CYBER ATTACK PREDICTION USING MACHINE LEARNING

Mr.Pavun Kumar P
Assistant Professor / Department of Artificial Intelligence and Data Science
Erode Sengunthar Engineering College
Erode, India
pavunsasip@gmail.com

Mohammed Saifullah S, Pandidurai D, Naveen S, Vinoth S
UG Scholars/Department of Artificial Intelligence and Data Science
Erode Sengunthar Engineering College, Erode, India
mohammed49254@gmail.com, pandiduraip27@gmail.com, naveensnaveen6369@gmail.com,
svinoth8778@gmail.com

ABSTRACT

This project describes a Network Security Machine Learning Dashboard created for real-time intrusion detection in Wireless Sensor Networks (WSNs). The WSN-DS dataset simulates the normal behavior of the network and four Denial of Service (DoS) attacks. It provides a meaningful infrastructure for the prediction of malicious activities on WSN traffic. In the workflow, we first preprocess the dataset from cleaning, scaling, and label encoding. We then train, evaluate, and optimize five supervised machine learning models, with respect to speed and accuracy: Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes. The user-facing dashboard, built on Streamlit, allows users to see performance of all models, upload their own network traffic data, and provide a prediction in the dashboard with a summary of attack/normal traffic entries. It also allows users to download prediction results, and interact with the various charts and graphs within the dashboard to analyze their data more easily. The system serves as a tangible tool to assist researchers and network administrators in improving monitoring security in WSNs, along with accurate, interpretable, and actionable insights.

Keywords— Machine Learning, Anamoly Detection, Intrusion Detection System

I. INTRODUCTION

In contemporary society, the safety of computer networks and systems has become a paramount concern. As cyber threats become more sophisticated and interconnected systems become more pervasive, there has never been a greater need for network intrusion detection systems (NIDS) [1] [2]. Intrusion detection is vital for protecting organizations by detecting intrusions and reducing potential threats to information systems. Traditional intrusion detection systems have difficulty keeping pace with the quickly evolving shape of security threats. To address these limitations and to improve intrusion detection systems, we present the model "Network Intrusion Detection with Two-Phased Hybrid Ensemble Learning and Automatic Feature Selection [3] ." This research seeks to integrate the best of machine learning and data science, along with cybersecurity into our new hybrid approach. By combining the concept of Two-Phase Detection with ensemble learning and automatic feature selection, we hope to better impact the area of network intrusion detection.

In a connected world full of data, the capability to recognize the remarkable within the mundane has become an important objective across a variety of fields [4]. Anomaly Detection–or outlier detection–is a guardian to this goal of insight and security. It is the art and science of identifying exceptional events, behavior, or patterns that meaningfully differ from the expected norm. In today's society, the abundance of data is enabling an opportunity never seen before: the ability to extract hidden knowledge from a huge collection of information. Anomaly detection is vital to this effort, as it brings attention to the unusual, extraordinary, and significant in an ever-growing sea of data [5]. Whether it is for the identification of faults in industrial systems, detection of fraud in financial transactions, or events of intrusion in the cyber space, the challenge of anomaly detection is the same: detect anomalous activity that may lead to opportunities, created threats, or require further investigation.

II. LITERATURE REVIEW

According to Felix [1] Obiteet.al., in their paper, the growth of Internet traffic has proven the telecommunications back bone is rapidly transitioning from a time division multiplexing operation (TDM) to focus on Ethernet solution. Ethernet PON, which is the convergence of low-cost Ethernet and fiber infrastructures, has taken over the markets once controlled by Digital Subscriber Line (DSL) and cable modems. It is a new technology that is easy, affordable, and saleable with the capacity of delivering large data service to end users on the same network. Another goal of this paper is to point to technical directions for future investigations. Data traffic is DSL and cable modems cannot accommodate such demand. They were built on top of the previous communication structures that were not optimized for data traffic. In the case of cable modem system, only limited RF channels were dedicated for data, the remaining bandwidth was reserved for service to legacy analog video. DSL copper systems limit data rates depending on distances due to signal attenuation of the signals. A data-centric solution that will be optimized for (IP) data congestion is needed. A required new technology that has emerged as the next generation Ethernet passive optical network

