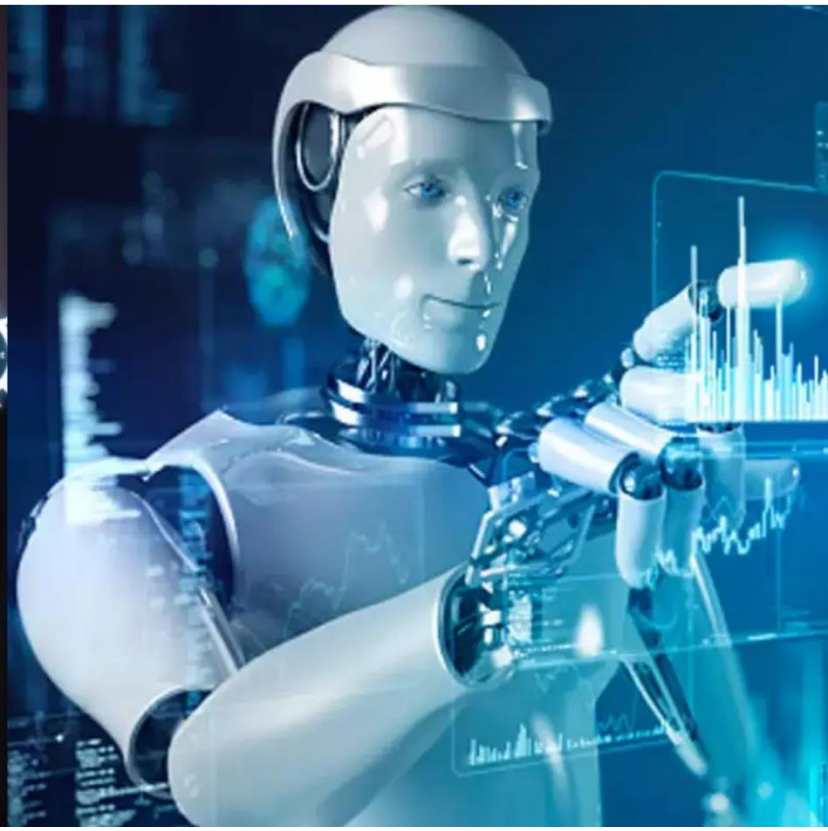
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 4, April 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504176|

Detection of Network Intrusion System

R. Femimol, Thammarapu Sarayu, Thammi Manideep, Thota Varun, Tivanala Vyshnavi

Department of Computer Science and Engineering, Bharath Institute of Science and Technology, BIHER, Selaiyur, Chennai, Tamil Nadu, India

Department of Computer Science and Engineering, Bharath Institute of Science and Technology, BIHER, Selaiyur, Chennai, Tamil Nadu, India

Department of Computer Science and Engineering, Bharath Institute of Science and Technology, BIHER, Selaiyur, Chennai, Tamil Nadu, India

Department of Computer Science and Engineering, Bharath Institute of Science and Technology, BIHER, Selaiyur, Chennai, Tamil Nadu, India

Department of Computer Science and Engineering, Bharath Institute of Science and Technology, BIHER, Selaiyur, Chennai, Tamil Nadu, India.

ABSTRACT: The rapid growth of networked systems has significantly increased the risk and complexity of cyber-attacks, making traditional signature-based security mechanisms inadequate for modern environments. This paper presents the design and implementation of an intelligent Network Intrusion Detection System (NIDS) using a machine learning approach based on the Random Forest algorithm. The proposed system utilizes the NSL-KDD dataset, which provides a refined and balanced representation of network traffic, to train and evaluate the model.A comprehensive pre-processing pipeline is employed to convert categorical attributes into numerical form and to transform multi-class labels into a binary classification problem distinguishing normal and malicious traffic. The Random Forest classifier, configured with multiple decision trees, effectively captures complex patterns in high-dimensional data and enhances detection performance.

Experimental evaluation demonstrates that the model achieves high accuracy with improved detection capability while maintaining a low false positive rate. Performance is validated using standard metrics such as confusion matrix and ROC-AUC analysis. The results indicate that ensemble learning techniques provide a robust and scalable solution for intrusion detection.The proposed system offers an efficient framework for identifying anomalous network behavior and can serve as a foundation for future real-time and large-scale cybersecurity applications.

