Digital Object Identifier

Explainable Cyber Attack Detection in IoHT)

ABSTRACT The Internet of Things has impacted a plethora of industries and its role in healthcare services is quite significant. But circumstantial evidence points toward various kinds of attacks aimed toward this sector, this has given rise to requirements for increased security and countermeasures, especially considering the fact that the industry and its data can directly correlate with the lives of patients. Since, the value of health information and data is extremely high and a loss of that can directly result in lives being affected, a Network Based Intrusion Detection system (NIDS) or a Neural Network or Machine Learning Algorithm is needed for identifying and preventing cyber-attacks. Relying on Artificial Intelligence (AI) to help solve this global issue, we have created a Deep Neural Network (DNN) that uses the Dataset ECU-IoHT to detect the different types of attacks that might be implemented by anyone with malicious intent toward the industry. The system proposed in the paper achieves an extremely high percentage of accuracy with a low false positive rate as well. This provides a viable alternative to other detection methods.

I. INTRODUCTION

Internet of Things (IoT) is a term that is used to refer to the connection of electronic devices over the internet. This can be applied to various environments spanning from large scale mega-industries or even to a small home. IoHT or Internet of Health Things is a sub-type of IoT as a whole, this refers to all the possibilities of IoT in the health industry [1]. The goal is to connect all the possible electronics equipment and services and provide them to the patient as required and when required. These equipment can consist of MRI machines, ECG and Heart Rate monitors and even simpler equipment like blood pressure monitors.

This IoHT system is also used to store and transfer the patient data from the equipment to a data center [2]. The use of IoHT leads to an effective digitalization of hospital management data and makes a lot of processes smooth and stress-free.

The issue arises when security violations due to malicious attacks take place. These ‘cyber-attacks’ are all aimed at disrupting internal workings and may even act to steal patient data for ransom and the like [3]. The major classifications of the type of attack that might happen are, DoS or Denial of Service, Virus, Trojans, Worms and Botnet. Each aiming to achieve a different result in harming the IoHT service.

Therefore creating a secure environment is quite necessary.

In response to this, various detection methods have been created that perform various kinds of analysis and rely on advanced techniques like artificial neural networks or machine learning algorithms [4].

By relying on the deep learning method, a greater performance in detection can be observed, along with constant improvement due to its self-learning and adapting capabilities to detect even unknown attacks. Deep Neural Networks(DNN) is therefore used to detect any attacks to a system. This network can also be repurposed to help detect both inconsistencies, such as wrong or invalid data as well as detect health issues beforehand. These systems can also be used to detect any issues doctors might have failed to see or recognize, such as extremely early stages of cancer and even recognize early symptoms of dementia and other brain related health issues [5].

By taking note of all the above mentioned information, this paper aims to provide a working DNN model which can detects various kinds of cyber-attacks. The dataset we have used for this network is ECU-IoHT which has 111,207 samples, which helps to better train the model with enough samples to both train and test.

II. RELATED WORKS

The paper from Vijayakumar et.al [6] states the importance of developing a system that can detect intrusions and cyber-attacks in the healthcare domain. And hence, they have relied on the domain of Artificial Intelligence and have employed a deep level neural network based cyber-attack detection system. In doing so with the ECU-IoHT data set, they have achieved a high level of accuracy in detecting any intrusions.

Though this paper has put forth great claims and reliable evidence supporting the claims, there is still space for further growth in this domain. Our goal is in tangent to the one in this paper, better functionality as well as better reliability and accuracy.

Harb et.al. state the use of IoT and sensors to detect biological changes [7]. They rely on real-time data analysis to detect, adapt to and even try to predict any changes in a patient's body. The ideas and methodology used in this paper can be used alongside a neural network system to detect as well as predict any complications that might occur in the future. The neural network can then be used to accomplish two things simultaneously, detecting threats as well as monitor patient health. This type of monitoring can make it so that the patient in question's health data can always be monitored and stopgap measures can be implemented to alert surrounding nurses in the case of an emergency, this can be in the form of detecting lower heart rates, irregular heartbeats, sudden increase in temperature and such.

The paper by Ozkan et.al, checks the reliability of different methods of artificial intelligence in the detection and the prevention of cyber-attacks [8]. By testing the relative strength of different methods of this somewhat recent technological development. The article expresses the different ways Machine Learning, Deep Learning and Reinforcement Learning can be applied in the domain of cybersecurity and to what extent they can be relied on. They have even proposed ChatGPT as a tool that can be used to enhance cybersecurity. The conclusion of the article states the high potential behind the use of Artificial Intelligence in cybersecurity but also indicates that there is space for great improvements in this domain.

