'National Conference on Advancements in Computer Science and Engineering'
International Journal of Scientific Research in Computer Science,
Engineering and Information Technology | ISSN : 2456-3307 (www.ijsrcseit.com)

A Cyber Threat Multi-step-ahead Intelligence Sharing Scheme for IoT Bot Detection Using Ensemble Learning

Ezra Vethamani[1], Nikil Krishnan[2], Sudha Sirisha[2], Pradeep Kumar P[2]
[1]Assistant Professor, [2]Final year UG student
Department of Computer Science Engineering SRM Institute of Science and Technology Chennai, Tamil Nadu, India

ABSTRACT
Network traffic analysis is that bots within a botnet typically exhibit similar traffic behavior, which can be characterized and classified using specific attributes to differentiate them from non- malicious traffic. Unlike content-based packet analysis, traffic analysis is not affected by encryption. However, many users are unaware of whether bots have compromised their devices, leading to numerous security incidents. To address this issue, we propose an IoT-based monitoring system comprised of cost-effective devices and wireless communication methods. These IoT devices can connect with each other and transmit data to a unified system for bot detection. To resolve this problem, we propose to focus on Internet of Things (IoT)-based monitoring system made up of affordable hardware and wireless communication technologies. These Internet of Things (IoT) devices may communicate with one another and provide data to a centralized bot detection system.This article focuses on the critical security challenges associated with IoT devices, such as cybercrime and countermeasures. Since IoT devices lack adequate safety mechanisms, they are frequently the target of cyberattacks. As such, this paper highlights the need for confidentiality, integrity, and authentication in IoT devices to ensure their safety.

Keywords- IoT-based monitoring system, Bot detection, Cybersecurity

I. INTRODUCTION

The Internet of Things (IoT) has enabled smart gadgets to interact and make decisions via the Internet with little or no human intervention. This technical innovation has tremendously improved the quality of our lives. The number of IoT devices has grown at an unprecedented rate in recent years. However, as more IoT applications are deployed across various sectors, such as the climate system and smart city energy, detecting network attacks becomes an increasingly important task. Due to limited processing resources, many IoT devices have inherent security risks.

An intrusion detection system (IDS) detects intrusions is used to identify prospective assaults and abnormalities in order to solve this issue. IoT devices connect to a local access gateway, which feeds real-time local data to a global security gateway in an IDS configuration. The global gateway then provides the relevant anomaly detection model to the local gateways for intrusion detection. The IDS continuously monitors IoT device communications and detects unusual communication behavior using anomaly detection models.

The Internet of Things (IoT) is a fast increasing network that connects hardware, software, and devices through complex linkages to collect and distribute data. As more people use IoT devices across different sectors, security and privacy issues arise. One important area of study is detecting unusual data patterns, known as anomalies or outliers, in IoT data streams. Anomalies can indicate problems or rare events that may be caused by errors or malicious attacks. Detecting and addressing anomalies is crucial for maintaining the integrity and security of the entire IoT network. Anomaly detection can help identify device malfunctions, cyber-attacks, unusual industrial processes, and even financial fraud. It is a critical tool for ensuring the smooth operation and security of IoT networks.

Machine learning techniques are widely used to discover anomalies in the IoT due to their ability to swiftly and effectively analyse vast volumes of data. These approaches, which include supervised and unsupervised learning techniques like as support vector machines, decision trees, clustering algorithms, and autoencoders, can be applied to a variety of data sources such as sensor data, network traffic data, and financial transaction data. It is possible to detect malicious data, prevent network attacks, and discover anomalous behaviours in the IoT environment by employing anomaly detection algorithms to recognise patterns in IoT data streams. Algorithms based on statistics, proximity, machine learning, and deep learning are all common ways for detecting anomalies in the IoT.

II. PROPOSED SYSTEM

Anomalies, often known as outliers, are odd patterns in big data. Anomaly detection is the process of identifying patterns that behave unexpectedly or differently from what is predicted. Anomalies are typically identified during data cleansing, anomaly detection, on the other hand, can detect anomalies in real-world scenarios such as fraud, intrusion, damage detection, or abnormal health problems.

