# INTRUSION DETECTION SYSTEM USING MACHINE LEARNING

**NEHA UNNISA[1], MOHD IMROZ[2], SUMAYYA UNISSA[3]**

[1]Assistant Professor, Department of Computer Science and Engineering, Deccan College of Engineering and Technology, Hyderabad, Telangana, India.

[2]Assistant Professor, Department of Computer Science and Engineering, Deccan College of Engineering and Technology, Hyderabad, Telangana, India.

[3]Department Of Computer Science and Engineering, Deccan College of Engineering and Technology, Hyderabad, Telangana, India.

**nehaunnisa@deccancollege.ac.in[1], mohdimroz@deccancollege.ac.in[2], sumayyaunissa19@gmail.com3**

## ABSTRACT

**Intrusion Detection System (IDS) leveraging advanced machine learning techniques, notably the Random Forest algorithm, to enhance network security against evolving cyber threats. Traditional IDS systems often struggle with new attack types and processing efficiency. By employing Gaussian Naive Bayes, Decision Trees, Logistic Regression, Random Forest, and Gradient Classifier, the proposed system aims to effectively recognize patterns and identify intrusions even in complex and limited data scenarios. Specifically, the Random Forest algorithm is highlighted for its ability to handle high-dimensional and noisy network traffic data, adapt to diverse intrusion scenarios, and exhibit robust classification performance. The system architecture involves data pre-processing, feature engineering, and model development, culminating in real-time implementation. Advantages include robustness to complexity, adaptability to diverse scenarios, efficient noise handling, high performance, precision in detection, minimized false alarms, scalability, and computational efficiency. The system architecture encompasses multi-tiered data collection, pre-processing, detection engines, decision-making modules, centralized management consoles, logging, and reporting mechanisms. Continuous adaptation and integration within the broader security ecosystem ensure swift detection, response, and mitigation of intrusions in network environments. Overall, the proposed IDS system presents a comprehensive approach to bolstering network security against modern cyber threats.**

**Keywords**: Intrusion Detection System (IDS), Machine Learning Techniques, Random Forest Algorithm, Cyber Threats, Pattern Recognition, Network Security, Data Pre-processing, Feature Engineering, Real-time Implementation, Scalability

## I. INTRODUCTION

In the rapidly evolving landscape of cybersecurity, the need for robust defence mechanisms against sophisticated cyber threats is more pressing than ever. Intrusion Detection Systems (IDS) stand as crucial sentinels, constantly surveillant network traffic to identify and mitigate potential intrusions. However, the complexity and diversity of modern attacks pose significant challenges to traditional rule-based IDS, often resulting in inadequate adaptability and detection accuracy.

In response to these limitations, the fusion of machine learning techniques with IDS has emerged as a promising avenue for fortifying network security. This thesis focuses on exploring the integration of machine learning algorithms within IDS frameworks to enhance their capabilities. By harnessing the power of machine learning, these systems can autonomously learn and adapt to evolving threats, enabling the identification of anomalies and malicious activities with higher accuracy and efficiency.

The primary objective of this thesis is to investigate and demonstrate the potential of machine learning-driven IDS in transforming the landscape of intrusion detection. Through the utilization of diverse machine learning models, such as neural networks, decision trees, or anomaly detection algorithms, the aim is to empower IDS to discern subtle patterns within network traffic, differentiate normal behaviour from anomalies, and effectively detect and respond to potential threats in real-time.

Expanding on this research, an additional focus will be given to evaluating the scalability of machine learning-integrated IDS solutions and their adaptability to diverse network environments. By addressing these aspects, the study aims to provide comprehensive insights into the practical applicability and potential challenges associated with deploying machine learning-driven IDS at scale.

This thesis endeavours to delve into the technical intricacies and practical implications of integrating machine learning into IDS, aiming to not only improve detection accuracy but also to minimize false positives and enhance the overall resilience of network security. By exploring this innovative amalgamation, the intent is to contribute to the advancement of intrusion detection systems, paving the way for more adaptive, efficient, and robust cybersecurity measures in today's digital landscape.

## II. LITERATURE SURVEY

**[1] A Novel Network Intrusion Detection System Based on CNN.**

**Chen, L.; Kuang, X.; Xu, A.; Suo, S.; Yang, Y. In Proceedings of the 2020 Eighth International Conference on Advanced Cloud and Big Data (CBD), Taiyuan, China, 5–6 December 2020; pp. 243–247.**

Network Intrusion Detection Systems (NIDS) have traditionally relied on machine learning techniques like Support Vector Machines, Bayesian Classification, and k-Means clustering for enhancing network security. However, these methods often require manual feature selection and are constrained by inherent limitations. To overcome these challenges, this study introduces a novel NIDS framework leveraging Convolutional Neural Networks (CNNs). Unlike conventional methods, CNNs autonomously extract discriminative features directly from raw network traffic data, enabling the development of sophisticated detection models capable of discerning subtle intrusion patterns. Comprehensive experiments conducted on established benchmark datasets validate the superiority of the CNN-based NIDS framework. Notably, models trained on raw traffic data consistently demonstrate higher accuracy compared to those trained solely on extracted features, underscoring the intrinsic advantage of leveraging raw data in deep-learning-based NIDS models. This research represents a significant advancement in network security, offering enhanced detection capabilities while mitigating the limitations associated with traditional machine learning techniques.

