IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 12, Issue 2, 2024

Classification of Anomaly Detection Attacks in IoT Devices using Machine Learning

Ms. Shilpa Keshri1, Computer Engineering1, shilpakeshri12@gmail.com

Dr. Sunil Wanjari2, Associate Professor2, Dept. of Computer Engineering2, swanjari@stvincentngp.edu.in

Dr. Kapil Gupta3, Assistant Professor3, Dept. of Computer Engineering3, kaps04gupta@gmail.com

St. Vincent Pallotti College of Engineering and Technology, Nagpur.123

Abstract: Concerns about the security of Internet of Things (IoT) devices are growing, and the abstract emphasizes how vulnerable they are to anomaly attacks. The project suggests a way to find strange things in the Internet of Things (IoT) using Machine Learning (ML) methods like Support Vector Machine (SVM) and Random Forest (RF), along with votes and stacked models. Experiments using the NSL-KDD dataset show that RF and stacked models can get high accuracy rates with few false positives. RF significantly beats current material, which shows how useful it could be. Some ensemble methods, like Voting Classifier (RF + AB) and Stacking Classifier (RF + MLP with LightGBM), are very good at detecting and preventing anomalies because they have high accuracy, memory, and precision. In addition, the project includes user testing through a front end built on the Flask framework and user identification, which makes IoT anomaly detection more useful in real life.

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

1. INTRODUCTION

With the Internet of Things (IoT), more than just standard gadgets can connect to the internet. This makes it easier for users, businesses, and other groups to send data to each other. IoT devices, which range from toasters to freezers, are divided into three groups: consumer, business, and industrial. They are also very different in size and function. Margaret Lee says that by 2025, there will be 64 billion Internet of Things (IoT) gadgets online [1]. Anomaly identification is an important part of IoT because it finds trends that don't behave the way they should, like outliers, exceptions, and irregularities [2]. This kind of analysis helps find technology problems or changes in how people act. Even though it's important, not much study has been done on Machine Learning (ML) methods to finding problems with IoT [2, 7, 8, 9, 10, 11, 12]. There are a lot of security issues because IoT devices can be attacked, like what happened with Western Digital's My Book Live, where hackers deleted data because of holes in the system [3]. Also, the number of IoT devices that were hacked doubled in Japan in 2018, showing how

ISSN 2347–3657

Volume 12, Issue 2, 2024

vulnerable they are [4]. Xu et al. stress that anomaly analysis is very important in many areas, such as data mining and machine learning, to find trends that don't match what is expected [5]. So, strong tools for finding strange behavior are needed to lessen the damage that could come from hacks or intrusions.

2. LITERATURE SURVEY

The Internet of Things (IoT) has grown very quickly, letting billions of smart gadgets use sensors to gather data for many uses. But because there are so many IoT devices, there are also more security risks [2, 7, 8, 9, 10, 11, 12]. Machine Learning (ML) has become an important tool for dealing with these problems because it opens up new areas for study and security technologies [1, 2, 3]. Machine learning methods, like Support Vector Machines (SVM), are very important for finding risks and strange things in IoT networks [6, 7]. For intrusion detection systems (IDS) to keep an eye on strange actions in smart node devices, SVM models like C-SVM and OC-SVM are used, which results in high classification accuracy rates [6]. ML has made a lot of progress in anomaly identification, which is a basic job for finding problems in data [2, 7, 8, 9, 10, 11, 12]. A Systematic Literature Review (SLR) was done in [7] that looked at 290 research papers from 2000 to 2020 and showed different machine learning models and datasets that were used to find anomalies. Because they work so well, unsupervised anomaly detection methods are becoming more popular in study [7]. Using machine learning in real-life situations, like healthcare, can also change the way anomalies are found, giving faster answers and important information about bodily data [8]. Even though a lot of study has been done on finding anomalies, not much has been done on analyzing trained and unstructured methods used on physiological datasets [8]. Also, detecting intrusions is still an important part of cybersecurity study, and getting rid of false alerts is still a problem [10]. Anomaly detection is an important part of intrusion detection systems because it finds changes from regular behavior that could mean an attack or a problem [10]. Even though there are still problems, new study points the way to better ways to handle finding anomalies using both controlled and unstructured methods.