KEYWORDS: Network Intrusion Detection System (NIDS), Machine Learning, Random Forest, NSL-KDD Dataset, Cybersecurity, Ensemble Learning.

I. INTRODUCTION

The rapid advancement of information technology and the widespread use of computer networks have significantly transformed modern society. From financial institutions and healthcare systems to government organizations and educational platforms, almost every sector relies heavily on networked systems for communication, data storage, and operational efficiency. However, this increasing dependence has also made networks highly vulnerable to a wide range of cyber threats. Malicious activities such as Denial of Service (DoS) attacks, unauthorized access, probing, and data breaches have become more frequent and sophisticated, posing serious risks to data security and system integrity.

Traditionally, network security has been maintained using firewalls and signature-based intrusion detection systems (IDS). These systems operate by comparing network traffic against a database of known attack patterns or predefined rules. While they are effective in detecting previously identified threats, they suffer from significant limitations. In particular, they are unable to detect new or unknown attacks, commonly referred to as zero-day attacks. Additionally,

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504176|

these systems require constant updates to their signature databases and often generate a high number of false alerts, making them inefficient in handling large-scale and dynamic network environments.

To overcome these challenges, there has been a growing interest in the use of intelligent and adaptive techniques for intrusion detection. Machine learning has emerged as a promising solution in this domain due to its ability to learn patterns from historical data and make predictions on unseen data. Unlike traditional rule-based systems, machine learning models can identify hidden relationships and anomalies within network traffic, enabling the detection of both known and unknown attacks. This capability makes machine learning particularly suitable for modern cybersecurity applications, where the threat landscape is continuously evolving.
Among various machine learning algorithms, ensemble learning methods have gained attention for their improved accuracy and robustness. Random Forest, an ensemble-based classification technique, is widely recognized for its ability to handle high-dimensional data, reduce overfitting, and provide reliable predictions. It operates by constructing multiple decision trees using random subsets of data and features, and then combining their outputs through a majority voting mechanism. This approach enhances the model's generalization ability and makes it highly effective for complex classification tasks such as network intrusion detection.

In this work, a Network Intrusion Detection System (NIDS) is developed using the Random Forest algorithm. The system is trained and evaluated using the NSL-KDD dataset, which is an improved version of the widely used KDD Cup 99 dataset. The NSL-KDD dataset addresses key issues such as redundancy and class imbalance, providing a more realistic and challenging benchmark for evaluating intrusion detection models. The dataset includes multiple features that describe network traffic behavior, allowing the model to learn meaningful patterns associated with normal and malicious activities.

A comprehensive preprocessing phase is implemented to prepare the dataset for model training. This includes transforming categorical attributes into numerical values using label encoding and converting the multi-class attack labels into a binary classification problem. The processed data is then used to train the Random Forest model, which learns to distinguish between normal and attack traffic based on feature patterns. The performance of the system is evaluated using standard metrics such as accuracy, confusion matrix, and Receiver Operating Characteristic (ROC) analysis.

The proposed system aims to provide an efficient and scalable solution for detecting network intrusions with high accuracy and low false alarm rates. By leveraging the strengths of machine learning and ensemble techniques, the system enhances the capability of traditional intrusion detection methods and contributes to improved network security. Furthermore, the approach can be extended to real-time environments and integrated with advanced security frameworks to provide continuous monitoring and protection.

II. LITERATURE REVIEW

The field of Network Intrusion Detection Systems (NIDS) has evolved significantly over the years in response to the increasing complexity and frequency of cyber-attacks. Early intrusion detection techniques were primarily based on signature-based methods, where predefined patterns of known attacks were stored in a database and compared with incoming network traffic. While these methods were effective in detecting known threats with high accuracy, they were limited in their ability to identify new or unknown attacks. This limitation led to the development of anomaly-based detection systems, which focus on identifying deviations from normal network behaviour.