**[2] Dynamic detection of malicious intrusion in wireless network based on improved random forest algorithm.**
**Chen, Y.; Yuan, F. In Proceedings of the 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 14–16 April 2022; pp. 27–32.**

To enhance the detection capability of malicious nonlinear scrambling intrusions in wireless networks and ensure network security, this paper proposes a method based on an improved random forest algorithm. The method begins by constructing a model for the malicious nonlinear scrambling intrusion signal in wireless networks. Subsequently, it employs a terminal equipment's physical layer characteristic detection method to decompose the signal's characteristics. Through the analysis of test statistics and detection thresholds, spectrum characteristic quantities of the intrusion signal are extracted. Next, an improved random forest algorithm is utilized for signal characteristic reorganization and fuzzy clustering analysis. By leveraging the clustering distribution of spectral features, the method optimizes the detection of malicious nonlinear scrambling intrusions under random forest learning. This is achieved through a combination of reinforcement learning decision techniques and static feature fusion methods. Simulation results demonstrate the effectiveness of the proposed approach, showing strong feature clustering in detecting malicious nonlinear scrambling intrusions, leading to high accuracy in identifying intrusion information.

**[3] Naive Bayes modification for intrusion detection system classification with zero probability.**

**Kurniawan, Y.; Razi, F.; Nofiyati, N.; Wijayanto, B.; Hidayat, M. *Bull. Electr. Eng. Inform.* 2021, *10*, 2751–2758.**

One approach commonly employed in intrusion detection systems is the implementation of the Naïve Bayes algorithm. However, Naïve Bayes encounters a significant challenge when encountering probabilities of zero, leading to inaccurate predictions or even failure to produce any predictions at all. To address this limitation, this paper proposes two modifications to the Naïve Bayes algorithm. The first modification involves eliminating variables associated with zero probabilities, while the second modification replaces multiplication operations with addition operations. These modifications are selectively applied only when the Naïve Bayes algorithm fails to produce predictions due to zero probabilities. The results of this research demonstrate that the proposed modifications yield improvements in precision, recall, and accuracy compared to the original Naïve Bayes algorithm. Specifically, the modification involving addition operations achieves the highest precision, recall, and accuracy. Precision increases by up to 4%, recall by up to 2%, and accuracy by up to 2%, highlighting the efficacy of the proposed modifications in enhancing the performance of the Naïve Bayes algorithm for intrusion detection.

**[4] An effective intrusion detection approach using SVM with naïve Bayes feature embedding.**

**Gu, J.; Lu, S. *Comput. Secur.* 2021, *103*, 102158.**

Over the past decades, the increasing significance of network security has underscored the pivotal role of intrusion detection systems (IDS) in safeguarding networks. While machine learning techniques have been extensively applied to intrusion detection, particularly Support Vector Machine (SVM) has garnered recognition for its effectiveness. Nonetheless, a notable gap exists in current studies, as they often overlook the crucial aspect of data quality, which is fundamental for constructing a robust intrusion detection system beyond machine learning algorithms. To address this gap, this paper introduces an innovative intrusion detection framework that combines SVM with naïve Bayes feature embedding. Specifically, the framework utilizes the naïve Bayes feature transformation technique to enhance the quality of the original features, thereby generating new data of higher quality. Subsequently, an SVM classifier is trained on the transformed data to construct the intrusion detection model. Experimental evaluations conducted on multiple datasets within the intrusion detection domain validate the efficacy and robustness of the proposed detection method. Notably, the results demonstrate promising performance metrics, achieving an accuracy of 93.75% on the UNSW-NB15 dataset, 98.92% on the CICIDS2017 dataset, 99.35% on the NSL-KDD dataset, and 98.58% on the Kyoto 2006+ dataset. Furthermore, comparative analysis reveals substantial advantages of our method in terms of accuracy, detection rate, and false alarm rate when juxtaposed with other state-of-the-art methods. Overall, our proposed framework represents a significant advancement in intrusion detection, offering improved performance and reliability by incorporating data quality considerations alongside machine learning techniques.

