2022 25th International Conference on Computer and Information Technology (ICCIT)
17-19 December, Cox's Bazar, Bangladesh

IoT Network Attack Detection Using XAI and Reliability Analysis

Sabrina Tabassum
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
sabrina.tabassum@g.bracu.ac.bd

Nazia Parvin
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
nazia.parvin@g.bracu.ac.bd

Nigah Hossain
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
nigah.hossain@gmail.com

Anika Tasnim
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
tasnimanika59@gmail.com

Rafeed Rahman
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
rafeedrahmansham2015@gmail.com

Muhammad Iqbal Hossain
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh
iqbal.hossain@bracu.ac.bd

***Abstract*—IoT has emerged as one of the most sophisticated techniques in recent years. But inadequate security controls are the most typical barrier to IoT expansion as the devices transmit a huge amount of data. Nowadays different machine learning and deep learning models are used to detect various IoT attacks. In our previous research, we applied Decision Tree, Random Forest, AdaBoost, XGBoost, ANN, and MLP to the IoT/IIoT dataset of TON IoT datasets to classify IoT network attacks. In binary classification, we got above 96% accuracy for all methods. In contrast, AdaBoost and ANN underperformed in multiclass classification. As accuracy improves, models get more complicated, and these models are frequently seen as black boxes that are difficult to interpret. Though these models give highly precise results, an explanation is required in order to comprehend and accept the models' decisions. Here comes XAI which emphasizes a variety of ways for breaking the black-box nature of Machine Learning and Deep Learning models as well as delivering human-level explanations. In this article, we have extended our work by analyzing different machine learning and deep learning methodologies using XAI to explain the categorization of IoT network attacks. LIME, SHAP, and ELI5 approaches have been used to interpret and explain which will increase transparency and reliability.**

Keywords— IoT attacks, XAI, Machine Learning, Deep Learning, LIME, SHAP, ELI5

I. Introduction

IoT technology can collect, measure and understand environmental data, allowing modernizations that raise living standards [1]. IoT systems are intricate and contain a number of interconnected components. As a result, maintaining the security requirement in an IoT system with a huge area of attack is difficult. Solutions must adopt a holistic approach to satisfy the security needs [2]. Moreover, some lightweight IoT devices have a lack of memory management units (MMU). Consequently, it is very easy for attackers to infiltrate these devices by exploiting memory vulnerabilities. The typical high-end security protection solution cannot function correctly in a system with limited resources and diverse components [3].

Different machine learning (ML) and deep learning (DL) approaches are great techniques to safeguard connected IoT devices from sophisticated network threats. In the previous research [4], a detailed analysis was performed using six different machine learning and deep learning models for both binary and multiclass classification. A highly precise result was achieved from most of the models. But the approaches from the ML and DL models are difficult to comprehend and analyze due to their diversity and nature [5]. AI models are frequently referred to as "black boxes" in the literature, especially when they come from ML or DL [6]. The lack of transparency in such approaches negatively influences user to embrace these models [7] . Traditional methods are no longer adequate because of the black-box nature brought on by ML and DL; instead, new rules and strategies are required [8]. And here comes the concept of Explainable AI (XAI).

XAI emphasizes a variety of ways for breaking the black box nature of machine learning and deep learning models as well as delivering a human level explanation [9]. XAI techniques seek to understand what the AI system discovered during training and how decisions are made for specific or new occurrences during the prediction process [5]. In domains where explanation and openness of the model's operation are essential, XAI has been studied more. Some domains are- cybersecurity-Intrusion Detection Systems (IDS) [10], finance-risk management [11], classifying military target [12].

In this paper, we have explained the predictions that were made by the Machine Learning and Deep Learning models for finding different network attacks on the IoT devices, using the XAI. We have used LIME, SHAP, and ELI5 approaches to interpret and explain the results and output created from machine learning algorithm. LIME is a local interpretable model that offers a local explanation for each prediction, whereas SHAP uses Shapely values to provide both a local and global explanation. Furthermore, the 5 in ELI5 corresponds to a 5-year-old kid, implying that the individual seeking clarification has limited knowledge of the subject. ELI5 shows the feature weight from the most significant features to the least significant features both locally and globally. It also incorporates colors that range from dark to light in order to highlight the contribution of the features.

