# A MACHINE LEARNING-BASED METHODOLOGY FOR CLASSIFYING ATTACKS ON INTERNET OF THINGS DEVICES

[1]BELLAMKONDA SAI HEMANTH, [2]Dr.P SURI BABU

[1]PG Scholar, Dept. of CSE, KMM Institute of Technology & Science, Ramireddipalle, Tirupati, AP, India.

[2]Professor, Dept. of CSE, KMM Institute of Technology & Science, Ramireddipalle, Tirupati, AP, India.

**Abstract:** The abstract highlights the inherent security flaws associated with Internet of Things (IoT) devices, highlighting the possible risks posed by hackers and intruders. IoT devices are susceptible to exploitation due to their interconnected features, especially through anomaly attacks. The project offers a way to identify anomalous attacks in IoT devices by using machine learning methods, namely Support Vector Machine (SVM), Random Forest (RF), stacking classifiers, and voting classifiers. The effectiveness of these algorithms in feature selection and detection led to their selection. The NSL-KDD dataset in ARFF format is used in the study's testing. The chosen approaches, RF and stacking classifier, demonstrate excellent accuracy rates of approximately. The focus is on false positive rates, which show a very low occurrence in all cases. This highlights the positive outcomes of the recommended approach, particularly the higher accuracy achieved with Random Forest when compared to the existing literature. Promising accuracy, recall, and precision are offered by the stacking classifier and Random Forest, underscoring their potential effectiveness in identifying and thwarting anomalous assaults in Internet of Things devices. Furthermore, ensemble methods were used, including the Voting Classifier (RF + AB) and the Stacking Classifier (RF + MLP with LightGBM), which combine the predictions of several different models to provide a final prediction that is more reliable and accurate. Both the Stacking Classifier and the Voting Classifier attained 100% accuracy. In addition, we created the front end with user authentication for IoT anomaly detection using the Flask framework for user testing.

***Index terms*** *– IOT devices, Support Vector Machine (SVM) and Random Forest (RF).*

## 1. INTRODUCTION

The Internet of Things (IoT) expands web connection to include a wide range of often non-internet-enabled physical equipment and commonplace things in addition to standard smart devices like computers, laptops, smartphones, and tablets. Users, organizations, and other entities can get actionable sensor data from IoT devices. Consumer, enterprise, and industrial are the three main groups. There are several IoT devices on the market today, and users choose these devices based on their features, conditions, and cost. IoT devices specifically include a wide range of products, from small devices like toasters to major appliances like refrigerators. According to Margaret Lee (2020), 64 billion Internet of Things (IoT) devices will be online by 2025.

A data pattern that deviates from expected behavior is referred to as an anomaly. Outliers, exceptions, abnormalities, surprises, and inconsistencies in an IoT function are also related. By analyzing trends, anomaly detection would let the software programs in these IoT devices spot any unusual behavior. Unusual data may point to important events, like technical problems, or possible patterns, such shifts in consumer behavior. According to the publication [2], data leakage, fraud detection, and intrusion prevention systems are separate causes of anomalies. Many IoT applications, including network security and smart cities, use anomaly detection.

There is a dearth of research on machine learning techniques for anomaly detection in Internet of Things devices. Nowadays, a lot of researchers ignore other aspects, including the security of the device, which have a big impact on choosing the perfect equipment. Furthermore, because of their built-in internet connectivity, IoT devices have security flaws that make hackers more likely to take advantage of them. The attack on Western Digital's My Book Live was one of the other attacks, according to the study report [3]. My Book Live devices are used for individual cloud storage. A technological flaw that allowed hackers to reset the devices without a password allowed them to successfully remove all of the data from the devices' storage.

The number of hacked Internet of Things (IoT) devices and cryptocurrency networks in Japan roughly quadrupled in 2018 compared to previous years, according to another analysis [4], which has shown that IoT devices have serious vulnerabilities. As a result, anomaly detection need to be implemented in situations where it might lessen possible damage from hackers or intruders. According to Xu et al. (2019) [5], anomaly analysis is essential in a number of research fields, such as machine learning and data mining. Its goal is to find data areas whose patterns or behaviors deviate from expected values.