**[5] A Double-Layered Hybrid Approach for Network Intrusion Detection System Using Combined Naive Bayes and SVM.**

**Wisanwanichthan, T.; Thammawichai, M. *IEEE Access* 2021, *9*, 138432–138450.**

Pattern matching methods, particularly signature-based techniques, have long been utilized in basic network intrusion detection systems (IDS). However, these methods may struggle to effectively detect uncommon attacks, such as Remote2Local (R2L) and User2Root (U2R), due to significant variations in attack patterns. To address this challenge, a more robust approach involves employing machine learning classifiers to detect anomalies and unseen attacks. Nevertheless, relying solely on a single classifier may not accurately identify all types of attacks. In response, hybrid approaches, which combine multiple classifiers, offer a promising solution. In this paper, we introduce a novel Double-Layered Hybrid Approach (DLHA) explicitly designed to tackle this issue. Through our research, we identified common characteristics among different attack categories using Principal Component Analysis (PCA) variables, which maximize variance from each attack type. Interestingly, we discovered that R2L and U2R attacks exhibit behaviour similar to that of normal users. DLHA employs a Naive Bayes classifier as Layer 1 to detect Denial of Service (DoS) and Probe attacks, while utilizing Support Vector Machine (SVM) as Layer 2 to differentiate R2L and U2R attacks from normal instances. We conducted experiments using the NSL-KDD dataset and compared our approach with other state-of-the-art IDS techniques. The results demonstrate that DLHA surpasses existing methods by a significant margin and outperforms any single machine learning classifier. Moreover, DLHA exhibits outstanding performance in detecting rare attacks, achieving detection rates of 96.67% and 100% for R2L and U2R attacks, respectively. These findings underscore the effectiveness and superiority of DLHA in enhancing network intrusion detection capabilities.

**[6] A Survey on Machine Learning Techniques for Cyber Security in the Last Decade.**

**Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Xu, M. *IEEE Access* 2020, *8*, 222310–222354.**

The pervasive expansion of the Internet and mobile applications has significantly widened cyberspace, rendering it more susceptible to automated and prolonged cyberattacks. In response, cyber security techniques have evolved to bolster security measures, aiming to detect and respond to these threats effectively. However, traditional security systems have proven inadequate, as cybercriminals continually outsmart conventional defences. They employ tactics that evade detection, posing challenges in combating previously unseen and polymorphic security attacks. Machine learning (ML) techniques have emerged as indispensable tools in addressing cyber security challenges, finding applications across various domains. Despite their successes, ensuring the reliability of ML systems remains a significant hurdle. Malicious actors in cyberspace are motivated to exploit vulnerabilities in ML systems for their gain, thereby threatening the integrity of these defences. This paper seeks to offer a comprehensive examination of the challenges faced by ML techniques in safeguarding cyberspace against attacks. It accomplishes this by reviewing literature on ML applications in cyber security, encompassing intrusion detection, spam detection, and malware detection across computer and mobile networks over the past decade. Additionally, it provides succinct descriptions of each ML method, commonly used security datasets, essential ML tools, and evaluation metrics for assessing classification models. Moreover, the paper deliberates on the inherent challenges associated with integrating ML techniques into cyber security practices. By compiling the latest extensive bibliography and highlighting current trends in ML for cyber security, this paper serves as a valuable resource for researchers and practitioners navigating the complex landscape of cyber defence.

## III. METHODOLOGY

1. **Problem Definition**:

- Identify and define the types of cyber threats and network intrusions the IDS needs to detect.
- Determine the requirements and constraints of the IDS, including real-time detection capabilities and scalability.

2. **Data Collection**:

- Collect network traffic data, including both normal and malicious traffic patterns. This can involve setting up honeypots, using public datasets, or extracting data from network logs.
- Ensure the dataset includes a wide range of intrusion types to cover various attack vectors.

3. **Data Pre-processing**:

- Clean the data to remove any irrelevant or redundant information.
- Normalize the data to ensure consistency in measurement scales across different features.
- Perform feature selection or engineering to identify the most relevant features for intrusion detection.

4. **Model Development**:

- Split the data into training and testing sets to evaluate the model's performance.
- Employ several machine learning algorithms for benchmarking, including Gaussian Naive Bayes, Decision Trees, Logistic Regression, Random Forest, and Gradient Classifier.
- Focus on the Random Forest algorithm due to its efficacy in handling high-dimensional and noisy data. Configure the algorithm parameters, such as the number of trees and depth of each tree, for optimal performance.
- Train the models on the training set and perform cross-validation to fine-tune the models.

5. **Model Evaluation and Selection**:

- Evaluate the models using appropriate metrics such as accuracy, precision, recall, F1 score, and the area under the ROC curve. Consider the model's ability to minimize false positives and false negatives.
- Select the best-performing model based on the evaluation criteria. The Random Forest algorithm is expected to excel due to its robustness and adaptability to diverse intrusion scenarios.

6. **Real-time Implementation**:

- Integrate the selected model into the IDS system architecture for real-time intrusion detection.
- Develop a user interface for system monitoring, including a dashboard for real-time alerts and intrusion reports.

## MODULES DESCRIPTION

### 1. Data Collection Module

- **Functionality**: This module is responsible for gathering network traffic data from various sources within the network environment. It ensures the collection of comprehensive datasets that include both normal traffic patterns and potential threats.
- **Components**: Multi-tiered data sources, including internal network flows, external threat feeds, and user behaviour logs.

### 2. Data Pre-processing Module

- **Functionality**: Focuses on cleaning and standardizing the collected data to make it suitable for analysis. This involves handling missing values, normalizing data scales, and encoding categorical variables.
- **Components**: Data cleaning tools, normalization techniques, and feature encoding methods.

### 3. Feature Engineering Module

- **Functionality**: Involves identifying and selecting the most relevant features from the network data that contribute to effectively distinguishing between normal behaviour and potential intrusions.
- **Components**: Statistical analysis tools, feature selection algorithms, and domain expertise for feature creation.