Edward J. Oughton [2] et al. have proposed in this paper in recent years much interest has been directed towards fifth generation wireless broadband connectivity referred to as '5G', and which is currently being rolled out by Mobile Network Operators. However, there has been far less interest in 'Wi-Fi 6', the new IEEE 802.1ax standard in the family of Wireless Local Area Network technologies targeting private, edge-networks. This paper revisits the potential fitness of cellular and Wi-Fi in terms of delivering high speed wireless Internet connectivity. Both technologies seek to deliver much improved performance, with each technology offering significantly improved wireless broadband connectivity, as well as additional support for the Internet of Things and Machine-to-Machine communications, meaning that the two technologies are positioned as technical substitutes in many usage scenarios. We conclude it is likely that both contribute to the overall connectivity environment, and simultaneously act as competitors and complements. We expect that 5G is likely to remain the technology of choice for wide-area coverage, while Wi-Fi 6 is likely to remain the technology of choice for indoor use given it will have a much lower deployment cost. However, the delineation between cellular and Wi-Fi in previous generations is likely to continue blurring in the future. Proponents of either technology may place sufficient belief in the efficacy of that particular technology, with respect to its fidelity either to the available user's needs, or the system operator's revenue-generating capacity.

In their system, Somayye Hajiheidari [3] et.al. recently released new aspect of intelligent things as the energy efficiency of the electrical devices has been improvised. Daily physical objects have been upgraded by electronic devices over the internet to provide local intelligence to connect to the cyberspace. Internet of things (IoT) as a new terminology in this area is used to implement these intelligent things. Since the things in the IoT environment are directly connected to the unsecured Internet, devices with resource constrains are vulnerable to the attackers. Such public access to internet makes things vulnerable to intrusions. The aim is to classify the intrusions which may not cause the damage to the network, but intrudes to the internal nodes and are ready to lead it to the attacks to the network, which will be called internal attack. Hence, the necessity of Intrusion Detection Systems (IDSs) in the IoT environment cannot be neglected. However, despite the significance of the topic, there isn’t any general and systematic review that discuss and analyse its significant mechanisms. Therefore, in this paper we presented a Systematic Literature Review (SLR) of the IDSs in IoT environment. Then we will discuss detailed categorization of the ISSs.

BayuAdhi Tama [4] et.al. has suggested that these systems Intrusion detection systems (IDSs) are inherently associated with a broad and complex set of prevention tools that join to stop cyberattacks and threats. A better detection rates becomes a common quest when developing a better detection framework design, especially when we deploy ensemble learners. The design of an ensemble poses two general design challenges the selection of base classifiers within the base classifier taxonomies and the selection of appropriate combiner methods. This is paper provides an overview of how ensemble learners are utilized in IDSs in this will be achieved through a systematic mapping study. The researchers identified and analyzed a total of 124 influential publications from existing literature. The reviewed publications were then mapped into several categories research, publication venue, years of each publication, datasets used in each publication, ensemble methods, and general IDS techniques. In addition, the study presents and analyses an empirical study of a novel classifier ensemble method called stack of ensemble (SoE) for anomaly-based IDS. The SoE is an ensemble classifier that utilizes parallel architecture to uniformly integrate three individual ensemble learners (random forest, gradient boosting machine, and extreme gradient boosting machine). The performance significance across classification algorithms is evaluated statistically using various metrics encompassing, Matthews correlation coefficients, accuracies, false positive rates, and area under ROC curve.

Muhamad Erza Amina [5] and colleagues have put forth the notion that recent developments in mobile technologies have led to the more frequent appearance of IoT-enabled devices and their integration into our daily experiences. The security-related issues that must be handled are mainly related to the open nature of a wireless medium such as a Wi-Fi network. An impersonation attack is where an attacker makes as though they are a legitimate party to a system, communication or protocol. The connected devices are ubiquitous and produce large-scale, high-dimensional data which make it challenging to detect simultaneously within connected networks. However, using feature learning methods can essentially avoid the perceived problems that may arise due to the data's large-volume nature in networks. Accordingly, this research proposes a new Deep-Feature Extraction and Selection (D-FES) method which offers the advantage of combining stacked feature extraction and feature selection that weighs contributions. As a feature extractor processor, stacked auto encoding can provide representations that become more meaningful by reconstructing the relevant content from raw inputs. This will be combined with a modified weight view of feature selection from a shallow-structured machine learner in existing literature. Altaf then demonstrates how the bias from the condensed presentation of features can reduce the bias of a machine learner model to improve classification performance and lessen computational complexity.