## 2. LITERATURE SURVEY

### 2.1 Anomaly Detection: Glimpse into the Future of IoT Data:

**ABSTRACT:** In a time where data is as important as sunshine, modernizing businesses requires effectively managing enormous volumes and extracting pertinent insights. One area where the flood of IoT data may provide insights is anomaly detection. IoT may be examined for issues with the physical devices it is connected to, much as traditional business applications and IT infrastructure. For example, the use of IoT devices to automate and modernize processes in manufacturing or industrial settings may show that anomalies signal that some gear needs maintenance. A prompt notification of a possible problem reduces unexpected downtime. However, gathering all of that data might be difficult since moving it from the device to a central computer platform could be problematic. Does every gadget have to send all of its data? Edge computing can improve the effectiveness of IoT anomaly detection and reduce data saturation. An insightful look into new network designs and how they will advance organizations in the future may be obtained by thoroughly examining that technique.

### 2.2 Attack and Anomaly Detection in IoT Networks using Machine Learning Techniques: A Review:

**ABSTRACT:** One of the modern era's fastest-growing technologies is the Internet of Things (IoT). Through a variety of sensors, this technology allows billions of sentient objects, or "Things," to collect a wide range of data about themselves and their surroundings. They may then distribute it to authorized organizations for a variety of uses, such as improving commercial services or operations or regulating and supervising industrial services. However, the Internet of Things is currently facing previously unheard-of security risks. Significant technical advancements in machine learning (ML) have opened up a number of new research avenues to solve current and upcoming IoT issues. In intelligent devices and networks, machine learning is a useful technique for identifying risks and questionable activities. Based on a thorough literature review of machine learning methods and the significance of IoT security with regard to numerous possible attacks, this study evaluates several machine learning algorithms for attack and anomaly detection. Furthermore, possible IoT security methods based on machine learning have been suggested.

### 2.3 Japan: Hacked IoT Devices and Cryptocurrency Networks Doubled in 2018:

**ABSTRACT:** The number of hacked cryptocurrency networks and Internet of Things (IoT) devices in Japan nearly quadrupled in 2018 compared to the previous year. Asahi, a local English-language media site, published a story on March 7. According to the article, data from the Japanese Police Agency showed an average of 2,752.8 incursions per sensor per day in the previous year, a 45 percent rise. Furthermore, the data shows that about 90% of the assaults came from outside. According to the report, there were an average of 1,702.8 invasions per sensor per day in Bitcoin networks and IoT devices in 2018—more than twice as many as there were in 2017 (875.9).

## 3. METHODOLOGY

### i) Proposed Work:

The machine algorithms are made to detect unusual attacks in Internet of Things devices. The methods that were selected include voting classifier, stacking classifier, Random Forest (RF), and Support Vector Machine (SVM). Robust supervised learning techniques like SVM and RF are used for feature selection as well as for detection. For testing, the NSL-KDD dataset—a common anomaly dataset—was used. We compare the accuracy, recall, precision, and F1-score of the suggested model with previous findings to evaluate its effectiveness. The Voting Classifier, which combines Random Forest and AdaBoost, and the Stacking Classifier, which combines Random Forest, Multi-Layer Perceptron, and LightGBM, were the two ensemble techniques that were described in the paper. The voting classifier and stacking both obtained 100% and 99% accuracy, respectively, demonstrating their effectiveness in enhancing the predictive performance of the individual models. Additionally, a user-centric front-end interface was developed using the Flask framework in order to facilitate user testing and practical implementation. By ensuring secure access, user authentication makes it easier to apply the recommended anomaly detection technique in Internet of Things devices.

### ii) System Architecture:

The research process framework is shown in Figure 1. The Weka tool software will be used to compare the suggested implementation of the two algorithms

with earlier implementations that are most relevant to this study. Random Forest and Support Vector Machine are the two algorithms. A reliable supervised learning technique used for classification and regression applications is the Support Vector Machine (SVM). However, it is mostly used in machine learning for classification problems [2, 7, 8, 9, 10, 11, 12]. On the other hand, the random forest algorithm is an approachable and flexible machine learning technique. To solve regression and classification issues, it makes use of ensemble learning.