With the rapid development of IoT, the threats affecting it are also many. They aim to disrupt the networks and its services. The paper by A.Maghrabi et.al develops the use of Bald Eagle Search Optimization with a hybrid deep learning based botnet detection system. The goal is to use the system to detect botnets in an IoT environment [9]. The BESO idea originates from a study in 2019, where an algorithm is likened to the strategy employed by eagles in hunting fish. The algorithm consists of three stages; evaluating the optimization of the code after benchmarking, evaluating the optimization after comparing it to the various other algorithms and evaluating based on standard metrics (mean, standard deviation etc..). The integration of hybrid deep learning leads it to utilize the intricate nature of botnet attacks.

M. Centenaro et.al, in their paper, explain the usage of satellites to help an IoT system cover a wide geographical area for flexible and affordable coverage and connectivity [10]. The paper is a survey to provide solutions for IoT services in villages and remote areas. The more reliable solution is highlighted in the form of Low orbital satellites which offer an efficient solution. Of course there still exist issues like the overall loss of information due to the use of satellites, but these can be ironed out with further technological developments.

Most heart attacks result in death before patients can get any sort of viable treatment. This is due to the passive nature of how healthcare works. In their paper, C. Li et.al explain the necessity of being able to provide services using IoT techniques to change the mode of service to an active one instead of a reactive one [11]. This is needed to prioritize physical status over their feelings. The paper proposes a system consisting of two parts; data acquisition and data transmission. The goal isn't to process the data but recognize variations and transmit the data based on risk, type of medical analysis required and other physical signs (blood pressure, heart rate, ECG, glucose and blood fat). Relying on this form of invasive healthcare can prove very useful but it does possess the risk of invading privacy for data collection.

The data transfer rate of Cloud-based IoT structures is extremely fast. But the speed is affected by the immense nature of the total data produced by the IoT devices. In response to this, Bosri et. al provide a solution in the form of edge computing [12]. This facilitates a fast interaction between the devices and data centers, showing a drastic increase in transfer rate within a limited bandwidth. But there exists the concerns of data theft, this raises questions for data security and privacy. A recent solution to this is the introduction of blockchain technology for secure data transmission. The paper proposes the use of blockchain to ensure data security in storing user's data transfer records

Zachos et al. in their paper, have proposed a system to detect intrusions using an anomaly based intrusion detection system (AIDS) in the domain of IoT and the medical field (IoMT). The proposed methodology aims to collect log files reliably from devices as well as gateways and traffic to the IoMT network by leveraging host and network based technologies. To test this system six different ML algorithms were used and the best one was chosen after comparison.

This paper from Manimurugan et al. has put forward a Deep-Belief Network based on a deep learning model to detect intrusions into an IoT system. They have used the CICIDS 2017 dataset in regards to train the model. The goal of the proposed model is to remove the main concerns of security and privacy in the development of a smart environment. The model after testing has given them better results in all factors of evaluation such as the accuracy, recall, precision, F1-score, and detection rate upon comparison to various other classifiers.

By using a Fog ecosystem and relying on a Deep learning approach, Diro. et.al have proposed a model to detect any foreign attacks. Attack detection at a fog level was found to be a lot more scalable when compared to centralized cloud for IoT applications. Upon testing this model they have seen success with greater effectiveness in attack detection than centralized detection systems.

The goal of this paper by Anithi et.al is to create an intrusion detection system called Pulse. Pulse is a novel IDS for IoT which uses various Machine Learning techniques and can successfully detect and identify network scanning and probing while also being able to detect simple DoS attacks.

This was proposed due to the usage of IoT devices as botnets for example, Mirai malware.

The paper on IoT's integration in healthcare from Thamilarasu et.al highlights its transformative potential and the urgent need for robust security measures. While Internet of Medical Things (IoMT) devices offer enhanced patient monitoring and treatment, they also introduce security risks. Researchers propose innovative solutions like mobile agent-based intrusion detection systems, which use hierarchical, autonomous architectures and machine learning to detect network intrusions and data anomalies. Through simulations, these systems show high detection accuracy with minimal resource use, promising to secure IoMT networks and safeguard patient care integrity.

Syed et.al have reviewed 322 articles applying ML in ICUs using MIMIC data offering insights into application areas and treatment outcomes, guiding further adoption of ML in clinical decision-making and offering valuable lessons for healthcare researchers. To achieve and understand the high volume of patient data poses challenges for real-time decision-making. Machine Learning (ML) techniques, leveraging datasets like MIMIC, show promise in early high-risk event detection.