Anomaly-based intrusion detection systems rely on statistical models and machine learning techniques to establish a baseline of normal activity. Any significant deviation from this baseline is considered suspicious and flagged as a potential intrusion. Although this approach provides the advantage of detecting previously unseen attacks, early implementations suffered from high false positive rates, which reduced their practical usability. To address these challenges, researchers have explored various machine learning algorithms to improve detection accuracy and reduce false alarms.

Among the earliest machine learning approaches, Naive Bayes gained attention due to its simplicity and computational efficiency. It applies probabilistic reasoning based on Bayes' theorem and assumes independence among features.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504176|

However, in the context of network traffic data, where features are often correlated, this assumption limits its performance. As a result, Naive Bayes tends to produce lower accuracy and higher false positive rates compared to more advanced methods.

Support Vector Machines (SVM) have also been widely used for intrusion detection due to their strong classification capabilities in high-dimensional spaces. SVM works by identifying an optimal hyperplane that separates different classes with maximum margin. Although SVM provides good accuracy, its computational complexity increases significantly with large datasets, making it less suitable for real-time or large-scale applications.

Decision Trees are another popular approach due to their interpretability and ease of implementation. They classify data by recursively splitting it based on feature values, creating a tree-like structure of decisions. While Decision Trees are fast and easy to understand, they are prone to overfitting, especially when dealing with complex and high-dimensional datasets such as network traffic logs. This limitation affects their ability to generalize well on unseen data.

To overcome the drawbacks of individual classifiers, ensemble learning techniques have been introduced. Ensemble methods combine multiple models to improve overall performance and stability. Random Forest is one of the most widely used ensemble algorithms for intrusion detection. It constructs multiple decision trees using random subsets of data and features, and the final prediction is determined through majority voting. This approach significantly reduces overfitting and enhances classification accuracy. Studies have shown that Random Forest consistently outperforms traditional machine learning models in terms of detection rate and false alarm reduction.

The availability of benchmark datasets has played a crucial role in evaluating intrusion detection systems. The KDD Cup 99 dataset was one of the earliest and most widely used datasets for this purpose. However, it contained a large number of redundant records, which led to biased learning and overly optimistic performance results. To address these issues, the NSL-KDD dataset was introduced as an improved version. It eliminates duplicate records and provides a more balanced distribution of data, making it a more reliable benchmark for evaluating machine learning models.

Recent research trends have also explored the use of deep learning techniques, such as Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), for intrusion detection. These methods are capable of learning complex patterns and temporal dependencies in network traffic. However, they require large amounts of data, high computational resources, and longer training times, which can limit their practical deployment in real-time systems.

In comparison, ensemble-based machine learning approaches, particularly Random Forest, offer a balance between performance, efficiency, and scalability. They are capable of handling high-dimensional data, providing feature importance insights, and achieving high detection accuracy with relatively lower computational cost. These advantages make Random Forest a suitable choice for developing an effective Network Intrusion Detection System.

Based on the analysis of existing literature, it is evident that machine learning techniques have significantly improved the performance of intrusion detection systems. The transition from traditional rule-based methods to intelligent data-driven approaches has enhanced the ability to detect both known and unknown threats. This study builds upon these advancements by implementing a Random Forest-based NIDS using the NSL-KDD dataset to achieve reliable and efficient intrusion detection.

III. METHODOLOGY

The proposed Network Intrusion Detection System (NIDS) follows a structured machine learning pipeline consisting of data acquisition, preprocessing, model training, and performance evaluation. The objective is to accurately classify network traffic as either normal or malicious using the Random Forest algorithm.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504176|

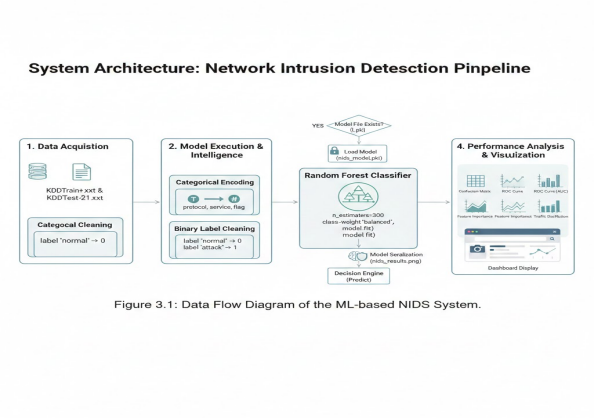


Figure 3.1: Data Flow Diagram of the ML-based NIDS System.