The following sections of the paper are organized in the following order: Section II discusses the literature review, which bears pertinent similarities to the concepts in our paper.

979-8-3503-4602-2/22/$31.00 ©2022 IEEE

The explainable artificial intelligence methodologies' along with visualizations of their successes and failures is discussed in Section III. The analysis and findings obtained from them and the future work of our paper is covered in Section IV's conclusion.

II. Related Works

In the previous research [4] , the ToN IoT dataset had been used which integrates data gathered from telemetry datasets of IoT and IIoT sensors, demonstrating the viability of the methods used in that research. Decision Tree, Random Forest, AdaBoost, XGBoost, ANN, and MLP had been applied to analyze the IoT network attack. Decision Tree, which works on the decision made on the decision node, had given the highest performance of more than 99% in both binary and multiclass classification. Random Forest, an extremely random approach of using decision trees in predictive data mining, produced more than 99% in both binary and multiclass classification. ANN which is a machine learning algorithm was also successful in giving a high performance in the binary classification but failed to classify in multiclass classification. AdaBoost, which provides a simple and efficient way to create ensembles, also takes the performance of each individual classifier as well as the variety of the ensemble's members' classifiers into account. AdaBoost also underperformed for multiclass classification. Extreme Gradient Boosting and Multilayer Perceptron were also used with a moderate outcome. Furthermore, the dataset was balanced using Undersampling, Oversampling and SMOTE.

This work [14] applies the XGBoost classifier to the IoT Intrusion Dataset, a dataset that supports both binary and multi-class classification in order to provide a model and practical tool for detecting IoT attacks. By detailing the detection model's method for predicting attacks and articulating the variables or features that influence the associated prediction, Explainable Artificial Intelligence is utilized to enhance the suggested model's accuracy and human understanding.

III. Methodology and Result

In our previous work [4], the applied Machine Learning and Deep Learning models in the datasets were evaluated based on matrices like accuracy, precision, recall and F1-Score, for both Binary and Multiclass classifications. Here Fig. 1 shows the accuracy of the Binary class classification of datasets and Fig. 2 shows the accuracy of multiclass classifications of datasets.

The explainable artificial intelligence methodologies tools that we have applied here are LIME, SHAP and ELI5.

LIME (Local Interpretable Model-agnostic Explanations) is a resource for comprehending and interpreting the underlying ML model at the same time being model-neutral [15]. The idea behind its introduction by [16], is that it can approximate the machine learning models with an easily comprehendible model [17]. It displays the local model that it generates by globally approximating the black-box model [16], [18].

Fig. 3 and Fig. 4 show the LIME explanations of the predictions for two input instances using Decision Tree and XGBoost classifier on the Modbus dataset and GPS Tracker dataset respectively. We can see from these two figures that each of these explanations is divided into three parts: prediction probabilities for all possible outputs, the bar charts for all outputs which represent the weights and contribution of each feature in the prediction and a feature value table. According to Fig. 3, the Decision Tree classifier has accurately predicted the output for a particular input data where the prediction probability of the "backdoor" is 1.00 and 0.00 for others. From the feature value table, we can observe that features with Maroon color support 'backdoor' as a prediction outcome whereas features with Blue support the outcome of 'not backdoor'. Therefore, the backdoor is predicted as an output for the features 'FC3_Read_Holding_Register_n' and 'FC4_Read_Coiln'. These two features have a positive impact on the prediction since the feature value of 'FC4_Read_Coiln' is less than or equal to -0.50 and the feature value of 'FC3_Read_Holding_Register_n' is larger than -0.50. The 'date', 'time', 'FC_Read_Input_Register_n' and 'FC2_Read_Discrete_Value_n' have a negative impact on predicting the output as a 'backdoor'.