This paper aims to help detect and prevent the Mirai malware, which has been used in several high- profile DDoS attacks. Mirai creates and controls botnets of IoT devices. The paper analyzes the malware's code, explain its parts, and creates a virtual environment for dynamic analysis of Mirai. A honeypot implementation is proposed to detect and report telnet attacks on IoT devices, including manual and Mirai-based attacks. The detection system can identify bots with dynamic or polymorphic behavior without prior knowledge of malicious signatures or profiles. The paper also proposes analyzing network flow characteristics and quantifying evidence with the apriori algorithm to detect the presence of bots.

Benefits-

- Proposes a novel strategy to decreasing computing performance and memory usage during outlier detection; the proposed anomaly detection technique outperforms the baseline approaches in terms of accuracy.
- Fast and efficient, but also as accurate as state-of-the-art algorithms.
- Supports location awareness of anomalies.
- Demonstrating high robustness and imperceptibility.
- The strategy reduced the number of false positives while increasing the rate of balanced detection.
- Enhance operational effectiveness.

A. Proposed Algorithm

Ensemble learning is a strong machine learning approach for improving predicted accuracy and resilience by combining numerous models. It can handle complex, high-dimensional, and diverse data by leveraging the strengths of different algorithms and models. Ensemble learning algorithms can continuously improve over time, adapting to changes in the data and improving accuracy with each iteration. This approach has demonstrated great success in various data analytics applications, including image classification, natural language processing, and financial forecasting. By combining the strengths of different models, ensemble learning can improve predictive accuracy, reduce overfitting, and enhance the generalization of models.

Benefits-

- Managing data that is multidimensional and diverse.
- Continuous progress.
- Have had great success in a variety of data analytics applications.