This paper from Fernandes et.al aims to comprehensively examine anomaly detection, covering background analysis and exploring key techniques, methods, and systems. Organized across five dimensions—network traffic anomalies, data types, intrusion detection system categories, detection methods, and open issues—the review provides an overview of this topic. It concludes by summarizing unresolved challenges and offering insights for future research in anomaly detection.

In this paper Zhang et.al explore how deep neural networks (DNNs) are used for anomaly detection, particularly in smart surveillance and industrial quality control. However, deploying DNNs on edge devices presents challenges due to resource constraints and latency requirements. Existing scaling techniques for DNN models often have long training times or disrupt the training process, leading to suboptimal performance. To address this, they introduce LightDNN, a method for scaling DNN models at the edge. LightDNN efficiently compresses DNN blocks and dynamically combines them, offering a large scaling space for high detection accuracy with minimal training and inference times.

The model proposed by Cauteruccio et.al delves into anomaly detection, expanding its scope beyond traditional domains to include social networking and internetworking. With the emergence of IoT and MIoT, anomaly detection gains significance. The paper introduces a methodological framework and defines a "forward problem" and an "inverse problem" in MIoT anomaly detection. It explores correlations between anomalies and factors like inter-node distances, network size, and centrality metrics of anomalous nodes.

This paper from Zhang et.al focuses on IoT network security, where traditional intrusion detection struggles. The paper proposes an adaptive intrusion detection model using a refined genetic algorithm (GA) and deep belief network (DBN). By dynamically adjusting network structures, the model enhances intrusion recognition on the NSL-KDD dataset while simplifying neural network complexity.

This paper from Banaamah et.al delves into the realm of cybersecurity, specifically focusing on its application in IoT devices. It highlights the significance of deep learning techniques, such as convolutional neural networks (CNNs), long short-term memory (LSTM), and gated recurrent units (GRUs), for intrusion detection. The study compares the performance of these methods and evaluates their effectiveness using a standard IoT intrusion detection dataset.

This paper from Almiani et.al explores cybersecurity in Fog computing for IoT. Fog computing brings cloud capabilities closer to IoT networks, improving data processing. Concerns over security have led to the development of intrusion detection systems, including AI-based models. This paper presents an automated intrusion detection system tailored for Fog security, using recurrent neural networks. Evaluation with the NSL-KDD dataset confirms its effectiveness, highlighting its suitability for Fog security applications.

III. PROPOSED METHODOLOGY

The design elements are from standard machine learning techniques. DNN followed by SHAP and LIME. Other models were used in a trial-and-error process and then dropped. Standard Data processing techniques, such as One-Hot Encoding in order to manipulate the data as required. By keeping effectiveness and quality in mind we have streamlined the working of the model and have focused on the practical element of the project. To refine our model's architecture, we employed an iterative feature selection process. Features were systematically dropped to assess their impact on the model's performance, which resulted in 'Source' and 'Destination' to be dropped as it decreased our model metrics and added noise to the model. We utilized LIME and SHAP techniques to interpret the model's decisions and identify the most influential features. Furthermore, we conducted hyperparameter tuning to optimize our model's performance. Parameters such as learning rate, dropout layers, optimizer selection, and batch size, Train-test splits were varied across different experiments. The performance of our model was evaluated using standard metrics such as Accuracy, Precision, Recall and F1-Score. These parameters provide a better insight into our model's effectiveness in classifying our instances. The proposed DNN consists of an input layer of 14 nodes, representing the input features after One-Hot Encoding the categorical features. This is followed by two dense layers of 64 nodes each with Rectified Linear Activation (Relu) activation function each. This was done to add non-linearity in the model, which is crucial in complex dataset. We utilized Adam optimizer to train the network as it provided the best results which we will see in the Experiment and Setup Section. Finally, the Soft-Max function was utilized at the output layer for multi-class classification.