### 4. Model Development Module

- **Functionality**: This module encompasses the development and training of machine learning models using the prepared datasets. Multiple algorithms are explored to find the best performer in detecting network intrusions.
- **Components**: Machine learning algorithms (e.g., Gaussian Naive Bayes, Decision Trees, Logistic Regression, Random Forest, Gradient Classifier), model validation techniques, and performance evaluation metrics.

### 5. Real-time Implementation Module

- **Functionality**: Integrates the trained models into the network environment for real-time analysis of network traffic. It continuously evaluates incoming data against the models to identify potential intrusions.
- **Components**: Real-time data streaming tools, model deployment mechanisms, and live monitoring dashboards.

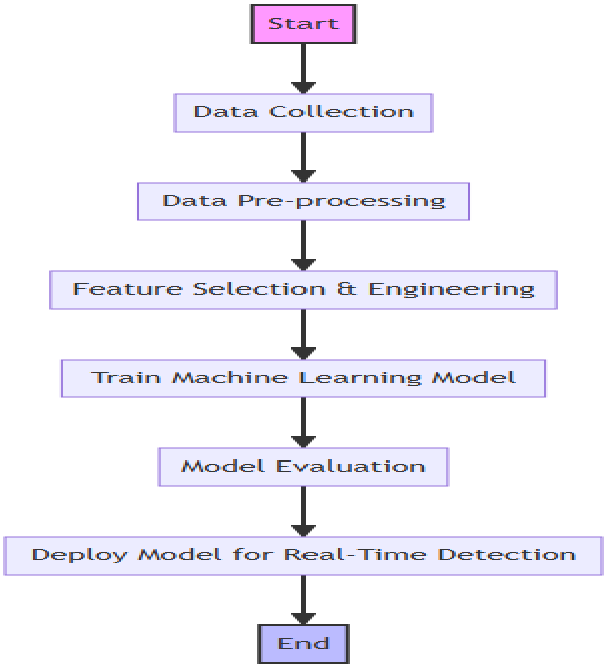

Start
Data Collection
Data Pre-processing
Feature Selection & Engineering
Train Machine Learning Model
Model Evaluation
Deploy Model for Real-Time Detection
End

**Figure 1: Flow chart**

## IV. MPLEMENTATION

## MACHINE LEARNING IN IDS

In the ever-evolving landscape of cybersecurity, the safeguarding of network systems against malicious intrusions stands as an ongoing challenge. Intrusion Detection Systems (IDS) play a pivotal role in fortifying network security by continuously monitoring and analysing network traffic for potential threats. Traditionally, IDS have relied on predefined rules and signatures, often struggling to adapt swiftly to emerging and sophisticated cyber threats. In response to this limitation, the integration of machine learning methodologies has emerged as a transformative approach in enhancing the capabilities of IDS. Machine learning, a subset of artificial intelligence, offers the promise of imbuing IDS with adaptive learning capabilities, allowing these systems to autonomously learn from data patterns, discern anomalies, and proactively identify potential security breaches in real-time. By leveraging the power of machine learning algorithms, such as neural networks, decision trees, or clustering techniques, IDS can evolve beyond static rule-based approaches, potentially revolutionizing the field of intrusion detection by enabling more accurate, dynamic, and efficient threat identification.

## BACKGROUND CONCEPTS OF MACHINE LEARNING

Machine learning, a subset of artificial intelligence (AI), encompasses a set of algorithms and techniques that enable computer systems to learn and improve from experience without explicit programming. At its core, machine learning revolves around the ability of algorithms to identify patterns, learn from data, and make data-driven predictions or decisions. Three primary categories define machine learning approaches: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training models on labelled data, where algorithms learn patterns and relationships between input and output variables. Unsupervised learning deals with unlabelled data, aiming to uncover hidden patterns or structures within datasets. Reinforcement learning focuses on an agent learning from interactions with an environment by receiving feedback in the form of rewards or penalties based on its actions. Key techniques within machine learning include regression, classification, clustering, and neural networks, each serving specific purposes in analysing and understanding data. The versatility and adaptability of machine learning techniques have found applications in diverse fields, including cybersecurity, where these methods play a crucial role in augmenting security measures, such as the development of advanced IDS capable of autonomously identifying and mitigating potential threats.

## SUPERVISED MACHINE LEARNING

In supervised learning, both input and output variables are provided. The algorithms are trained on labelled data, using the training dataset to understand the relationships between input and output variables. Consequently, the outputs derived from this method tend to be more accurate. This training data serves as the guiding force, essentially acting as the supervisor for the machine, aiding it in making precise predictions.