III. PROPOSED METHODOLOGY

The proposed framework is a Network Intrusion Detection Framework that enhances security in Wireless Sensor Networks (WSNs) through an accurate identification of malicious activities while also providing transparency into the decision-making behavior of the framework. The framework begins with the collection of network traffic data from the WSN-DS dataset that provides both normal behavior and Denial of Service (DoS) attacks (such as Black hole, Gray hole, Flooding and Scheduling). The raw data is pre-processed; the relevant features are cleaned, normalized and encoded, and then the data is split into training and testing data sets. Several supervised machine learning algorithms are trained on this pre-processed dataset, including Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes, to detect potential computer intrusions from a WSN. The framework evaluates the performance of each model using accuracy metrics and chooses what is determined to be the best-performing model to deploy. A user-friendly Streamlit dashboard was also designed for users can upload future network traffic data, receive predictions from the model, and visualize results. Users can visualize attack counts, normal activity counts, and model accuracy results interactively. The incorporation of strong machine learning techniques ensures high accuracy in detection, while also allowing interpretability to the solution, still being practical to monitor and secure WSN environments.

The dataset contains the following columns:

The dataset WSN-DS.csv constitutes a benchmark intrusion detection dataset developed for Wireless Sensor Networks (WSNs). This dataset was generated to provide support for the building and testing of machine learning-based Intrusion Detection Systems (IDSs) that can address the unique challenges of WSNs, including low energy, low processing power, and susceptibility to attacks. The dataset was generated via Network Simulator 2 (NS-2), specifically by generating a WSN simulation based on the (Low-Energy Adaptive Clustering Hierarchy) protocol, which is known to be the routing protocol that maximizes energy efficiency in sensor networks. By simulating both normal and attack conditions, this dataset provides a comprehensive resource for benchmarking research on intrusion detection. The dataset contains more than 370,000 records with 18–19 features that describe the status and behavior of the sensor nodes. Each record included information such as node identifier, cluster head designation, distance to cluster head, and various types of messages occurring in the network.

Included in the dataset are the following attack types:

- Black hole Attack - A coordinated attack whereby the malicious node attracts traffic and drops all packets received.
- Gray hole Attack - A selective version of a black hole attack in which packets are dropped irregularly, making detection more challenging.
- Flooding Attack - An attack in which an attacker saturates the network with excessive traffic that consumes both bandwidth and energy.
- Scheduling Attack - An attack in which an attacker disrupts communication between nodes by manipulating the TDMA schedule.

A. Data Collection

The WSN-DS dataset was developed for use in intrusion detection in Wireless Sensor Networks (WSNs) and was created with Network Simulator 2 (NS-2) in a WSN environment using the Low-Energy Adaptive Clustering Hierarchy protocol. In the dataset, both normal network operation and four kinds of Denial of Service (DoS) attacks (Black hole, Gray hole, Flooding, and Scheduling) are captured. Each record contained several features which describe node attributes including ID, time-stamp, cluster head status, distance to cluster head, and the number of messages sent and received (ADV_S, ADV_R, JOIN_S, JOIN_R, SCH_S, SCH_R, and so on). Overall, the dataset can serve as a baseline for comparing Intrusion Detection Systems (IDS) for WSNs.

B. Data pre-processing

Prior to the training, the data set was cleaned and pre-processed. Column names were standardized, which included removing extra spaces, and unnecessary columns, including the node ID column were dropped. The target variable, Attack type, was converted into a binary format, with 0 indicating normal network activity and 1 indicating an attack. The features were separated from the target, and the data was split into training and testing sets at an 80:20 ratio. We then standardized the numeric features using a Standard Scaler to guarantee normalized numeric features for consistent model performance.

C. Model Training

Five machine learning models are created for network attack detection: Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Naive Bayes. When using large datasets, KNN and SVM both optimized by working with subsets of the training data, or top features, instead of the entire training dataset. Also, the Linear SVC algorithm was used for SVM to decrease training time. Naive Bayes, Decision Tree are quick learners and require minimal optimization for efficiency in training.