3. METHODOLOGY

i) Proposed Work:

Support Vector Machine (SVM) and Random Forest (RF) are two machine learning techniques that have been suggested for finding strange activity in Internet of Things (IoT) devices. Other methods include stacking and voting classifiers. The NSL-KDD dataset [12] was used to test these algorithms and show that they are good at both recognition and feature selection. To rate how well a model works, we use evaluation measures like accuracy, recall, precision, and f1-score. Notably, the vote predictor is 100% accurate, while stacking is only 99% accurate. This shows how well they work to improve prediction skills. Also, to make it easier to use in real life, a front-end interface made with the Flask framework that is easy for anyone to use makes sure that entry is safe through user registration.

ii) System Architecture:

A plan for the study process is shown in Figure 1. The suggested way to use the two algorithms in the Weka tool program to check how they compare to

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 12, Issue 2, 2024

other ways they have been used in the past that are most relevant to this study paper. The names of these two methods are Support Vector Machine and Random Forest. Support Vector Machine, or SVM, is a strong Supervised Learning method that can be used for both regression and classification problems. That being said, it is mostly used in [2, 7, 8, 9, 10, 11, 12] Machine Learning for Classification problems. On the other hand, the random forest algorithm is a Machine Learning method that is easy to use and flexible. It uses group learning to deal with problems like regression and classification.

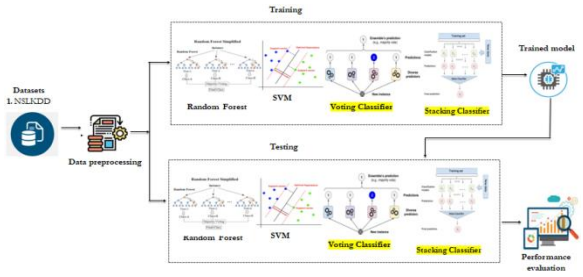

Fig 1 Proposed architecture

iii) Dataset collection:

The main goal of the project is to look at the NSL-KDD dataset [12], which is a common set of anomalies used to compare intrusion detection systems. The NSL-KDD dataset comes from the KDD cup99 dataset [Tavallaee et al., 2009] and has 42 training intrusion strikes and 41 characteristics. Notably, 21 characteristics are about the link itself, and 19 are about the nature of the relationship within the same host [Tavallaee et al., 2009]. To rate how well a model works, we use evaluation measures like precision, recall, F-measure, accuracy, false positive rate, and true positive rate. The statistical study of the cup99 dataset showed problems that make intruder detection less accurate [Tavallaee et al., 2009].

	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 handling is the process of turning unstructured data into knowledge that businesses can use. In general, data scientists handle data, which means they gather it, organize it, clean it, check it, analyze it, and turn it into forms that can be read, like graphs or papers. There are three ways to handle data: by hand, mechanically, or electronically. The goal is to make knowledge more useful and decision-making easier. This helps companies run better and make smart strategy decisions more quickly. This is made possible in large part by automated data handling tools, like computer programs. It can help turn big data and other types of data into useful information for decision-making and quality control.

v) Feature selection:

A very important part of feature engineering is choosing which features are the most useful to feed into Machine Learning methods [2, 7, 8, 9, 10, 11, 12]. By getting rid of traits that aren't needed or aren't important, this process aims to improve model performance and make computations simpler. Feature selection improves the performance of forecasting

ISSN 2347–3657

Volume 12, Issue 2, 2024

models by gradually decreasing the size of datasets. It lets models focus on the most important traits, which makes predictions more accurate and easier to understand. This proactive method makes sure that models are taught on the most useful traits, which makes Machine Learning systems work better and faster.

vi) Algorithms:

A well-known guided learning method called Random Forest can be used to solve both Classification and Regression issues. It uses ensemble learning to make predictions more accurate by mixing several decision trees on different parts of the information. Random Forest is tested with 10 and 20 folds using k-FOLD cross-validation, going through training and testing groups over and over to get an idea of how well it does in generalization. Random Forest was chosen because it can handle large datasets. Its group nature helps it find different patterns, which makes it good for IoT apps and lowers the risk of overfitting.

The Support Vector Machine (SVM) is a well-known guided learning method that is mostly used for sorting jobs. Its goal is to find the best decision border, also known as a hyperplane, to divide n-dimensional space into groups. SVM uses K-Fold Cross-Validation with 10 and 20 folds to split the data into groups that can be used for training and testing over and over again. SVM was chosen because it is good at dealing with large amounts of data. It works well in IoT settings, especially when looking for strange behavior, where there are many possible choices.