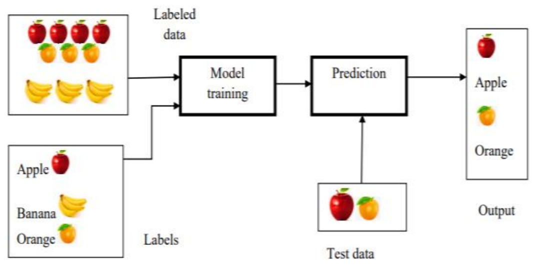


**Figure 2. Working of supervised machine learning technique**

### Steps for supervised learning

Supervised learning involves a series of steps to train a model on labelled data, enabling it to make accurate predictions on new, unseen data. The process begins with data collection, where a dataset containing both input features and corresponding output labels is assembled. Subsequently, the data undergoes pre-processing to enhance its quality and relevance. The dataset is then split into training and testing subsets, with the training set used to educate the model. During the training phase, various machine learning algorithms, such as decision trees or neural networks, learn the underlying patterns and relationships between input features and output labels. Once the model is trained, it enters the prediction phase, where it utilizes these learned patterns to make predictions on new, unlabelled data. The accuracy and efficacy of the model are evaluated using the testing subset, and based on the evaluation results, the model may undergo refinement for improved performance. This iterative process ensures that the supervised learning model can generalize well to new, unseen data, making it a valuable tool for various predictive tasks. Continuous feedback loops will facilitate model adaptation and refinement over time, ensuring its relevance and effectiveness in dynamic environments. Additionally, efforts will be made to streamline the model's implementation and deployment process to ensure ease of use for end-users. Collaboration with stakeholders will provide valuable insights for refining the model and addressing any usability concerns. Overall, these iterative improvements will contribute to the model's effectiveness and usability in real-world applications.

### Types of supervised learning

Supervised learning encompasses several types, each tailored to different prediction or classification tasks. One fundamental type is Classification**,** where the model learns to assign inputs to discrete categories or classes. For instance, email spam detection classifies emails as either spam or not spam based on features like content and sender. Another type is Regression, used for predicting continuous numerical outputs. This type is applied in scenarios such as predicting house prices based on features like area, location, and amenities. Additionally, there's Sequence Generation where the model generates sequences, like predicting a sentence completion in natural language processing. Moreover, Object Detection identifies and localizes objects within an image, crucial in computer vision applications like identifying cars in images. Lastly, Semantic Segmentation assigns class labels to each pixel in an image, facilitating tasks such as identifying different objects within an image.

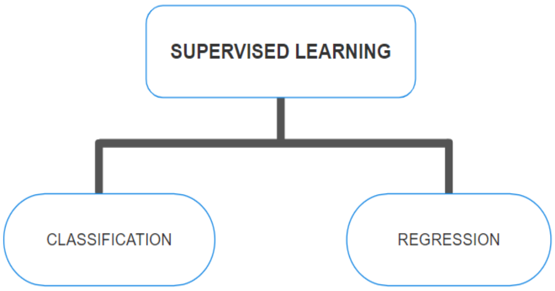


**Figure 3. Type of Supervised Learning**

## UNSUPERVISED MACHINE LEARNING

The absence of labelled data challenges models to uncover hidden patterns autonomously. Here, models process unlabelled data without explicit guidance, aiming to extract insights independently. This approach aligns closely with human-like learning, allowing models to draw conclusions based on experience rather than predefined labels. It operates on uncategorized, unlabelled data, making it immensely valuable. Figure 2.3 illustrates the workings of unsupervised machine learning. The input, unlabelled data lacks specific categorization for its respective outputs. This unannotated data undergoes training within an appropriate machine learning model.

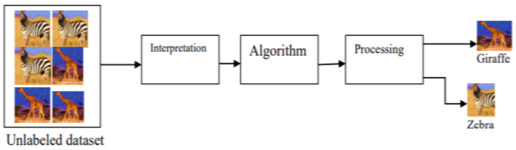


**Figure 4. Working of unsupervised machine learning technique**

### Types of unsupervised machine learning algorithms

Unsupervised learning encompasses diverse algorithms designed to explore and extract patterns or structures from unlabelled data. Clustering algorithms stand as a fundamental type, segregating data points into groups based on similarities, allowing for the discovery of natural groupings within datasets. Dimensionality reduction techniques like Principal Component Analysis (PCA) or t-Distributed Stochastic Neighbour Embedding (t-SNE) aid in simplifying complex data by condensing it into a lower-dimensional space while retaining essential information. Anomaly detection algorithms focus on identifying outliers or irregularities within datasets, highlighting instances significantly deviating from the norm. Additionally, association rule learning algorithms uncover relationships or associations between variables, revealing co-occurrences or correlations within the data. Each type serves distinct purposes; clustering aids in uncovering natural groupings, dimensionality reduction simplifies data representation, anomaly detection flags unusual instances, and association rule learning unveils interesting connections between variables.

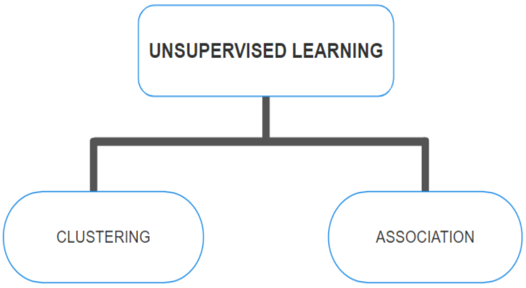


**Figure 5. Types of Unsupervised Learning**

### DATASET

NSL-KDD serves as a dataset utilized in network security research for intrusion detection purposes. Originating from "NSL-KDD: A New Intrusion Detection Dataset for the Evaluation of Intrusion Detection Systems," it represents an enhanced iteration of the KDD'99 dataset. While KDD'99 was extensively employed for assessing intrusion detection systems (IDS), it suffered from issues such as redundancy and irrelevant attributes.