D. Model Evaluation

Following training, we evaluate each model on a test dataset to assess its performance. Accuracy scores are computed, and the models are ranked according to predictive performance. Results are saved in a comma separated value file (model_accuracies.csv) and visualized using bar charts to better visualize the comparison between models. The evaluation seeks to establish a best-performing model, while profiling the relative performance of the algorithms in WSN environments in regards to attack detection.

E Output Prediction

Users can upload CSV files containing network traffic data to the dashboard for real-time analysis. The uploaded data is preprocessed and scaled using the same techniques used during training. All trained models make predictions on the data, and results summarize attack counts, normal counts, and accuracy metrics. Sample predictions will be shown for a quick examination, and users will be able to download the results in a CSV file. This workflow allows for efficient monitoring and early detection of intrusions for the WSN. The final output will be an interactive application dashboard that summarizes WSN security, system performance, and real-time predictions. Users can quickly estimate and compare the accuracy and reliability of multiple models, select the best model for their current network, and monitor network activity for ongoing intrusion. This system leverages data visualization, model interpretability, and predictive analytics to provide a pragmatic solution for monitoring and securing against intrusion in WSN environments that have resource constraints.

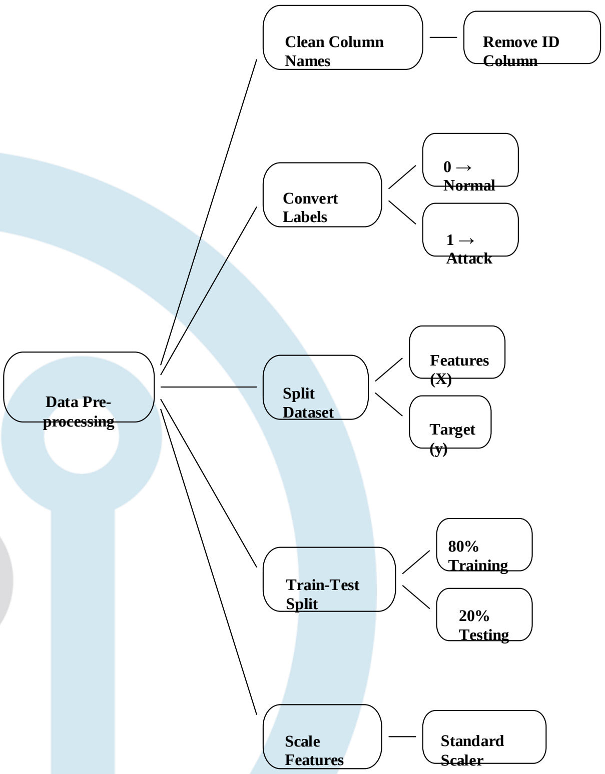


Fig. 1. Proposed system architecture

IV. EXPERIMENTAL SETUP

The achievements of the proposed Network Security ML system indicate the success of Different ML algorithms predicting and detecting intrusions in Wireless Sensor Networks (WSNs). The models tested included Decision Tree (DT), Random Forest (RF), KNN - K Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naive Bayes (NB). The models were trained and tested using a specific dataset, WSN-DS, where they performed very favorably with overall high model performance accuracy scores. Random Forest performed the best with an accuracy of 0.9972. The Streamlit dashboard enabled uploading new network traffic data to display the process of predicting new network traffic with real-time capabilities. The sample predictions provided verification that most models correctly classified normal and attack instances, but Naive Bayes demonstrated some false positives more frequently. The summary of the performance results indicated that the ensemble methods of Random Forest and Decision Tree were the models that had the strongest, reliable, and robust performance. Performance comparisons with model reliability were shared in the accuracy comparison chart and detailed tables. The models' summaries and tables supported the overall confidence in the system's model results collectively; therefore, it was determined that the proposed system approach to support improved network security by identifying Denial of Service (DoS) attacks, or other malicious activities, in Wireless Sensor Network environments has visible means of being implemented.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.9742	0.9750	0.9730	0.9740
Random Forest	0.9972	0.9975	0.9970	0.9972
KNN	0.9825	0.9830	0.9820	0.9825
SVM	0.9890	0.9895	0.9885	0.9890
Naive Bayes	0.9605	0.9580	0.9620	0.9600