Step1. Dataset Description

The system utilizes the NSL-KDD dataset, which is an improved version of the KDD Cup 99 dataset. It addresses issues such as redundancy and class imbalance, providing a reliable benchmark for intrusion detection research. The dataset contains 41 features representing different characteristics of network traffic, including protocol type, service, and traffic statistics.

Step2.Data preprocessing

Data preprocessing is a crucial step to ensure compatibility with machine learning algorithms. The dataset includes both categorical and numerical attributes. Categorical features such as protocol type, service, and flag are converted into numerical values using label encoding.The target variable, which originally contains multiple attack types, is transformed into a binary classification problem by mapping normal traffic to one class and all attack types to another. Additionally, irrelevant attributes are removed, and the dataset is checked for inconsistencies to improve data quality.

Step3.Data Integration and Splitting

To maintain consistency in encoding, training and testing datasets are combined during preprocessing and later separated. The processed data is divided into training and testing sets, where the training data is used to build the model and the testing data is used to evaluate its performance.

Step4.Model development

The Random Forest algorithm is employed as the core classification model. It is an ensemble learning technique that constructs multiple decision trees using random subsets of data and features. Each tree independently predicts the class label, and the final output is determined by majority voting.

The model is configured with multiple estimators to improve accuracy and reduce overfitting. This approach enables the system to handle high-dimensional data and capture complex patterns in network traffic.

Step5.Model Evaluation

The performance of the proposed system is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. A confusion matrix is used to analyze classification results, and the Receiver Operating Characteristic (ROC) curve is used to assess the model's ability to distinguish between classes.

Step6.Model persistence

To enable efficient reuse, the trained model is saved using serialization techniques. This allows the system to load the trained model without retraining, making it suitable for deployment in real-world environments.

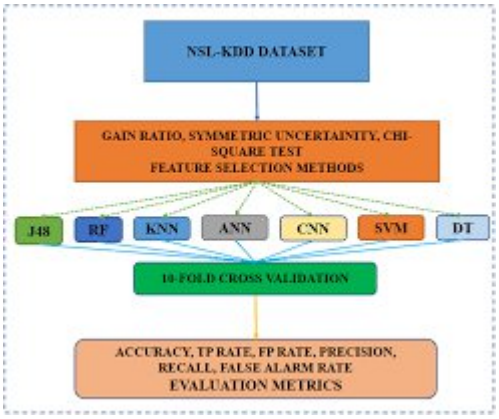
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

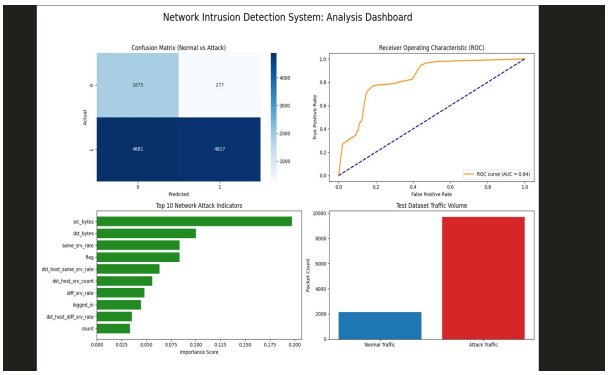
|DOI: 10.15680/IJIRSET.2026.1504176|



IV. RESULTS

The proposed Text-to-SQL system was implemented to allow users to retrieve database information using natural language queries. The system was developed using a Streamlit interface, a Large Language Model for query interpretation, and a MySQL database for storing and retrieving data. After integrating these components, the system was tested with different user queries to evaluate its functionality and performance.During testing, users entered questions in natural language through the Streamlit interface. The system successfully interpreted the input and converted it into corresponding SQL queries. These queries were executed on the MySQL database, and the retrieved results were displayed in a structured format on the user interface. The results showed that the system was able to correctly identify relevant database tables and attributes for most user queries.