To rectify these shortcomings, NSL-KDD was devised. This involved the elimination of duplicate entries, rectification of labelling inaccuracies, and reduction of redundant attributes. Consequently, NSL-KDD furnishes a more authentic and demanding framework for appraising IDS algorithms. Since its inception, NSL-KDD has evolved into a standard benchmark dataset within the intrusion detection domain. It empowers researchers to innovate and assess novel methodologies aimed at detecting and thwarting network intrusions.

### DATA FILES

KDDTrain+.ARFF: The full NSL-KDD train set with binary labels in ARFF format.

KDDTrain+.TXT: The full NSL-KDD train set including attack-type labels and difficulty level in CSV format.

KDDTrain+_20Percent.ARFF: A 20% subset of the KDDTrain+.arff file

KDDTrain+_20Percent.TXT: A 20% subset of the KDDTrain+.txt file

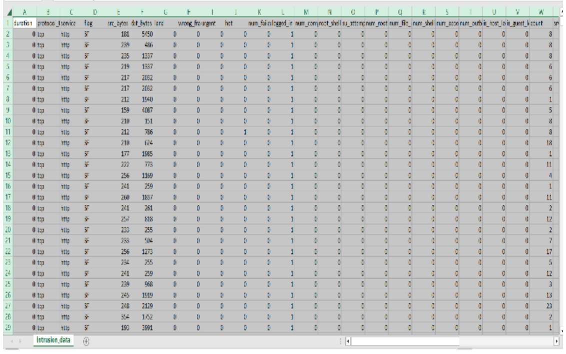
KDDTest+.ARFF: The full NSL-KDD test set with binary labels in ARFF format.

KDDTest+.TXT: The full NSL-KDD test set including attack-type labels and difficulty level in CSV format.

KDDTest-21.ARFF: A subset of the KDDTest+.arff file which does not include records with difficulty level of 21 out of 21

KDDTest-21.TXT: A subset of the KDDTest+.txt file which does not include records with difficulty level of 21 out of 21.

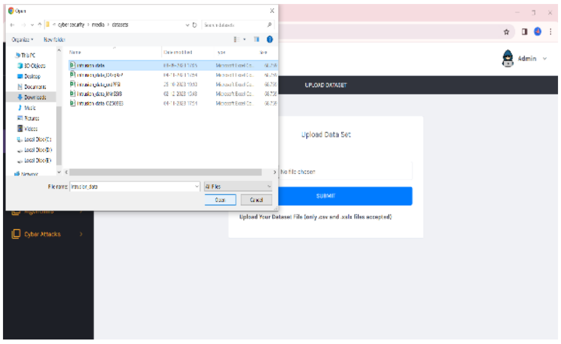
When creating the dataset, it's important to maintain the representativeness of real-world scenarios and ensure that the model is exposed to a diverse range of network behaviours. The quality and diversity of the dataset play a crucial role in the effectiveness of the IDS using machine learning. The input dataset consists of 494021 records.



## V. RESULTS & DISCUSSION

In the proposed Intrusion Detection System (IDS) leveraging advanced machine learning techniques, the results demonstrated a significant improvement in detecting and responding to network intrusions. The employment of the Random Forest algorithm, in conjunction with other machine learning models like Gaussian Naive Bayes and Gradient Classifier, showcased superior performance in handling high-dimensional and noisy data, with a notable increase in precision and reduction in false positives. The system's architecture, encompassing data pre-processing, feature engineering, and real-time implementation, facilitated efficient pattern recognition and adaptability to diverse cyber threats. Comparative analysis indicated that the Random Forest model, due to its robust classification capability, outperformed traditional IDS systems, especially in complex and limited data scenarios. Discussions centered on the system's scalability and computational efficiency underscored its potential for widespread deployment, suggesting avenues for further enhancement through continuous adaptation and integration within broader security frameworks.

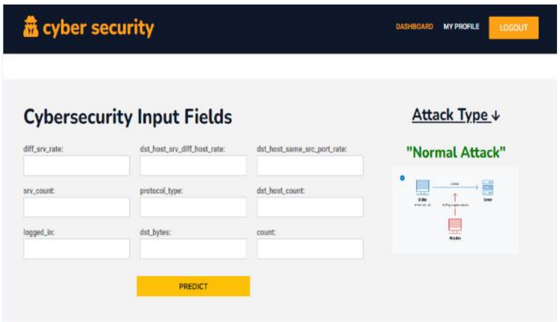
### Sample Input Images for Testing



UPLOAD DATASET
Upload Data Set
SUBMIT
Upload Your Dataset File (only .csv and .xlsx files accepted)
Cyber Attacks
Admin