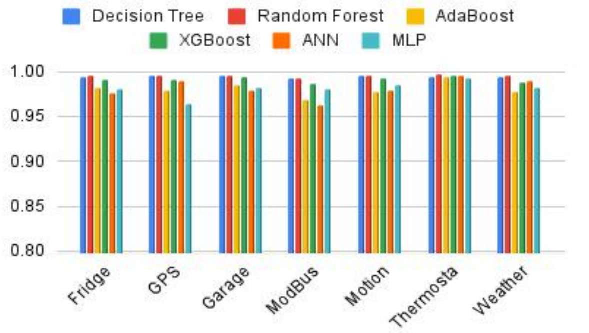


Fig. 1. Accuracy of Binary Classification of Datasets

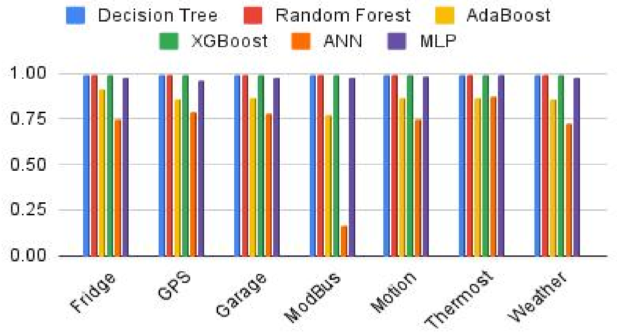


Fig. 2. Accuracy of Multiclass Classification of Datasets

From Fig. 4, we can see that for an input instance the XGBoost classifier could not predict the output correctly. It has predicted the output as "ddos" but the actual output would be "normal". The prediction probability is 0.85 for 'ddos', 0.14 for 'normal' and 0.00 for other possible outputs. From the feature value table, we can observe that 'time', 'date' and 'latitude' are the features that have a positive impact on the predicted output and for these features' contribution this prediction is made. The feature 'longitude' has a negative impact on the prediction, whereas it has a positive impact for 'normal' output since it has a feature value of 0.31which is in the range of 0<longitude<=0.47.