B. System Architecture

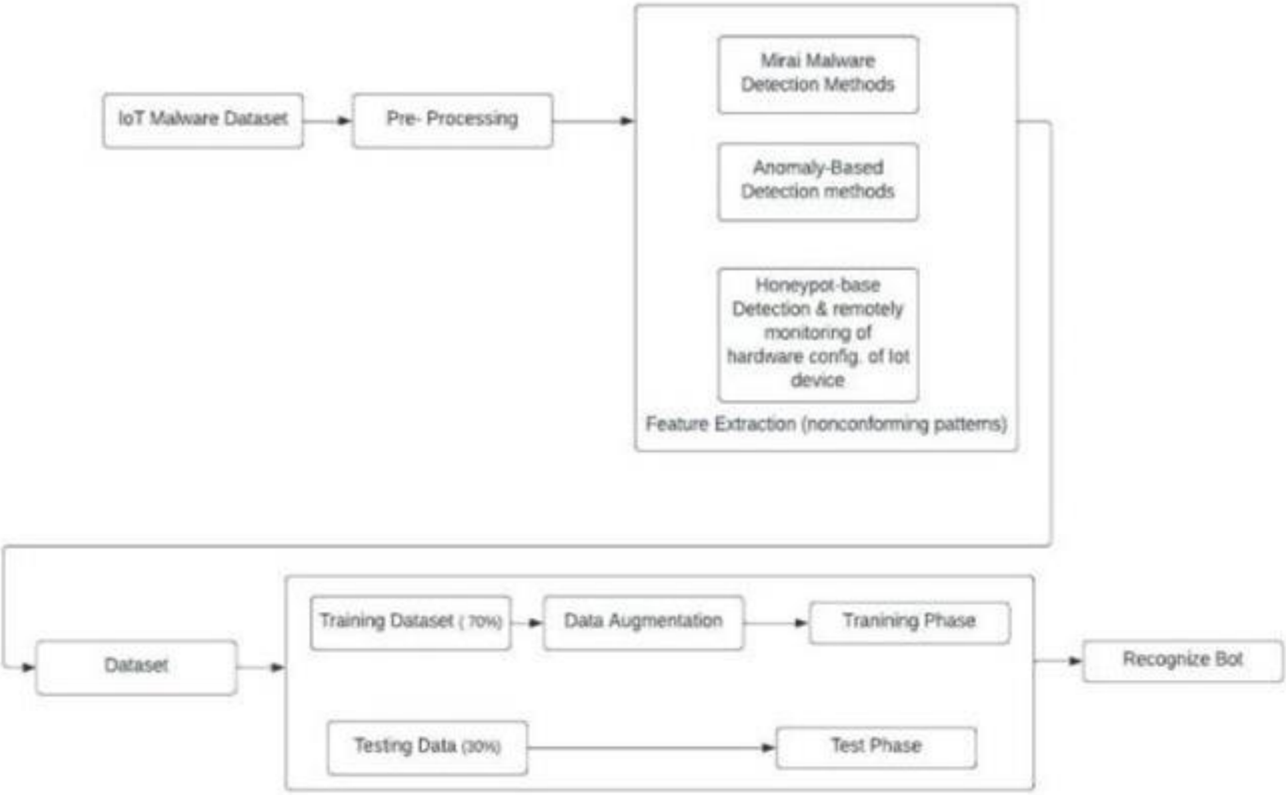

IoT Malware Dataset
Pre- Processing
Mirai Malware Detection Methods
Anomaly-Based Detection methods
Honeypot-base Detection & remotely monitoring of hardware config. of Iot device
Feature Extraction (nonconforming patterns)
Dataset
Training Dataset (70%)
Data Augmentation
Tranining Phase
Testing Data (30%)
Test Phase
Recognize Bot

Figure 2-System Architecture

III. RELATED WORKS

[1] Chunhui Zhao et al. demonstrates a new approach for identifying abnormalities in hyperspectral pictures by integrating fractional Fourier transform (FrFT) with saliency-weighted collaborative representation (SWCR). This method is more accurate and robust against noise and interference. It can detect small and

subtle anomalies that are not detected by other methods. On two hyperspectral datasets. This approach can be useful for environmental monitoring, mineral exploration, and military surveillance.

[2] Zhaoyang et al. provides a novel approach for identifying insulator anomalies based on few-shot learning. The approach collects information from insulator photos with a convolutional neural network (CNN) and then detects abnormalities with a few- shot learning algorithm. The suggested technique may learn from a small number of samples, reducing the requirement for a huge training dataset. Experiment findings demonstrate that the suggested technique detects insulator abnormalities with high accuracy and outperforms standard machine learning methods.. The method has potential applications in the field of power grid inspection, where accurate and efficient detection of insulator anomalies is critical for ensuring power grid safety and reliability.

[3] Kisan Sarda et al. suggested a multi-step anomaly detection technique for the steel sector based on robust distances.The proposed method first applies a clustering algorithm to group similar data points together and then calculates robust distances to identify anomalous data points. The method can detect both isolated and contextual anomalies. The suggested technique beats existing state-of-the-art anomaly detection methods in terms of accuracy and efficiency, according to experimental data. The method has potential applications in the steel industry, where accurate and efficient detection of anomalies is critical for maintaining product quality and safety.

[4] Ning Yan et al. proposes an online yarn breakage detection method based on reflection- based anomaly detection. The method uses a camera to capture images of the yarn and then analyzes the reflection patterns to detect anomalies caused by yarn breakage. The suggested technique detects breakages in real time and can discriminate between broken and unbroken yarns with excellent accuracy. The method has potential applications in the textile industry, where real-time detection of yarn breakages is critical for maintaining product quality and efficiency.

[5] Osama Abdelrahman and Pantea Keikhosrokiani present a machine learning-based technique for identifying and analysing abnormalities in assembly line data. The process entails detecting abnormalities in data using supervised and unsupervised machine learning algorithms, followed by root cause analysis to determine the underlying reasons of the anomalies. The suggested technique detects abnormalities with high accuracy and gives useful insights into the underlying causes of the anomalies. The proposed method outperforms current state-of-the-art anomaly identification and root cause analysis methods in experiments. The method has potential applications in various industries, where accurate and efficient detection and analysis of anomalies are critical for maintaining product quality and efficiency.

[6] Junfeng Tian et al. present the Moving Things Outlier Detection Algorithm (MTOD) for detecting inconsistencies in the Internet of Moving Things (IoMT). While traditional trajectory outlier detection algorithms can detect common anomalies, they are ineffective at detecting generalised anomalies such as those caused by gravity or magnetic fields. This problem is addressed by MTOD, which analyses both location trajectory and multi-sensor data using a three-step methodology of generalisation, partition, and detection. The authors use experimental results to demonstrate the efficiency and accuracy of MTOD. However, there are still some issues to work out, such as determining the relationship between the parameters and effective parameter reduction, and implementing the algorithm in real-time online settings.