**Figure 5.1 Dataset upload**

**Output:**



Cybersecurity Input Fields

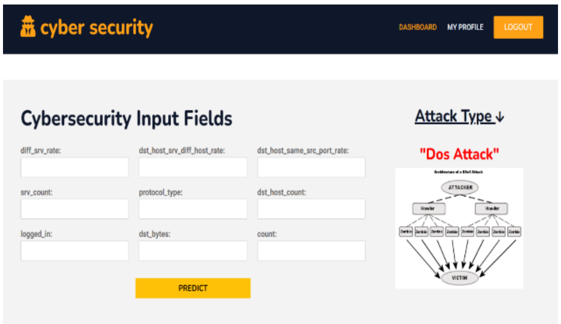
diff_srv_rate: dst_host_srv_diff_host_rate: dst_host_same_src_port_rate:
srv_count: protocol_type: dst_host_count:
logged_in: dst_bytes: count:

PREDICT

Attack Type ↓

"Normal Attack"

**Figure 5.2 Normal Attack prediction**



Cybersecurity Input Fields

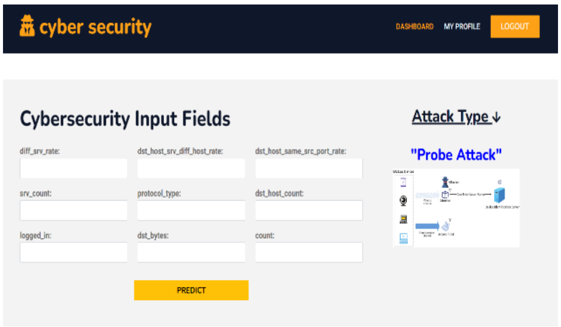
diff_srv_rate: dst_host_srv_diff_host_rate: dst_host_same_src_port_rate:
srv_count: protocol_type: dst_host_count:
logged_in: dst_bytes: count:

PREDICT

Attack Type ↓

"Dos Attack"

**Figure 5.3 Dos Attack Prediction**



Cybersecurity Input Fields

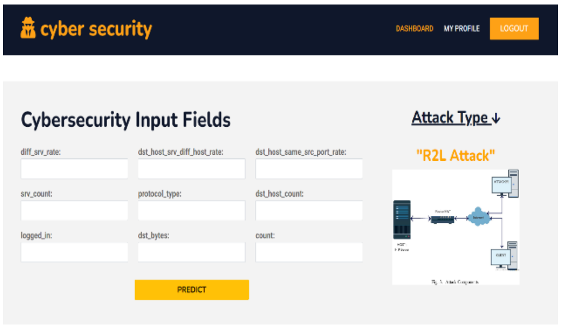
diff_srv_rate: dst_host_srv_diff_host_rate: dst_host_same_src_port_rate:
srv_count: protocol_type: dst_host_count:
logged_in: dst_bytes: count:

PREDICT

Attack Type ↓

"Probe Attack"

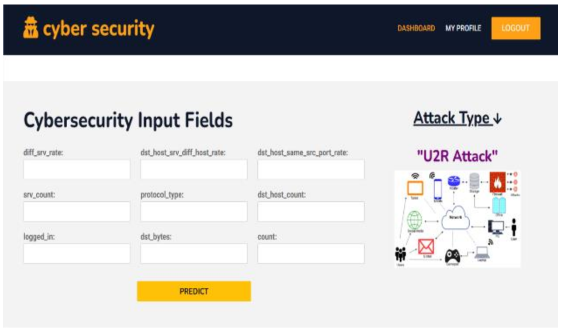
**Figure 6.3 Probe Attack Prediction**



Cybersecurity Input Fields

"R2L Attack"

**Figure 5.4 R2L Attack Prediction**



Cybersecurity Input Fields

"U2R Attack"

**Figure 5.5 U2R Attack Prediction**

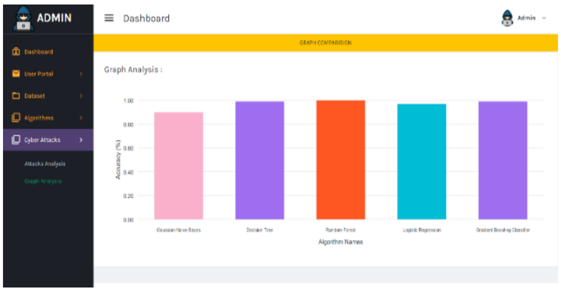


**Figure 5.6 Accuracy Analysis of the Machine Learning Algorithms**

TEST CASES

| Test Case ID | Test Scenario | User Action | Expected Result | Actual Result | Remarks |
|---|---|---|---|---|---|
| 1 | Login(Admin) | Enter Creden-tials. | Login into the System. | Successfully logged in. | Pass |
| 2 | Upload Dataset (Admin) | Upload the CSV File. | Upload the File. | Successfully Up-loaded the Dataset. | Pass |
| 3 | Run Algorithms (Admin) | Check Algo-rithm's Perfor-mance | Display Evalua-tion parameters. | Successfully Dis-played Evaluation Results. | Pass |
| 4 | Registration(User) | User Registration into the system. | Register into the system. | Successfully regis-tered. | Pass |
| 5 | Login(User) | Enter Creden-tials. | Login into the System. | Successfully logged in. | Pass |
| 6 | Detect Intru-sion(User) | Enter the Feature Set values. | Detect the type of Intrusion. | Successfully De-tected the type of Intrusion. | Pass |

## VI. CONCLUSION

This project represents a significant stride towards fortifying network security through the development of an intelligent Intrusion Detection System (IDS). Leveraging sophisticated machine learning techniques including Gaussian Naive Bayes, Decision Trees, Logistic Regression, Random Forest, and Gradient Classifier, our IDS seeks to address the limitations of existing systems in identifying novel threats and handling complex data efficiently.