Since there is a lack of publicly available datasets due to

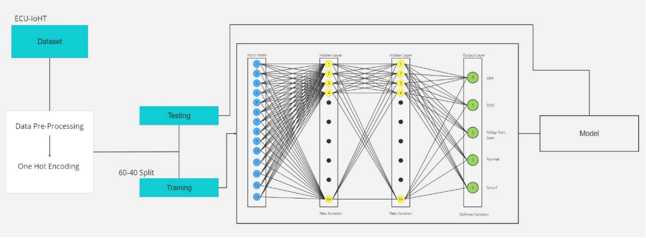

FIGURE 3.1. Proposed Model Architecture

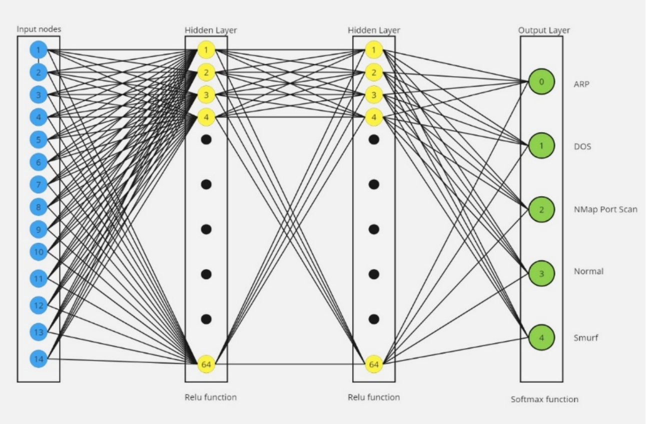

FIGURE 3 .2. DNN Model

privacy concerns, the dataset for this study, named ECU-IOHT, was obtained from an online Research Repository at Edith Cowan University. The dataset was specifically designed to aid the healthcare community in analyzing cyber-attacks and preventing various exploits. 4 The ECU-IOHT dataset comprises samples of 4 different types of attacks, including Smurf Attack, Denial of Service, ARP Spoofing, N-map Port Scan and Normal. With 111,207 rows and 9 columns, the dataset encompasses various features such as Time, Source, Destination, Protocol, Length, Info, Type, Type of Attack, and Sequential Identifier (No.), which will be disregarded for our analysis. Notably, only the "Length" feature contains numerical data, while the remaining features are categorical.

We excluded the ‘Serialized index’ and ‘Time’ due to it being considered unnecessary for our analysis. Additionally, due to the presence of 24, 000 unique values in ‘Info’ feature, which resulted in computational overhead when encoding we decided to remove this feature. Features such as ‘Source’ and ‘Destination’ were also dropped which are explained in Experimental Setup section.

Category	Count
ARP Spoofing	2359
DOS Attack	639
NMAP Port Scan Attack	6,836
Normal	23,453
Smurf Attack	77,920

TABLE 3.1. Description of the classes

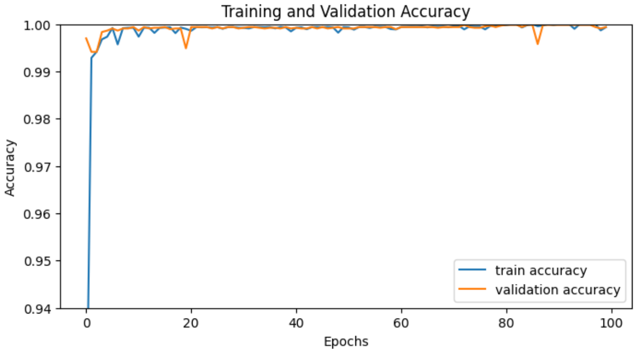

FIGURE 4.1. Training and Validation Accuracy

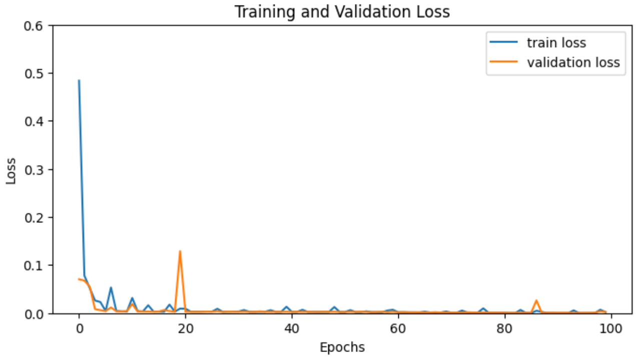

FIGURE 4.2. Testing and Validation Loss

Feature	Type	No.of Unique values
Source	Categorical	69
Destination	Categorical	71
Protocol	Categorical	11
Length	Numerical	-
Type	Categorical	2
Type of Attack	Categorical	5

TABLE 3.2. Description of the Features

To pre-process the remaining data, we employed One-Hot Encoding: Protocol, Type and Type of Attacks were One-hot encoded and ‘Length’ feature was normalized. This pre-processing lead to produce 111207 rows and 14 features for the independent variables and 111207 rows and 5 features for the dependent variable. We opted for One-Hot Encoding due to its ability to handle large numbers of unique values present in our categorical columns.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The model was trained with 100 epochs with 60-40 split which gave us the maximum accuracy and minimum validation loss as seen in figure 4.1 and 4.2.