[7] Xiaokang Zhou, et al. proposes a variational long short-term memory (VLSTM) learning model for intelligent anomaly detection in industrial big data (IBD) environments, and an estimation network is used to identify anomalies in IBD. In imbalanced inflammatory bowel disease (IBD) datasets, the challenge of inconsistency between dimensionality reduction and feature retention is tackled by the proposed VLSTM model. Results from experiments conducted on a publicly available IBD dataset demonstrate that the VLSTM model effectively handles issues related to imbalance and high dimensionality, and achieves significant improvements in accuracy and false alarm rates for anomaly detection in IBD.

[8] Kyung-Su Kim et al. presents CSIP (contrast- shifted instances via patch-based percentile) for detecting diseased lung shadowing in chest X-ray data obtained exclusively from healthy subjects. The suggested method improves on previous one-class classifiers (OCCs) by employing a patch-based percentile strategy to boost the network's sensitivity to detecting changes in shadowing density across different local locations of the lung. The results suggest that the proposed technique outperforms existing OCC methods in terms of diagnostic performance (with an average AUC of 0.96 for diverse lung illnesses). This method has the potential to aid in the early detection of anomalies associated with emerging infectious illnesses, such as coronavirus variations, for which training data is scarce.

[9] Andrea Castellani et al. proposes fragile direct approaches to anomaly detection for industry- oriented settings, utilizing Digital Twin technology to generate raw data for training. The proposed Siamese Autoencoder (SAE) outperforms the anomaly detection approaches on real data from an existing monitoring model, with robust results across different hyperparameter settings. The methods make use of a limited number of classified anomalous measurements from the actual equipment along with simulated normal operation data. The study demonstrates the potential of Digital Twin technology for generating synthetic datasets for the development of effective anomaly detection algorithms.

[10] Xu Fang et al. proposes a new method on detecting faults in sewer pipelines by ML based anomaly techniques. Unlike existing methods, this approach does not require pre-existing annotated inefficient sample data for training, making it computationally efficient. The method was tested on real data gathered in Macau, accuracy rate exceeding 85% was attained. The proposed method for surveying urban sewer pipelines is a novel and efficient technique that reduces the need for manual identification or data annotation, thereby enhancing productivity and reducing costs. The code and data utilized in this study are publicly accessible to promote future research in this field.

IV. RESULTS AND DISCUSSION

We created four different neural network models using TensorFlow's Keras.Each model has two hidden layers of 128 neurons each, as well as a final output layer of 10 neurons that uses the softmax activation function.

- The first model is a simple neural network with no noise that employs the Adam optimizer with a sparse categorical cross- entropy loss function.
- In the second model, Gaussian noise is added to the input layer to make the model more resilient to data noise. It employs the same Adam optimizer and loss function as the first model.
- The third model is identical to the second but use the NAdam optimizer instead of Adam. The NAdam optimizer is a faster- converging variation of Adam.

- The fourth model is the same as the first model but uses the NAdam optimizer instead of Adam.

The sparse categorical cross-entropy loss function is used to train these models by comparing the predicted output of the model with the real output labels. The optimizer is used to modify the model's weights and biases in order to minimise the loss function.

The accuracy measure is used to evaluate the performance of each model. The proportion of properly categorised instances in the test dataset is measured by the accuracy.

For each model, I trained it on a dataset and recorded the training and test accuracy over time. The results are shown in the graph below:

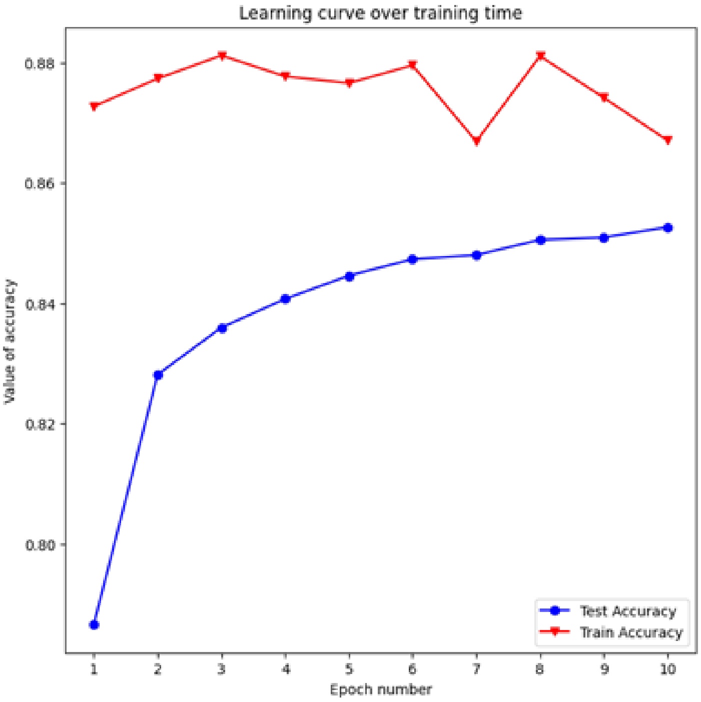


Figure 3- Learning curve over training time (model-1)

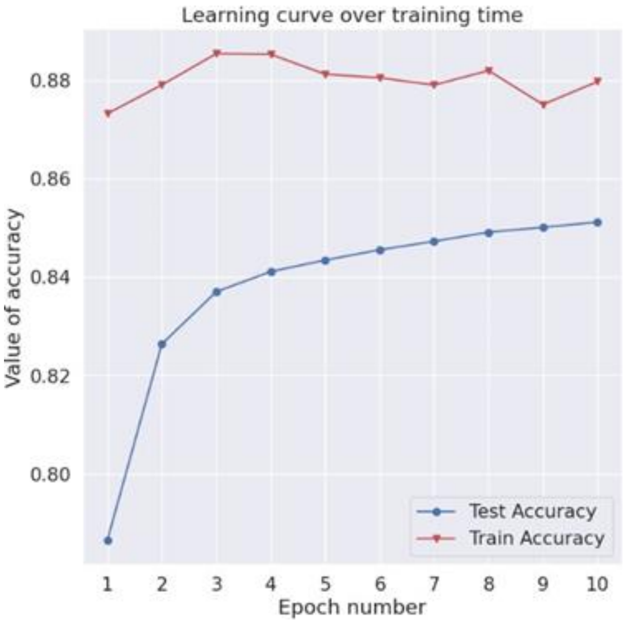


Figure 4- Learning curve over training time (model-2)

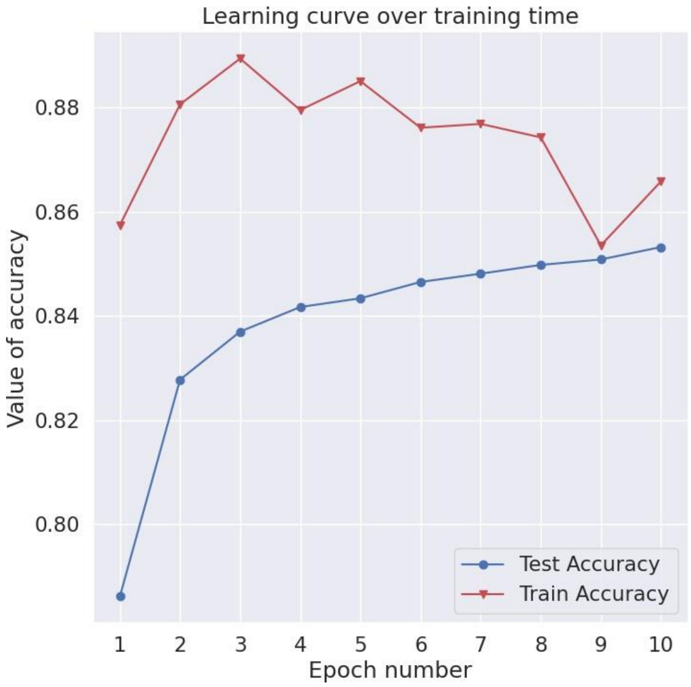


Figure 5- Learning curve over training time (model-3)