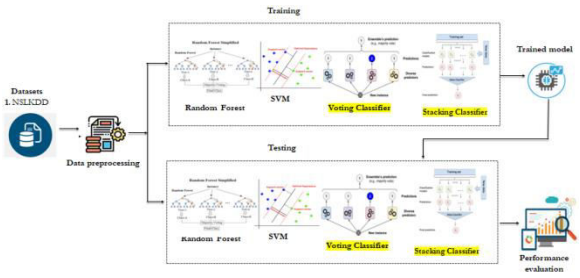


Fig 1 Proposed architecture

**iii) Dataset collection:**

Understanding the characteristics and structure of the NSL KDD dataset is emphasized. To comprehend its features, data types, and any trends, the dataset is loaded and analyzed. The NSL-KDD dataset [12], a benchmark anomaly dataset, is used in the study to compare different intrusion detection systems.Accuracy, false positive rate, true positive rate, precision, recall, and F-measure are among the metrics used by the proposed system to evaluate the model's performance. The older KDD Cup 99 dataset was used to create the publicly accessible NSL-KDD dataset (Tavallaee et al., 2009). A statistical examination of the cup99 dataset revealed important issues that significantly impact intrusion detection accuracy and result in a false evaluation of AIDS (Tavallaee et al., 2009). There are 41 characteristics (features) and 22 training intrusion assaults in the NSL_KDD dataset. According to Tavallaee et al. (2009), this dataset contains 19 properties that describe connections inside the same host and 21 attributes that relate to the connection.

| | duration | protocol_type | service | flag | src_bytes | dst_bytes | land | wrong_fragment | urgent | hot | ... | dst_host_same_srv_rate | dst_host_diff_srv_rate | dst_host |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | tcp | ftp_data | SF | 491 | 0 | 0 | 0 | 0 | 0 | ... | 0.17 | 0.03 | |
| 1 | 0 | udp | other | SF | 146 | 0 | 0 | 0 | 0 | 0 | ... | 0.00 | 0.60 | |
| 2 | 0 | tcp | private | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0.10 | 0.05 | |
| 3 | 0 | tcp | http | SF | 232 | 8153 | 0 | 0 | 0 | 0 | ... | 1.00 | 0.00 | |
| 4 | 0 | tcp | http | SF | 199 | 420 | 0 | 0 | 0 | 0 | ... | 1.00 | 0.00 | |

5 rows × 43 columns

Fig 2 NSL KDD dataset

**iv) Data Processing:**

Data processing is the process of transforming raw data into insights that businesses can use. Data processing, which includes gathering, organizing, cleaning, validating, analyzing, and transforming data into understandable representations like papers or graphs, is what data scientists usually do. Three ways are available for processing data: mechanical, electronic, and manual. Increasing the value of information and simplifying decision-making are the goals. This enables businesses to improve their operations and make quick strategic decisions. In this context, computer software development and other automated data processing technologies are essential. Large datasets, particularly big data, may be transformed into important insights for decision-making and quality control.

**v) Feature selection:**

Finding the most relevant, consistent, and non-redundant features for model building is known as feature selection. While the quantity and variety of datasets continue to grow, it is imperative to systematically reduce dataset sizes. Enhancing a predictive model's effectiveness while reducing modeling-related computing costs is the main goal of feature selection.

Finding the most important characteristics to feed into machine learning algorithms is known as feature selection, and it is a basic component of feature engineering. By eliminating superfluous or duplicate features, feature selection techniques reduce the number of input variables and narrow the set down to those that are most relevant to the machine learning model. The main benefits of selecting features in advance rather than letting the machine learning model decide how important a feature is on its own.

## 4. EXPERIMENTAL RESULTS

**Precision:** Precision assesses the proportion of accurately classified cases among those identified as positive. Consequently, the formula for calculating precision is expressed as:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

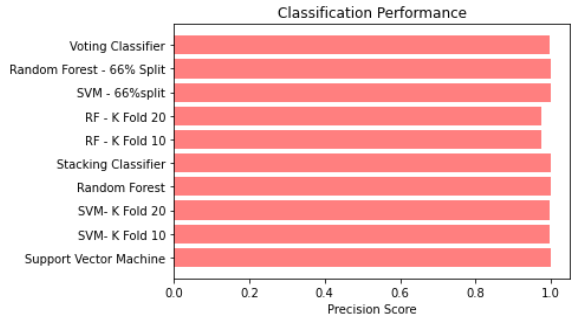


Fig 7 Precision comparison graph

**Recall:** In machine learning, recall is a measure of how well a model can find all relevant instances of a given class. It provides information about how well a model identifies instances of a particular class and is calculated as the ratio of correctly predicted positive observations to the total actual positives.