Figure 4.3 shows the ROC curve where numbers 0, 1, 2, 3, and 4 denote ARP Spoofing, DoS attacks, Nmap attacks, normal behavior, and Smurf attacks, respectively. The ROC curve gives us the tradeoff between sensitivity and specificity. From figure 4.3 ROC is nearly 1 indicating that the model can

Model Metrics	Value
ACCURACY	99.943
PRECISION	99.944
RECALL	99.932
F1-SCORE	99.942

TABLE 4.1. Evaluation Metrics

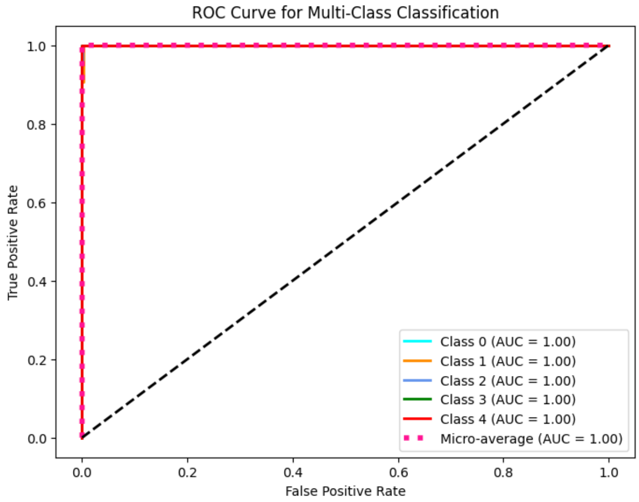

FIGURE 4.3. ROC Curve

distinguish between positive and negative class points.

Figure 4.4 displays the performance of the model using evaluation metrics. From our observation we can see that the model's performance is superior in detecting various attacks. As we can observe, only DOS attack has less performance metric than its counterparts and this would be due to the smaller number of samples: 639 out of 111,207 samples. 14 To better understand our model performance LIME was first used and the results are as follows: The LIME graphs show us which of the features is responsible for the model's output. As seen below the classification of ARP Spoofing is relied on Protocol ARP in classifying the output as ARP Spoofing whereas Protocol ICMP and Protocol TCP negatively impact the same, Fig 4.5. The predicted probabilities of 1.00 shows a strong presence towards class ARP Spoofing and 0 for the rest of the classes .