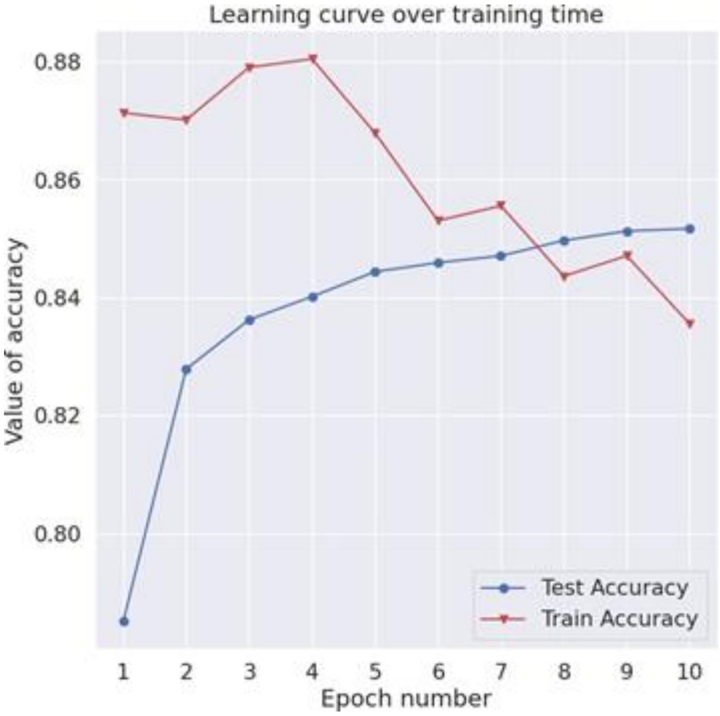


Figure 6- Learning curve over training time (model-4)

Heatmap of Dataset: -

We also generated a heatmap for the dataset used to train the neural network models. This heatmap visualizes the distribution of data points across different features, providing insights into potential patterns or correlations in the data. See the heatmap below:

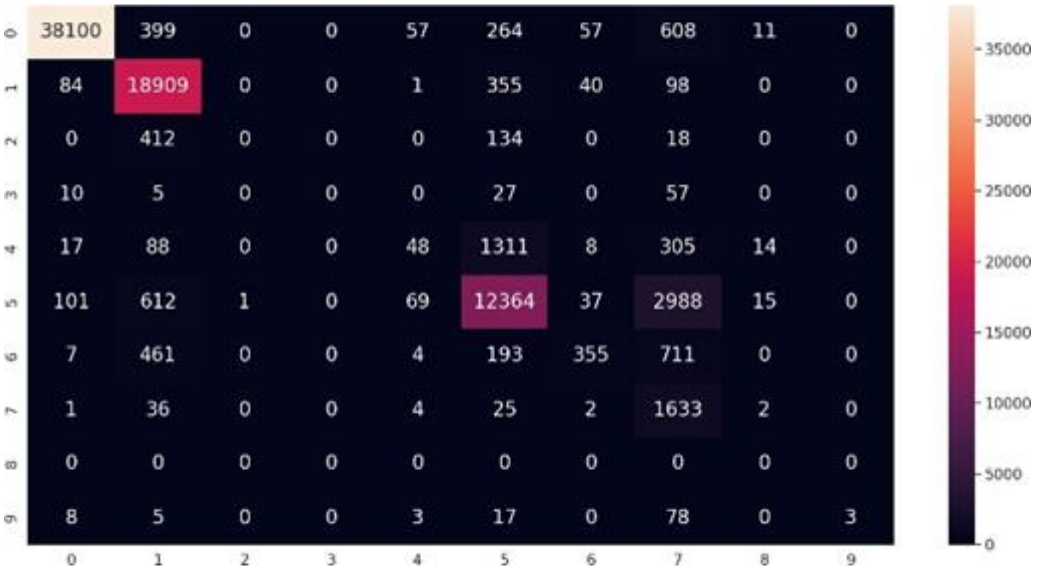
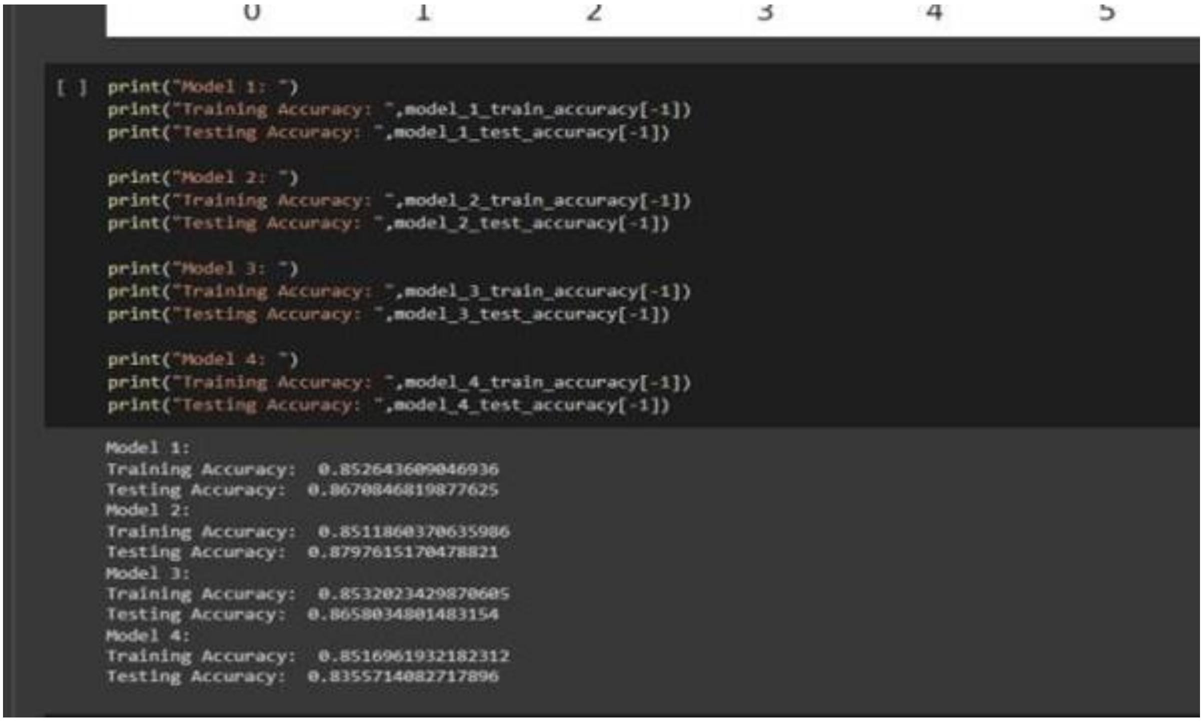


Figure 7- Heatmap of Dataset

V. RESULT



```
print("Model 1: ")
print("Training Accuracy: ",model_1_train_accuracy[-1])
print("Testing Accuracy: ",model_1_test_accuracy[-1])

print("Model 2: ")
print("Training Accuracy: ",model_2_train_accuracy[-1])
print("Testing Accuracy: ",model_2_test_accuracy[-1])

print("Model 3: ")
print("Training Accuracy: ",model_3_train_accuracy[-1])
print("Testing Accuracy: ",model_3_test_accuracy[-1])

print("Model 4: ")
print("Training Accuracy: ",model_4_train_accuracy[-1])
print("Testing Accuracy: ",model_4_test_accuracy[-1])
```

```
Model 1:
Training Accuracy:  0.852643609046936
Testing Accuracy:  0.8670846819877625
Model 2:
Training Accuracy:  0.8511860370635986
Testing Accuracy:  0.8797615170478821
Model 3:
Training Accuracy:  0.8532023429870605
Testing Accuracy:  0.8658034801483154
Model 4:
Training Accuracy:  0.8516961932182312
Testing Accuracy:  0.8355714082717896
```

Figure 8- The above figure shows the level of data accuracy for various models.

The given results show how well four different models perform on both the data they were trained on and new, unseen data. A higher training accuracy means the model is good at fitting to the data it has seen, while a higher testing accuracy means the model is better at generalizing to new data.

Model 2 has the highest testing accuracy, which means it performs better than the other models on new data. However, it's important to note that Model 1 and Model 3 also perform well on the testing data. Model 4 has the lowest testing accuracy, indicating it may not be the best choice for this task.

All four models have similar training accuracies, so they all fit the training data well. However, a high training accuracy does not always mean the model will perform well on new data.