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

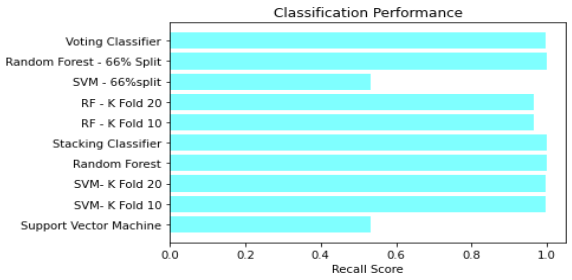


Fig 8 Recall comparison graph

**Accuracy:** Accuracy is the ratio of correct predictions in a classification test, assessing the overall precision of a model's predictions.

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

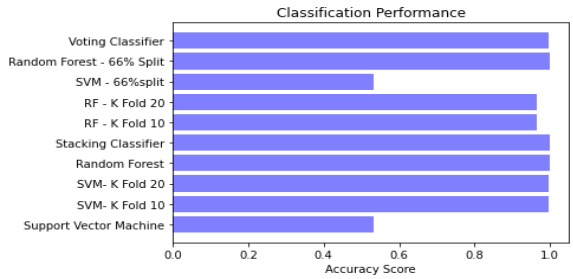


Fig 9 Accuracy graph

**F1 Score:** The F1 Score is the harmonic mean of accuracy and recall, providing a balanced metric that accounts for both false positives and false negatives, hence rendering it appropriate for imbalanced datasets.

$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$

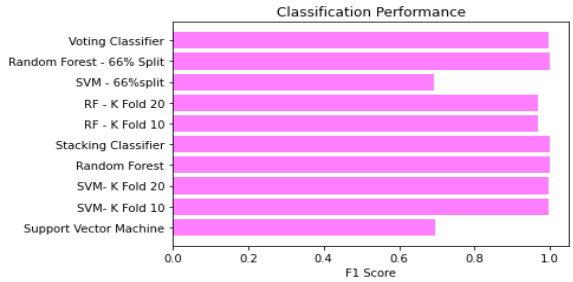


Fig 10 F1Score

| ML Model | Accuracy | Precision | Recall | F1 - score |
|---|---|---|---|---|
| Support Vector Machine | 0.534 | 0.999 | 0.534 | 0.696 |
| SVM- K Fold 10 | 0.998 | 0.998 | 0.998 | 0.998 |
| SVM- K Fold 20 | 0.998 | 0.998 | 0.998 | 0.998 |
| Random Forest | 1.000 | 1.000 | 1.000 | 1.000 |
| Stacking Classifier | 1.000 | 1.000 | 1.000 | 1.000 |
| RF - K Fold 10 | 0.966 | 0.975 | 0.966 | 0.970 |
| RF - K Fold 20 | 0.966 | 0.975 | 0.966 | 0.970 |
| SVM - 66%split | 0.533 | 0.999 | 0.533 | 0.694 |
| Random Forest - 66% Split | 1.000 | 1.000 | 1.000 | 1.000 |
| Voting Classifier | 0.998 | 0.998 | 0.998 | 0.998 |