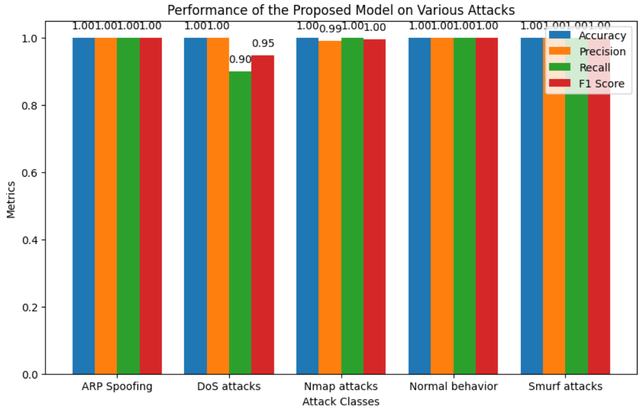

FIGURE 4.4. Performance of Model on various Attacks

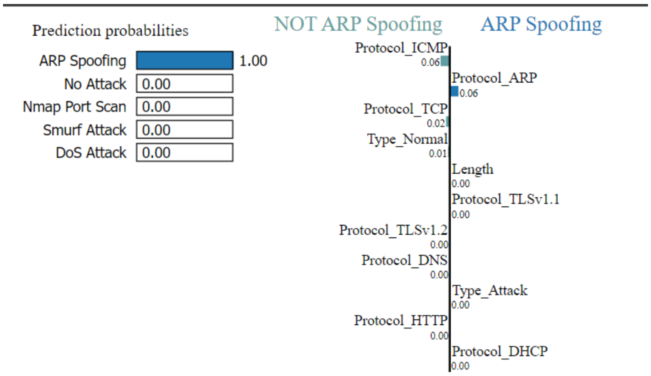

FIGURE 4.5. LIME plot on class ARP spoofing

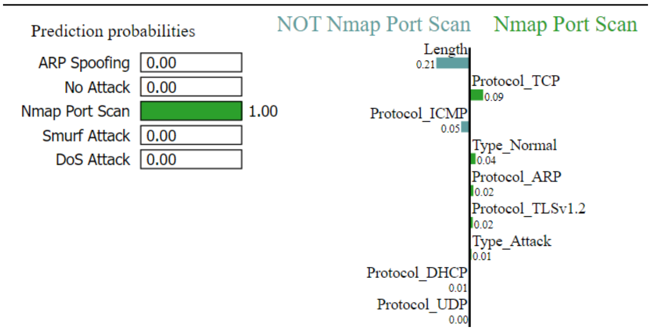

FIGURE 4.6. LIME plot on class NMAP Port Scan

For NMAP Port Scan at Fig 4.6 Protocol TCP, Type Normal and Protocol ARP have a positive effect on the models output whereas Protocol ICMP and Length have a negative impact. The predicted probabilities show 1.00 for NMAP Port Scan and 0 for the rest.

For DOS Attack Protocol ICMP, Type Attack, Protocol UDP and Protocol ARP are the positive features whereas Length, Protocol TCP, Type Normal are the negative features. The predicted probability score is 1.0 for NMAP Port Scan and 0 for the rest. Fig 4.7.

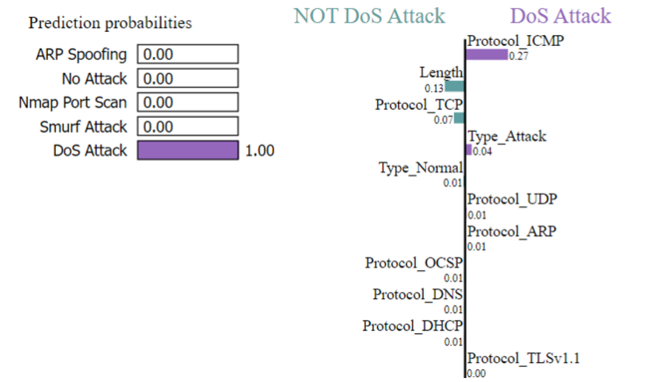

FIGURE 4.7. LIME plot on class DOS Attack

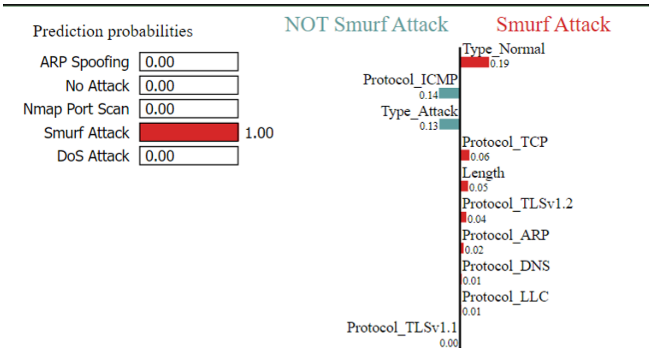

FIGURE 4.8. LIME plot on class Smurf Attack

For Smurf Attack Fig 4.8, Type Normal, Protocol TCP, Length, Protocol TSLv1,2, Protocol ARP, Protocol DNS, and Protocol LLC are the positive features and Protocol ICMP and Type Attack are the negative features. The predicted probability is 1.0 for Smurf Attack and 0 for the rest.

In the output class ’Normal,’ LIME encountered errors by failing to select relevant features. Additionally, LIME incorrectly identifies ’Type Normal’ as a positive indicator for cyberattacks, despite ’Type Attack’ being the correct indicator. This discrepancy is a significant limitation of LIME, largely due to the high dimensionality of the data. When employing perturbed data to predict local model behavior, LIME’s perturbed data fails to comprehensively represent the dataset, resulting in misclassifications and missing classifications. SHAP, on the other hand, addresses these challenges by providing more robust insights into model behaviors. The SHAP summary plot for class ARP Fig 4.9, Spoofing gives us the features which have a positive effect in model prediction, given in red. Removing the features which only blue streaks negatively affects the classification of DOS attack measured using evaluation metrics. The SHAP individual plot gives us the strength of that feature in predicting the model. As we can observe below:

A representation of the value predicted by the algorithm can be seen in Fig 4.10, with the contribution of each factor clearly visible. For the prediction ARP Spoofing protocol ARP is the sole contributor of the model prediction. Unlike in LIME where Protocol ICMP also contributed equally.

Similarly for class Nmap, SHAP summary plot gives us a clear indication of both Type Attack and protocol TCP being a key indicator of the model’s output as seen in Fig.4.11.