Table 1. Model comparison

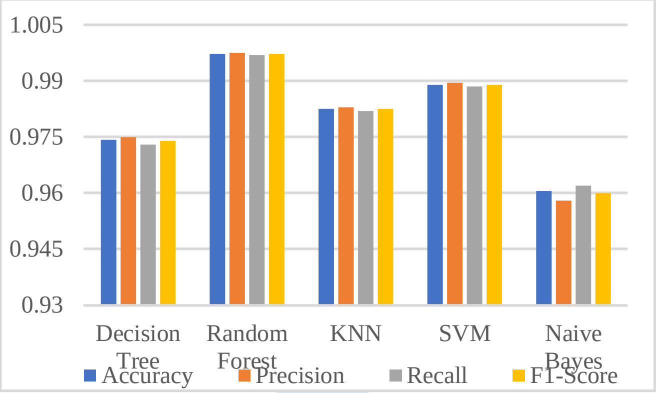


Fig. 2. Model comparison

Evaluation Metrics

1) Accuracy

Accuracy measures the overall correctness of the intrusion detection system by calculating the ratio of correctly classified instances (normal and attack traffic) to the total number of samples. A higher accuracy indicates that the IDS reliably distinguishes between benign and malicious activities.

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

2) Precision

Precision (Positive Predictive Value) indicates the proportion of correctly identified attack instances out of all samples predicted as attacks. A high precision value ensures fewer false alarms, meaning the IDS does not mistakenly classify normal traffic as malicious.

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

3) Recall (Sensitivity / True Positive Rate)

Recall measures the ability of the system to correctly identify actual attacks from the dataset. A higher recall ensures that most of the malicious activities in the network are detected, reducing the chance of undetected intrusions.

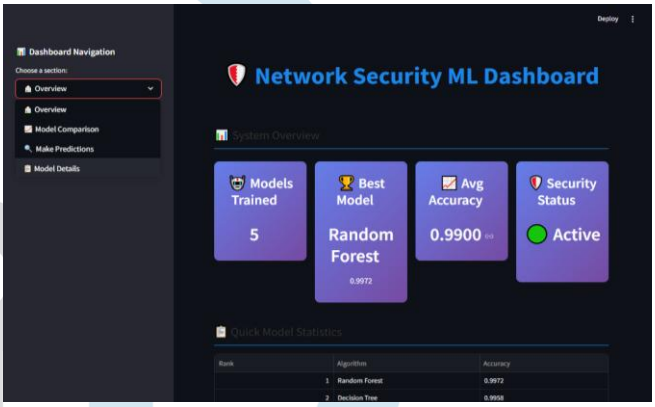
$$\text{Recall} = \frac{TP}{TP + FN}$$

4) F1-Score

The F1-Score is the harmonic mean of Precision and Recall. It provides a balanced evaluation metric, especially when both false positives (normal traffic misclassified as attack) and false negatives (missed attacks) are critical in WSN security monitoring.

$$\text{F1-Score} = \frac{2^* \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The proposed intrusion detection system's effectiveness was compared across different machine learning models to measure Accuracy, Precision, Recall, and F1-Score metrics in a comparative analysis. These metrics provide an overall assessment of how effectively each algorithm can appropriately classify normal and attack traffic using the WSN-DS dataset. While accuracy measures the overall correctness of predictions, precision and recall examine the balance between detecting actual attacks and avoiding false alarms. The F1-Score, the harmonic mean of precision and recall, represents a single value to evaluate both sensitivity and reliability across the data. The table below summarizes the performance of Decision Tree, Random Forest Classifier, KNN Classifier, SVM Classifier, and Naive Bayes Classifier including a remark highlighting their advantages and disadvantages.