Fig 11 Performance Evaluation

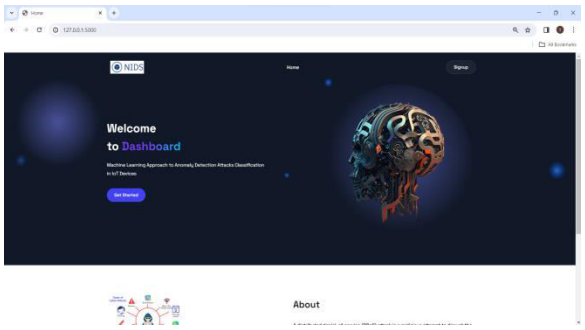


Fig 12 Home page

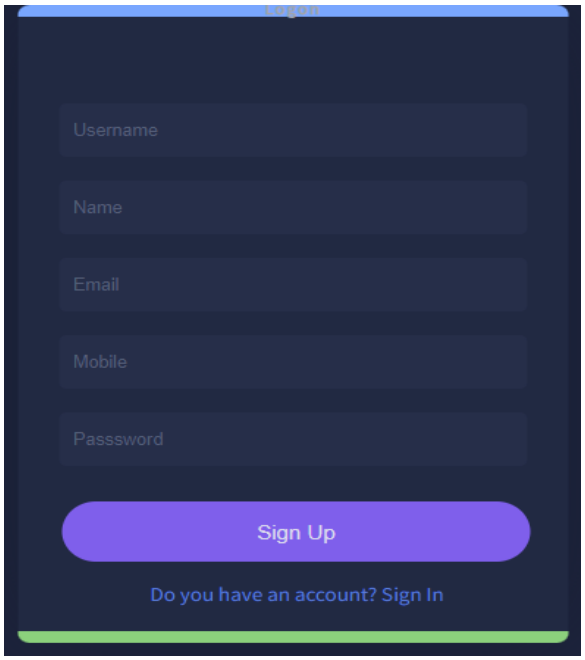


Fig 13 Signin page

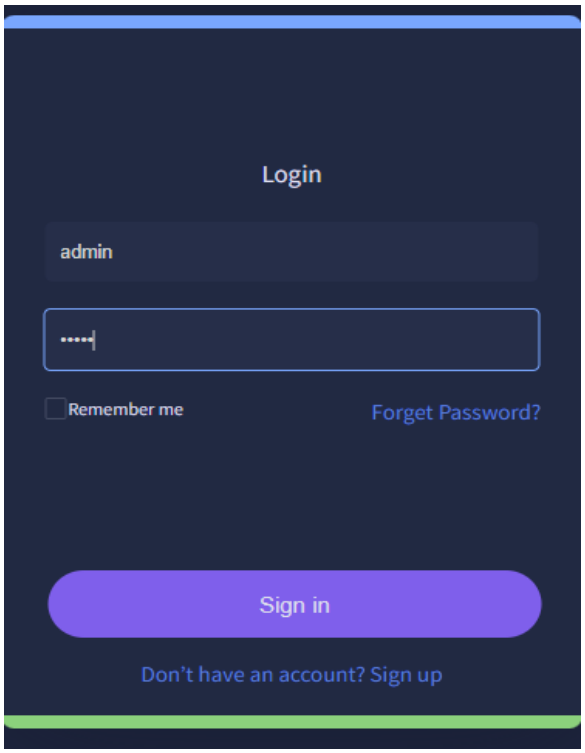


Fig 14 Login page


Fig 15 User input



Fig 16 Predict result for given input

## 5. CONCLUSION

Its capacity to detect and handle anomalous attacks in IoT devices is highlighted by the efficient application of machine learning techniques, including Support Vector Machine (SVM) and Random Forest (RF). The findings show that the suggested approach performs better than existing research. The achieved accuracy demonstrates the method's reliability in addressing unusual problems in IoT environments. By maintaining a consistently low false positive rate across several settings, the research highlights the method's reliability. In the diverse landscape of IoT devices, this consistency is essential for proper categorization of anomaly attacks [1, 2]. Using the NSL-KDD dataset, the study carefully evaluates the proposed machine learning methods [12]. This standardized dataset serves as a trustworthy starting point for assessing the usefulness of anomaly detection in the Internet of Things. With an astounding accuracy of 100%, the alternative method, which made use of ensemble techniques like the Voting Classifier and Stacking Classifier, performed exceptionally well. The algorithm's strength and reliability in real-world applications were proven by extensive front-end testing using feature values, indicating its efficacy in enhancing anomaly detection for IoT devices. By offering a method that successfully addresses the important problem of anomalous assaults, this study significantly improves IoT security and increases the overall resilience of IoT devices against possible threats.