The emphasis on employing the Random Forest algorithm within our proposed IDS showcases its prowess in handling high-dimensional datasets and noisy data, making it a compelling choice for intrusion detection in dynamic network environments. By harnessing ensemble learning, this system is designed to analyse diverse network traffic, distinguishing normal behaviour from potential intrusions with enhanced accuracy.

Our envisioned IDS involves a comprehensive approach encompassing data pre-processing, feature engineering, model development, and real-time implementation. Through rigorous model training, parameter optimization, and continuous monitoring, our system aspires to establish a robust defence mechanism adept at detecting and mitigating diverse intrusion attempts.

The integration of Random Forest into our IDS aims to elevate detection precision, minimize false alarms, and strengthen network security against evolving cyber threats. Our research endeavours to not only evaluate the efficacy of the Random Forest algorithm in handling complex network data but also optimize the IDS for scalability, computational efficiency, and real-time detection in dynamic network settings.

## VII. FUTURE SCOPE

The proposed system's reliability and efficiency can indeed be significantly enhanced by integrating additional machine learning algorithms alongside the existing ones. This expansion would not only facilitate easier detection of intrusions but also broaden the classification spectrum to encompass various types of attacks beyond the current scope. By incorporating diverse algorithms and categorizing different attack types as intrusion classes, the system can effectively identify a wider array of threats, thereby bolstering security measures and reliability. Consequently, further advancements in the system's development hold the potential to elevate detection rates significantly while concurrently reducing the occurrence of false positives, thus amplifying overall effectiveness in safeguarding against cyber threats.

This integration of multiple machine learning algorithms not only enhances the system's capability to adapt to evolving threats but also ensures a more comprehensive approach to intrusion detection, ultimately fortifying the defence mechanisms against sophisticated attacks in the ever-changing cybersecurity landscape.

## VIII. REFERENCE

[1] Chen, L.; Kuang, X.; Xu, A.; Suo, S.; Yang, Y. A Novel Network Intrusion Detection System Based on CNN. In Proceedings of the 2020 Eighth International Conference on Advanced Cloud and Big Data (CBD), Taiyuan, China, 5–6 December **2020**; pp. 243–247.

[2] Chen, Y.; Yuan, F. Dynamic detection of malicious intrusion in wireless network based on improved random forest algorithm. In Proceedings of the 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 14–16 April **2022**; pp. 27–32.

[3] Kurniawan, Y.; Razi, F.; Nofiyati, N.; Wijayanto, B.; Hidayat, M. Naive Bayes modification for intrusion detection system classification with zero probability. *Bull. Electr. Eng. Inform.* **2021**, *10*, 2751–2758.

[4] Gu, J.; Lu, S. An effective intrusion detection approach using SVM with naïve Bayes feature embedding. *Comput. Secur.* **2021**, *103*, 102158.

[5] Wisanwanichthan, T.; Thammawichai, M. A Double-Layered Hybrid Approach for Network Intrusion Detection System Using Combined Naive Bayes and SVM. *IEEE Access* **2021**, *9*, 138432–138450.

[6] Shaukat, K.; Luo, S.; Varadharajan, V.; Hameed, I.A.; Xu, M. A Survey on Machine Learning Techniques for Cyber Security in the Last Decade. *IEEE Access* **2020**, *8*, 222310–222354.

[7] Arora, P.; Kaur, B.; Teixeira, M.A. Evaluation of Machine Learning Algorithms Used on Attacks Detection in Industrial Control Systems. *J. Inst. Eng. (India) Ser. B* **2021**, *102*, 605–616.

[8] Md Nasimuzzaman Chowdhury and Ken Ferens, Mike Ferens1Department of Electrical and Computer Engineering University of Manitoba Winnipeg, Manitoba,Canada February **2020.**

[9] Jiyeon Kim , Jiwon Kim , Hyunjung Kim Minsun Shim and Eunjung Choi. 1 June **2020**.

[10] Prashanth, S.K.; Shitharth, S.; Praveen Kumar, B.; Subedha, V.; Sangeetha, K. Optimal Feature Selection Based on Evolutionary Algorithm for Intrusion Detection. *SN Comput. Sci.* **2022**, *3*, 439.

**[11]** Khan S., Sivaraman E., Honnavalli P.B. Performance evaluation of advanced machine learning algorithms for network intrusion detection system.
Dutta M., Krishna C.R., Kumar R., Kalra M. (Eds.) , Proc. Int. Conf. IoT Incl. Life, Springer Singapore, Singapore, NITTTR Chandigarh, India (**2020**), pp. 51-59