A common ensemble modeling method called "stacking" combines weak learners with meta-learners at the same time. The goal is to make better predictions about the future by figuring out the best way to combine the predictions from different models. Stacking improves the spotting of anomalies in IoT data by using the results of sub-models and a meta-classifier. It does this by recording a wider range of patterns, which helps with generalizing about how anomalies change and adapt in IoT settings.

A voting classifier takes results from several models and predicts an output based on the most likely class. It then uses a majority vote method to choose the final output. The Voting Classifier takes the best parts of different models and uses them together to make predictions. This improves overall performance and makes sure that decisions are fair, which makes the system more effective in IoT settings.

4. EXPERIMENTAL RESULTS

Precision: Precision is the percentage of correctly classified cases or samples compared to those that were correctly classified as hits. So, here is the method to figure out the precision:

$$\text{Precision} = \text{True positives} / (\text{True positives} + \text{False positives}) = TP / (TP + FP)$$

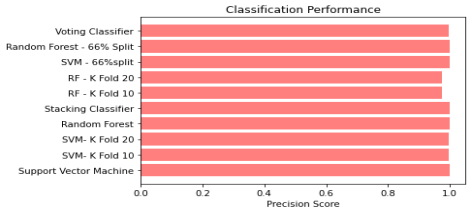

Fig 3 Precision comparison graph

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 12, Issue 2, 2024

Recall: In machine learning, recall is a parameter that shows how well a model can find all the important cases of a certain class. It shows how well a model captures cases of a certain class. It is calculated by dividing the number of correctly predicted positive observations by the total number of real positives.

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

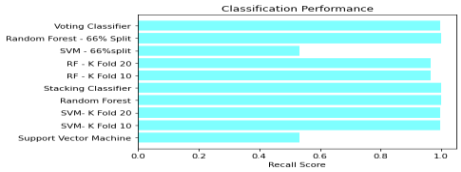

Fig 4 Recall comparison graph

Accuracy: Accuracy is the percentage of right guesses in a classification job. It shows how accurate a model's forecasts are generally.

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

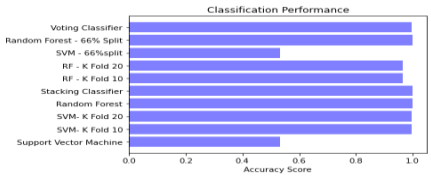

Fig 5 Accuracy graph

F1 Score: The harmonic mean of accuracy and recall is F1. This fair measure accounts for erroneous positives and negatives, therefore it may be used with unbalanced datasets.

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

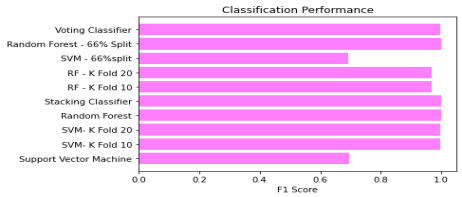

Fig 6 F1Score

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 7 Performance Evaluation

Fig 8 User input

Fig 9 Predict result for given input

5. CONCLUSION

Support Vector Machine (SVM) and Random Forest (RF) algorithms have been used to show that they can find and stop anomaly attacks in Internet of Things (IoT) devices [3]. The results are better than what has been written before, with high accuracy and

consistently low false positive rates, which are important for accurately classifying anomalies [1, 2]. Using the NSL-KDD dataset [12], the project carefully tests the performance of the algorithm, showing how reliable it is in real-world IoT settings. Ensemble methods, such as Voting Classifier and Stacking Classifier, are very accurate, which proves that they work. This project makes IoT security a lot better by tackling the important problem of anomaly threats and making devices more resilient.

6. FUTURE SCOPE

In the future, it might be possible to improve the ability to find anomalies by using more advanced Machine Learning methods, such as deep learning models [6, 11]. Real-time monitoring in IoT devices is very important, and to fight changing cyber dangers, we need algorithms that can handle data in real time [6, 11]. Adaptive models are needed to deal with new gadgets and strange situations, and they need to be improved all the time. In later versions, the focus may be on better security features like encryption and methods for responding to strange events in order to deal with more advanced threats.

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 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.

ISSN 2347–3657

Volume 12, Issue 2, 2024

[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).