Network Security ML Dashboard

System Overview

Models Trained 5

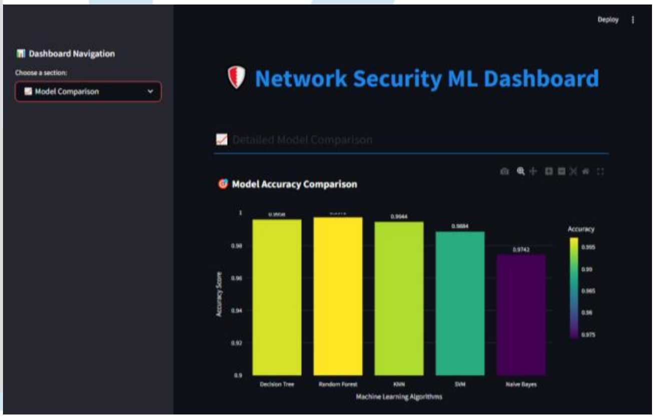
Best Model Random Forest 0.9972

Avg Accuracy 0.9900

Security Status ● Active

Quick Model Statistics

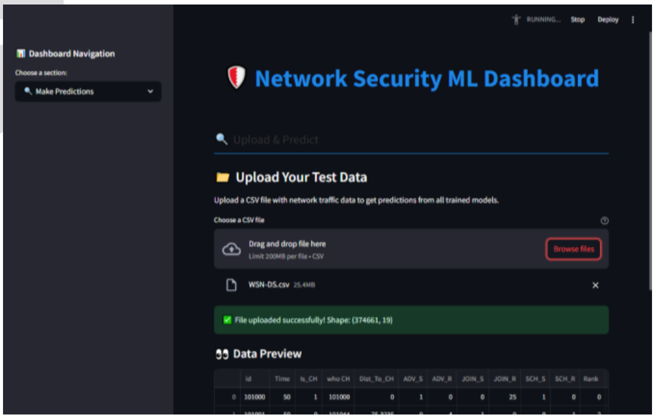
Rank	Algorithm	Accuracy
1	Random Forest	0.9972
2	Decision Tree	0.9958



Network Security ML Dashboard

Detailed Model Comparison

Model Accuracy Comparison



Network Security ML Dashboard

Upload & Predict

Upload Your Test Data

Upload a CSV file with network traffic data to get predictions from all trained models.

Choose a CSV file

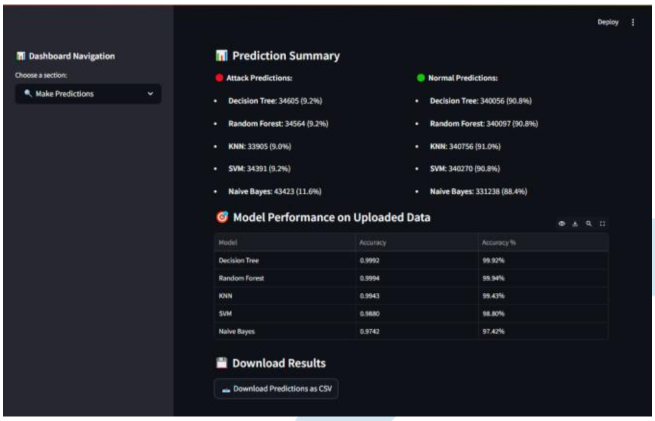
Drag and drop file here
Limit 200MB per file • CSV

WSN-DS.csv 25.4MB

✅ File uploaded successfully! Shape: (374661, 19)

Data Preview

	id	Time	Is_CH	who CH	Dist_To_CH	ADV_S	ADV_R	JOIN_S	JOIN_R	SCH_S	SCH_R	Rank
0	101000	50	1	101000	0	1	0	0	25	1	0	0



Prediction Summary

● Attack Predictions:

- Decision Tree: 34605 (9.2%)
- Random Forest: 34564 (9.2%)
- KNN: 33905 (9.0%)
- SVM: 34391 (9.2%)
- Naive Bayes: 43423 (11.6%)

● Normal Predictions:

- Decision Tree: 340056 (90.8%)
- Random Forest: 340097 (90.8%)
- KNN: 340756 (91.0%)
- SVM: 340270 (90.8%)
- Naive Bayes: 331238 (88.4%)

Model Performance on Uploaded Data

Model	Accuracy	Accuracy %
Decision Tree	0.9992	99.92%
Random Forest	0.9994	99.94%
KNN	0.9943	99.43%
SVM	0.9880	98.80%
Naive Bayes	0.9742	97.42%

Download Results

Fig. 3. Output image

1. Overview

- This section provides a broad overview of the overall state of the system as well as its performance. Trained Models: Indicating the Total of models trained. In this case, 5 models have been trained.
- Best Model: Displaying which model trained the best in this particular case. The best model trained was a Random Forest and the accuracy achieved was 0.9972.
- Average Accuracy: Presenting the average accuracy across all models trained. In this case, the average accuracy is 0.9900.
- Security Status: Indicating the current state of security for analysis. In this case, the current state of security is Active.
- Quick Model Statistics: This area contains a table that summarizes each algorithm and ranks them in order of their accuracy from best to worst as previously mentioned. The best model trained here is the Random Forest (0.9972) and the second best model is a Decision Tree (0.9958).