The SHAP force plot on the other hand in Fig.4.12 indicates that Protocol TCP has a negative impact on the model, and it isn’t clear which figure provides a clear indicator of our prediction.

One clear advantage SHAP has over LIME Is in its ability to explain the prediction ‘No Attack’ which LIME couldn’t.Fig.4.13 gives us the summary plot of ‘No Attack’ and we can observe type Normal is a key indicator of the model’s prediction in both as a positive indicator as well as a negative one as well as Protocol TCP.

The SHAP force plot at Fig.4.14 tells us that protocol ARP

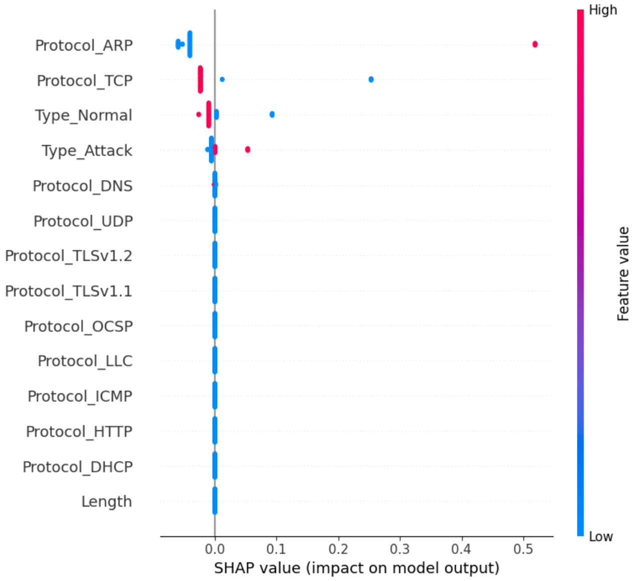

FIGURE 4.9. SHAP summary plot on ARP Spoofing

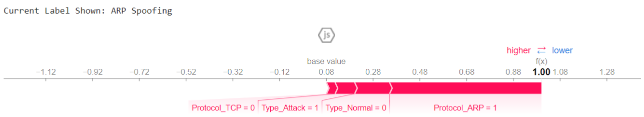
Current Label Shown: ARP Spoofing

FIGURE 4.10. SHAP force plot for ARP Spoofing

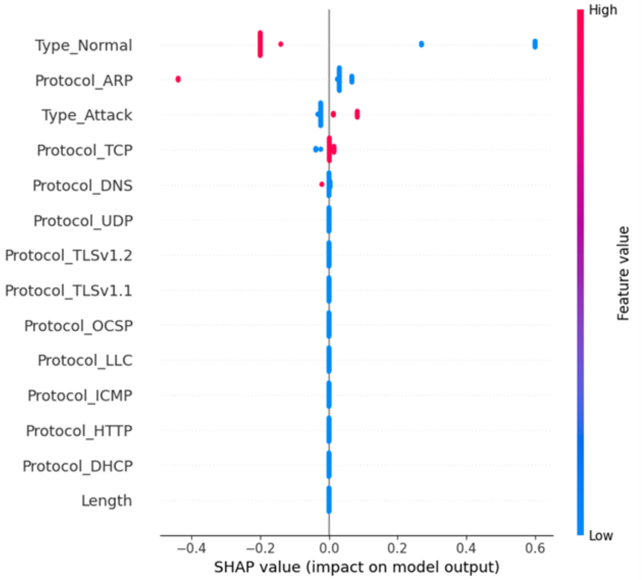

FIGURE 4.11. SHAP summary plot for Nmap Port Scan

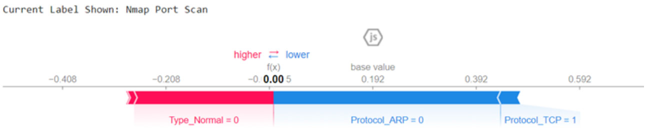
Current Label Shown: Nmap Port Scan

FIGURE 4.12. SHAP force plot for Nmap Port Scan

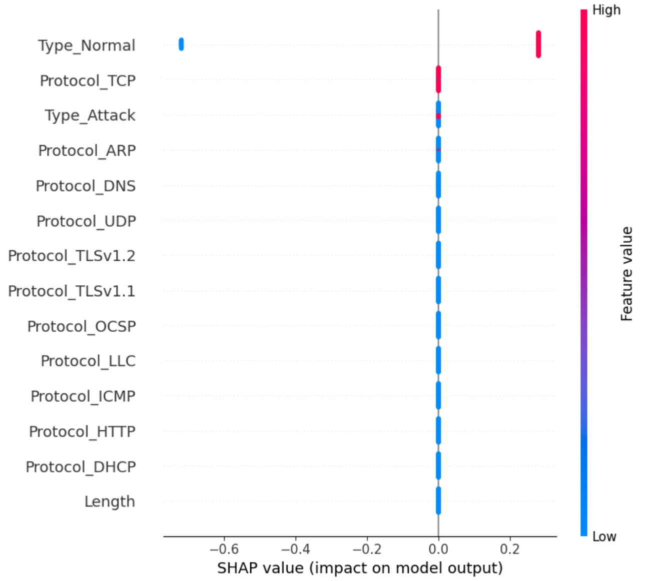

FIGURE 4.13. SHAP summary plot for Normal

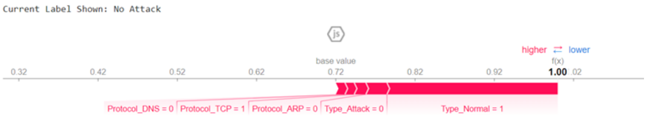

FIGURE 4.14. SHAP force plot for Normal

and Type Attack have a positive impact whereas Type Normal seems to have nullified as observed from summary plot.

This shows us that both algorithms were helpful in highlighting the contributions of the features in the model predictions. SHAP provided better results in terms of coherence and explainability of the model. Still the shapley values for the class DOS is a bit of a black box as it doesn't provide a clear explanation. This limitation is based on SHAP open-source platform where most of the features are towards trees rather than neural networks.

V. CONCLUSION

Our goal is to understand our black-box model using Explainable AI methods such as LIME and SHAP to protect healthcare devices from cyber-attacks. Both Explainable AI models give us interpretable results but LIME due its high dimensionality mismatches certain features whereas SHAP produces accurate results. The reduced DNN model with its less computations has increased accuracy than its predecessor.

REFERENCES

[1] M. Kumar, C. Kim, Y. Son, S. K. Singh, and S. Kim, "Empowering cyberattack identification in ioht networks with neighborhood component-based improvised long short-term memory," IEEE Internet of Things Journal, 2024.

