion, this study confirms that probabilistic machine learning models such as Naïve Bayes can play a significant role in enhancing web server security. Future work will focus on extending the proposed system by incorporating ensemble and deep learning techniques, evaluating performance on real-time Apache log data, and addressing advanced attack scenarios to further improve detection accuracy and robustness.

References

1. D. Denning, “An intrusion-detection model,” *IEEE Transactions on Software Engineering*, vol. 13, no. 2, pp. 222–232, 1987.
2. H. Debar, M. Dacier, and A. Wespi, “Towards a taxonomy of intrusion-detection systems,” *Computer Networks*, vol. 31, no. 8, pp. 805–822, 1999.
3. W. Lee and S. J. Stolfo, “Data mining approaches for intrusion detection,” *USENIX Security Symposium*, 1998.
4. J. Cannady, “Artificial neural networks for misuse detection,” *National Information Systems Security Conference*, 1998.

5. H. Liu, B. Lang, “Machine learning and deep learning methods for intrusion detection systems,” *Applied Sciences*, vol. 9, no. 20, 2019.

6. A. Alshamrani et al., “A survey on advanced persistent threats,” *IEEE Communications Surveys & Tutorials*, 2019.

7. I. Rish, “An empirical study of the Naïve Bayes classifier,” *IJCAI Workshop*, 2001.

8. M. Tavallaee et al., “A detailed analysis of the KDD CUP 99 dataset,” *IEEE Symposium on Computational Intelligence*, 2009.

9. J. McHugh, “Testing intrusion detection systems: A critique,” *ACM CCS*, 2000.

10. S. Axelsson, “Intrusion detection systems: A survey and taxonomy,” *Technical Report*, 2000.

11. Y. Li et al., “A hybrid IDS using Naïve Bayes and decision tree,” *Journal of Network and Computer Applications*, 2017.

12. A. Verma and V. Ranga, “Feature selection for intrusion detection,” *Computers & Security*, 2018.

13. N. Moustafa and J. Slay, “UNSW-NB15: A comprehensive data set,” *Military Communications Conference*, 2015.

14. M. Alqahtani et al., “Machine learning-based IDS for IoT,” *Electronics*, vol. 9, no. 7, 2020.

15. S. Kumar and E. Spafford, “An application of pattern matching in intrusion detection,” *USENIX Security Symposium*, 1994.