The SHAP (short for SHapely Additive explanation) is a

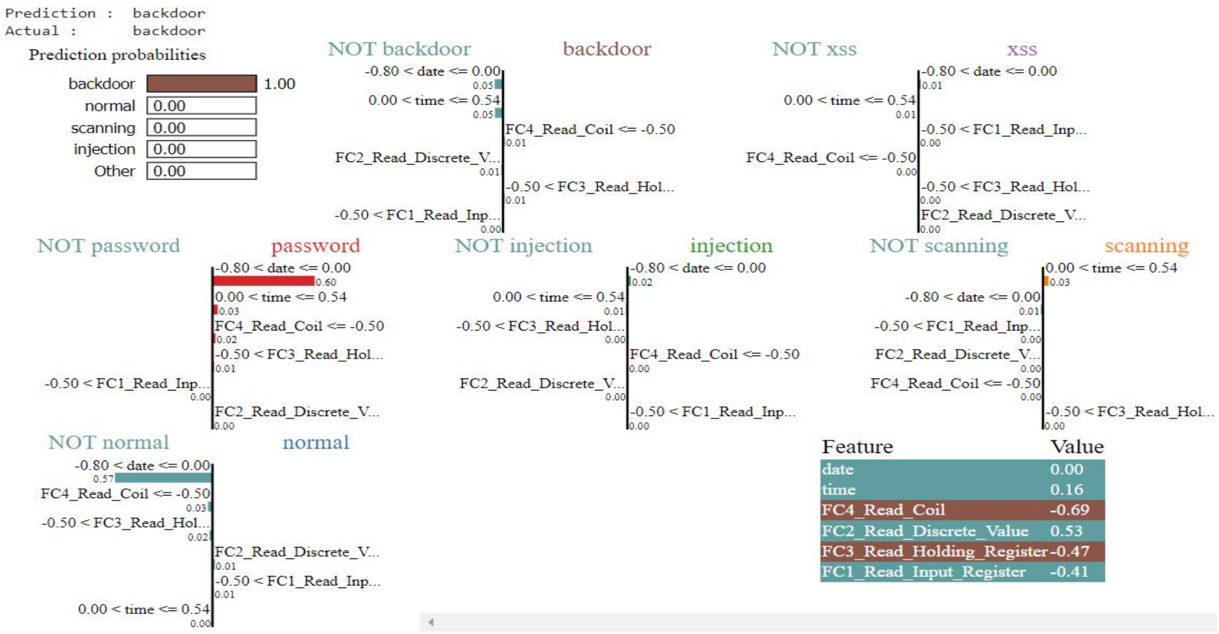


Fig. 3. LIME Explanations of Decision Tree on Modbus Dataset

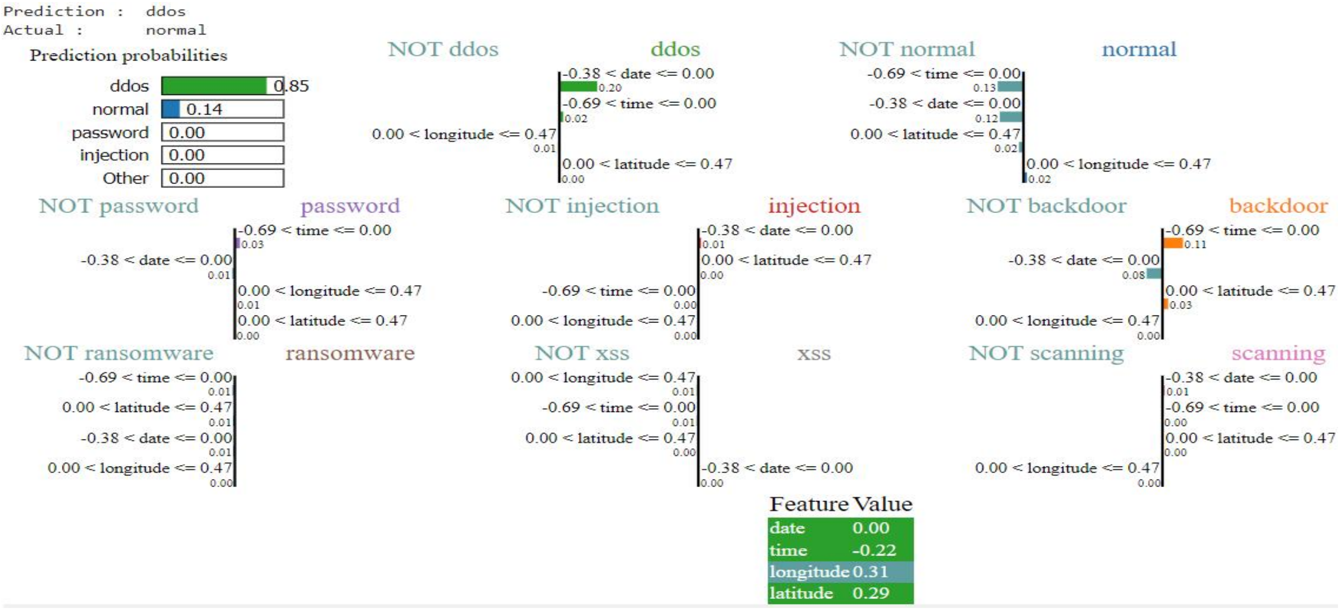


Fig. 4. LIME Explanations of XGBoost on GPS Tracker Dataset

framework of XAI. In SHAP, the variability of the prediction is split among the obtainable variables, which allows each explanatory variable's contribution to each point of prediction to be evaluated, whatever the prime model is. It uses Shapely values to describe how each feature affected the prediction [19].

Fig. 5 shows the average impact of each feature on the prediction of the IoT Fridge dataset using the Decision Tree model. This summary plot gives a global explanation of this model. The names of the features are displayed on the Y-axis in descending order of contribution, from highest (time) to lowest (fridge_temperature). The X-axis represents the

.

absolute means of SHAP values. In this figure, 7 various colors symbolize various classes. There are seven different classes in this figure. We can see from the figure that the 'time' feature contributes the most to all classes. Moreover, the 'time' and 'date' features have a significant impact on all classes. Whereas other classes have ignored this feature, 'temp_condition' has a minor influence on 'normal', 'ddos', and 'ransomware'. The 'fridge_temperature' has the least influence and almost all the classes have completely ignored this feature.

Fig. 6 shows the summary plot of SHAP for the Fridge dataset using Decision Tree. The dots in the plot are the data points of the IoT Fridge dataset. The color represents feature value, with red being high and blue representing low. On the Y-axis, the names of the features are given in descending order of contribution, from highest (date) to lowest (temp_condition). The X-axis shows SHAP values which shows whether the feature contribution has a positive impact or negative. As there are seven different classes in this dataset, there can be a total of seven possible outcomes. For each possible outcome, there is an array of SHAP values. So for this plot, we have used shap values for a possible outcome. From this figure, we can see that the 'date' is the most important feature, whereas 'temp_condition' is the least important. The feature 'date' has a high feature value which has a positive impact on the outcome. We can also observe that, as the feature value of 'time' is increasing, the positive impact on the prediction is also increasing. Though the feature 'temp_condition' has no effect on model prediction, there are two extreme cases where a high value has a negative impact and another high value has a positive impact on the model output.