[2] B. Amutha et al., "Iot revolutionizing healthcare: A survey of smart health-care system architectures," in 2023 International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE), pp. 1–5, IEEE, 2023.

[3] D. Sparrell, "Cyber-safety in healthcare iot," in 2019 ITU Kaleidoscope: ICT for Health: Networks, Standards and Innovation (ITU K), IEEE, 2019.

[4] S. Dasari and R. Kaluri, "An effective classification of ddos attacks in a distributed network by adopting hierarchical machine learning and hyperparameters optimization techniques," IEEE Access, 2024.

[5] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," nature, vol. 542, no. 7639, pp. 115–118, 2017.

[6] K. P. Vijayakumar, K. Pradeep, A. Balasundaram, and M. R. Prusty, "Enhanced cyber attack detection process for internet of health things (ioht) devices using deep neural network," Processes, vol. 11, no. 4, p. 1072, 2023.

[7] H. Harb, A. Mansour, A. Nasser, E. M. Cruz, and I. de la Torre Diez, "A sensor-based data analytics for patient monitoring in connected healthcare applications," IEEE Sensors Journal, vol. 21, no. 2, pp. 974–984, 2020.

[8] M. Ozkan-Ozay, E. Akin, Ö. Aslan, S. Kosunalp, T. Iliev, I. Stoyanov, and I. Beloev, "A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions," IEEE Access, 2024.

[9] L. A. Maghrabi, S. Shabanah, T. Althaqafi, D. Alsalman, S. Algarni, A. Abdullah, and M. Ragab, "Enhancing cybersecurity in the internet of things environment using bald eagle search optimization with hybrid deep learning," IEEE Access, 2024.

[10] M. Centenaro, C. E. Costa, F. Granelli, C. Sacchi, and L. Vangelista, "A survey on technologies, standards and open challenges in satellite iot," IEEE Communications Surveys & Tutorials, vol. 23, no. 3, pp. 1693–1720, 2021.

[11] C. Li, X. Hu, and L. Zhang, "The iot-based heart disease monitoring system for pervasive healthcare service," Procedia computer science, vol. 112, pp. 2328–2334, 2017.

[12] R. Bosri, A. R. Uzzal, A. Al Omar, M. Z. A. Bhuiyan, and M. S. Rahman, "Hidechain: A user-centric secure edge computing architecture for healthcare iot devices," in IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 376–381, IEEE, 2020.

• • •