Overall, Model 2 appears to be the best choice based on its high testing accuracy. However, more analysis and experimentation may be necessary to determine the best model for the specific needs of the task.

VI. CONCLUSION

The Internet of Things (IoT) is a current area of research with significant attention given to security issues. This study specifically examines devices that may not receive updates from their manufacturers, leaving them vulnerable. One of the challenges to ensuring IoT security is to ensure that the communication between devices and cloud services or applications is secure. Many IoT devices do not have the capability to encrypt messages until they are transmitted over the network.

VII. REFERENCES

[1]. Zhao, Chunhui & Li, Chuang & Feng, Shou & Su, Nan & Li, Wei. (2020). A Spectral– Spatial Anomaly Target Detection Method Based on Fractional Fourier Transform and Saliency Weighted Collaborative Representation for Hyperspectral Images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 13. 5982- 5997. 10.1109/JSTARS.2020.3028372.

[2]. Wang, Zhaoyang & Gao, Qiang & Li, Dong & Liu, Junjie & Wang, Hongwei & Yu, Xiao & Wang, Yipin. (2021). Insulator Anomaly Detection Method Based on Few-Shot Learning. IEEE Access. PP. 1-1. 10.1109/ACCESS.2021.3071305.

[3]. Acernese, Antonio & Sarda, Kisan & Nolè, Vittorio & Manfredi, Leonardo & Greco, Luca & Glielmo, Luigi & Del Vecchio, Carmen. (2021). Robust Statistics-based Anomaly Detection in a Steel Industry*.

[4]. N. Yan, L. Zhu, H. Yang, N. Li and X. Zhang, "Online Yarn Breakage Detection: A Reflection-Based Anomaly Detection Method," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-13, 2021, Art no. 5008813, doi: 10.1109/TIM.2021.3071227.

[5]. O. Abdelrahman and P. Keikhosrokiani, "Assembly Line Anomaly Detection and Root Cause Analysis Using Machine Learning," in IEEE Access, vol. 8, pp. 189661-189672, 2020, doi: 10.1109/ACCESS.2020.3029826.

[6]. Tian, Junfeng & Ding, Wei & Wu, Chunrui & Nam, Kwang. (2019). A Generalized Approach for Anomaly Detection From the Internet of Moving Things. IEEE Access. PP. 1- 1. 10.1109/ACCESS.2019.2945205.

[7]. Zhou, Xiaokang & Hu, Yiyong & Liang, Wei & Ma, Jianhua & Jin, Qun. (2020). Variational LSTM Enhanced Anomaly Detection for Industrial Big Data. IEEE Transactions on Industrial Informatics. PP. 1-1. 10.1109/TII.2020.3022432.

[8]. Kim, Kyung-Su & Oh, Seong & Cho, Hyun & Chung, Myung. (2021). One-Class Classifier for Chest X-Ray Anomaly Detection via Contrastive Patch-Based Percentile. IEEE Access. PP. 1-1.10.1109/ACCESS.2021.3136263.

[9]. A. Castellani, S. Schmitt and S. Squartini, "Real-World Anomaly Detection by Using Digital Twin Systems and Weakly Supervised Learning," in IEEE Transactions on Industrial Informatics, vol. 17, no. 7, pp. 4733-4742, July 2021, doi: 10.1109/TII.2020.3019788.

[10]. Fang, Xu & Guo, Wenhao & Li, Qingquan & Zhu, Jiasong & Chen, Zhipeng & Yu, Jianwei & Zhou, Baoding & Yang, Haokun. (2020). Sewer Pipeline Fault Identification Using Anomaly Detection Algorithms on Video Sequences. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.2975887.

[11]. Ma, Dandan, Yuan Yuan, and Qi Wang. 2019. "Hyperspectral Anomaly Detection Based on Separability-Aware Sample Cascade" Remote Sensing 11, no. 21: 2537.https://doi.org/10.3390/rs11212537

[12]. Tu, Bing, Nanying Li, Zhuolang Liao, Xianfeng Ou, and Guoyun Zhang. 2019. "Hyperspectral Anomaly Detection via Spatial Density Background Purification" Remote Sensing 11, no. 22: 2618.https://doi.org/10.3390/rs11222618.