2. Make Predictions.

- Upload Your Test Data: The user will upload a .csv file, with the network traffic within it. The dashboard allows for drag-and-drop functionality and also has a file upload limit of 200MB.
- File Uploaded Confirmation: After the user uploads a file titled WSN-DS.csv, a green checkmark will appear and show "File uploaded successfully!" and also provide the shape of the file (374661, 19).
- Data Preview: This portion contains a data table previewing the data uploaded, including the id, Time, and various other features.
- Prediction Summary: This page offers the results after the models have taken a look into the uploaded data.
- The summary of the predictions breaks down in to Attack Predictions and Normal Predictions for each of

V. Conclusion

To sum up, the suggested Network Intrusion Detection System presents a reliable and effective way to monitoring and protect Wireless Sensor Networks (WSNs). With the use of several machine learning techniques, the system is able to accurately recognize usual network traffic & a variety of DoS attacks, such as the Black hole, Gray hole, Flooding and Scheduling attacks. The addition of a user-friendly Streamlit dashboard adds dimension by supporting real-time data upload, prediction and visualization, thus making the system practical for use by both network administrators, as well as researchers.

VI. References

[1] R. Kumar, A. Malik, and V. Ranga, "A hybrid hunger games search and remora optimization algorithm based intelligent intrusion detection system for IoT wireless networks," Knowl.-Based Syst., vol. 256, Nov. 2022, Art. no. 109762.

[2] W. Wang, S. Jian, Y. Tan, Q. Wu, and C. Huang, "Network intrusion detection system using representation learning based capturing explicit and implicit feature interactions," Comput. Secur., vol. 112, Jan. 2022, Art. no. 102537.

[3] J. Oughton, W. Lehr, K. Katsaros, I. Selinis, D. Bubley, and J. Kusuma, "Wireless internet connectivity revisited: 5G versus Wi-Fi 6," Telecomm. Policy, vol. 45, no. 5, Jun. 2021, Art. no. 102127.

[4] B. A. Tama and S. Lim, "Intrusion detection systems and ensemble learning: A systematic mapping review and cross-benchmark study," Comput. Sci. Rev., vol. 39, Feb. 2021, Art. no. 100357.

[5] S. Lei, C. Xia, Z. Li, X. Li, and T. Wang, "HNN: A new model for understanding network intrusion detection based on multi-feature correlation and temporal-spatial analysis," IEEE Trans. Netw. Sci. Eng., vol. 8, no. 4, pp. 3257-3274, Oct. 2021.

[6] Y. Cheng, Y. Xu, H. Zhong, and Y. Liu, "Using a semi-supervised hierarchical stacking temporal convolutional network for anomaly detection of IoT communication," IEEE Internet Things J., vol. 8, no. 1, pp. 144-155, Jan 2021.

[7] X. Li, M. Zhu, L. T. Yang, M. Xu, Z. Ma, C. Zhong, H. Li, and Y. Xiang, ''Sustainable Ensemble Learning Driving Intrusion Detection Model,'' IEEE Trans. Dependable Secure Comput., vol. 18, no. 4, pp. 1591–1604, Jul./Aug. 2021.

[8] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, ''Building an efficient intrusion detection system based on feature selection and ensemble classifier,'' Comput. Netw., vol. 174, Jun. 2020, Art. no. 107247.

[9] G. Kumar, K. Thakur, and M. R. Ayyagari, ''MLEsIDSs: Machine learning-based ensembles for intrusion detection systems—A review,'' J. Supercomput., vol. 76, no. 11, pp. 8938–8971, Nov. 2020.

[10] B. A. Tama, L. Nkenyereye, S. M. R. Islam, and K. Kwak, ''An Enhanced Anomaly Detection in Web Traffic Using a Stack of Classifier Ensemble,'' IEEE Access, vol. 8, pp. 24120–24134, 2020.