The experimental results demonstrate that the system simplifies database interaction for users who do not have knowledge of SQL. Users were able to retrieve data using simple English questions without needing to understand database schemas or query syntax. The system also reduced the time required to access information compared to manual SQL query writing.However, the system may encounter limitations when processing highly complex queries or ambiguous user input. In such cases, additional improvements in prompt design, query validation, and database schema understanding can further enhance the accuracy and reliability of the system.



Overall, the results indicate that the proposed approach effectively bridges the gap between natural language interaction and structured database querying, making data access more efficient and user-friendly.

V. CONCLUSION

This paper presented the design and implementation of a machine learning-based Network Intrusion Detection System (NIDS) using the Random Forest algorithm. The system was developed to address the limitations of traditional signature-based security mechanisms by enabling the detection of both known and unknown cyber-attacks. By utilizing

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 4, April 2026

|DOI: 10.15680/IJIRSET.2026.1504176|

the NSL-KDD dataset and applying effective preprocessing techniques, the model was trained to accurately classify network traffic as normal or malicious.

The experimental results demonstrate that the proposed approach achieves high detection accuracy with a low false alarm rate. The use of an ensemble learning method enhances the system's ability to handle high-dimensional data and complex traffic patterns while reducing overfitting. Additionally, feature importance analysis influencing intrusion provides valuable insights into the key factors detection.

Overall, the proposed system offers a reliable and scalable solution for improving network security. It highlights the effectiveness of machine learning techniques in detecting intrusions and lays the foundation for further enhancements, such as real-time deployment and integration with advanced security frameworks.

VI. FUTURE ENHANCEMENT

The proposed Network Intrusion Detection System (NIDS) can be further enhanced to improve its performance, scalability, and real-world applicability. One potential extension is the integration of deep learning techniques, such as neural networks and recurrent architectures, to capture more complex temporal patterns in network traffic. Another important improvement is the transition from offline analysis to real-time intrusion detection by integrating the model with live network monitoring systems. This would enable immediate identification and response to cyber threats. Additionally, incorporating larger and more recent datasets can improve the model's ability to detect emerging and sophisticated attack patterns.The system can also be extended to support multi-class classification, allowing it to identify specific types of attacks rather than only distinguishing between normal and malicious traffic. Furthermore, deployment in cloud-based environments can enhance scalability and allow the system to handle high-volume network data efficiently.Finally, integrating automated response mechanisms, such as alert generation or attack mitigation strategies, can transform the system from a detection tool into a comprehensive network security solution.

REFERENCES

[1] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE Symp. Computational Intelligence for Security and Defense Applications*, 2009, pp. 1–6.
[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[3] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[5] Canadian Institute for Cybersecurity, "NSL-KDD dataset for network intrusion detection," University of New Brunswick, 2009.
[6] S. Axelsson, "Intrusion detection systems: A survey and taxonomy," *Technical Report*, Chalmers University, 2000.
[7] W. Stallings, *Network Security Essentials: Applications and Standards*, 6th ed. Pearson, 2017.
[8] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
[9] K. Scarfone and P. Mell, "Guide to intrusion detection and prevention systems (IDPS)," NIST Special Publication 800-94, 2007.
[10] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proc. IEEE Symp. Security and Privacy*, 2010, pp. 305–316.
[11] H. Liao, C. R. Lin, Y. C. Lin, and K. Y. Tung, "Intrusion detection system: A comprehensive review," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 16–24, 2013.
[12] S. Mukkamala, A. H. Sung, and A. Abraham, "Intrusion detection using ensemble of soft computing paradigms," in *Proc. International Conf. Intelligent Systems Design and Applications*, 2003, pp. 239–248.
[13] J. Zhang, M. Zulkernine, and A. Haque, "Random-forest-based network intrusion detection systems," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 38, no. 5, pp. 649–659, 2008.
[14] G. Creech and J. Hu, "A semantic approach to host-based intrusion detection systems using contiguous and discontiguous system call patterns," *IEEE Trans. Computers*, vol. 63, no. 4, pp. 807–819, 2014.
[15] A. Patcha and J. M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer Networks*, vol. 51, no. 12, pp. 3448–3470, 2007.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details