## 6. FUTURE SCOPE

In order to improve anomaly detection skills, the future scope involves researching and implementing

cutting-edge machine learning algorithms or techniques. This might entail using deep learning models or looking at cutting-edge methods that could improve performance. An important path for future development is to improve the project to allow real-time anomaly detection in IoT devices. Maintaining an advantage against rapidly changing cyber threats requires the use of algorithms and procedures that can handle and analyze data in real-time. The project's future scope includes developing adaptive models that can efficiently handle new device types, changing data patterns, and emergent anomalies due to the dynamic nature of IoT ecosystems [6, 11, 14]. This means that the anomaly detection system must be continuously improved and modified. The inclusion of sophisticated security protocols, such as anomaly response systems and encryption techniques, may be emphasized in later iterations of the project. As the project moves forward, adopting a comprehensive approach for IoT device security will be essential to addressing the issues brought on by increasingly sophisticated cyberattacks.

**REFERENCES**

[1] M. Lee. “Anomaly Detection: Glimpse into the Future of IoT Data.” The New Stack. https://thenewstack.io/anomaly-detection-glimpse-into-thefuture-of-iot-data/ 2022, January 24.

[2] S. H. Haji, & S. Y. Ameen, “Attack and Anomaly Detection in IoT Networks using[2, 7, 8, 9, 10, 11, 12] Machine Learning Techniques: A Review.” In (p. 46). 2021.

[3] Firedome (2021). Top Cyber Attacks on IoT Devices in 2021. https://firedome.io/blog/top-cyber-attacks-on-iot-devices-in-2021/. 2021, November 30.

[4] A. ZMUDZINSKI, “Japan: Hacked IoT Devices and Cryptocurrency Networks Doubled in 2018.”. Cointelegraph. https://cointelegraph.com/news/japan-hacked-iot-devices-andcryptocurrency-networks-doubled-in-2018. 2019, March 7.

[5] X. Xu, H. Liu, & M. Yao, Recent Progress of Anomaly Detection. Complexity, 2019, 1–11. https://doi.org/10.1155/2019/2686378. 2019.

[6] C. Ioannou, & V. Vassiliou, “Network Attack Classification in IoT Using Support Vector Machines.” https://www.mdpi.com/2224-2708/10/3/58/pdf . 2021.

[7] B. Nassif, A. Abu Talib, M., Nasir, & F. Dakalbab, “Machine Learning for Anomaly Detection: A Systematic Review.” Ieee Access 9 (2021): 78658-78700. 2021 May 24.

[8] C. Das, A. Rasool, A. Dubey, & N. Khare,. “Analyzing the Performance of Anomaly Detection Algorithms.” International Journal of Advanced Computer Science and Applications Vol. 12, no. 6 2021.

[9] Y. Gavrilova “Anomaly Detection in Machine Learning.” Software Development Company. https://serokell.io/blog/anomaly-detection-inmachine-learning. 2021 December 10.

[10] S. Benqdara, & M. A. Ngadi,. “Machine Learning Techniques for Anomaly Detection: An Overview.” International Journal of Computer Applications. Vol. 79, no. 2. 2013.

[11] M. Hasan, M. Islam, M. Md., I. Zarif, & M. M. A. Hashem. “Attack and Anomaly Detection in IoT Sensors in IoT sites using[2, 7, 8, 9, 10, 11, 12]

Machine Learning Approaches." Internet of Things, Vol. 7, p.100059. 2019.

[12] Mathworks, "Machine Learning.". Www.mathworks.com. https://www.mathworks.com/discovery/machinelearning.html#:~:text=Machine%20learning%20uses%20two%20types. n. d,

[13] T. Crunch. "The evolution of machine learning." TechCrunch. 2017 Aug 8. https://techcrunch.com/2017/08/08/the-evolution-of-machinelearning/ (16 January 2023).

[14] B. Posey, S. Shea "What are IoT Devices?' TechTarget.com. IoT Agenda. 2022 https://www.techtarget.com/iotagenda/definition/IoT device (Accessed 16 January 2023).

[15] A.W. S. Amazon, "What is IoT? - Internet of Things Beginner's Guide - AWS.". Amazon Web Services, Inc. 2022 https://aws.amazon.com/what-is/iot/ (Accessed 16 January 2023).