The summary plot of SHAP for the Fridge dataset using MLP is shown in Fig. 7. The data points from the IoT Fridge dataset are represented by the dots in the figure. The names of the characteristics are shown on the Y-axis in descending order of contribution, from highest (sphone_signal) to lowest (door_state). The X-axis displays SHAP values, which indicate whether the feature contribution has a positive or negative influence. According to this figure, the 'sphone_signal' property is the most significant, while the 'door_state' feature is the least important. The feature 'sphone_signal' has a low feature value, which has a favorable influence on the outcome. We can also see that when the feature value of 'door_state' increases, so does its positive influence on prediction.

ELI5 stands for the phrase, 'Explain Like I'm 5'. It is used to verify and explain machine learning classifier predictions. The main use of ELI5 is to debug algorithms. There are primarily two approaches to examine a classification or regression model: (1) Examine model parameters to see how the model works on a larger scale. (2) Examine a model's individual prognosis and try to understand why it makes the decisions it does.

In Fig. 8, we have used the show_weights() method of ELI5 to show the feature importance in descending order for the prediction of the Random Forest model on the IoT Fridge dataset. We can use this to find out how this model works on a global scale. According to the figure, the feature 'date' has the highest approximate weight of 0.4929, while the feature 'fridge_temperature' has the lowest importance with an approximate weight of 0.0090. We can also observe that the 'date' and 'time' features are in Green color but the 'temp_condition' and 'fridge_temperature' features are not. The Green color represents the most significant feature with a positive contribution whereas Red represents the least significant feature with a negative impact. Because there are no Red features in this table and the feature weights are all positive, we can conclude that all of the features have contributed to the model prediction.

In Fig. 9, we have used the show_prediction() method of

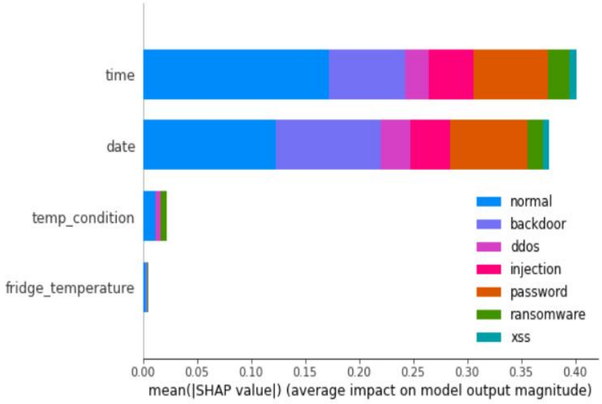


Fig. 5. Feature importance calculated by SHAP value of Fridge dataset

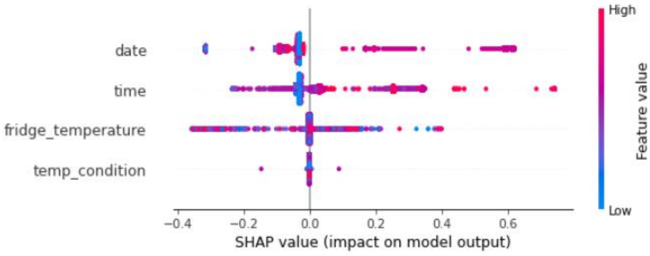


Fig. 6. SHAP Summary Plot for Fridge dataset using Decision Tree

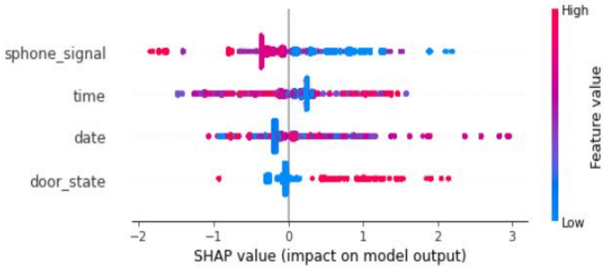
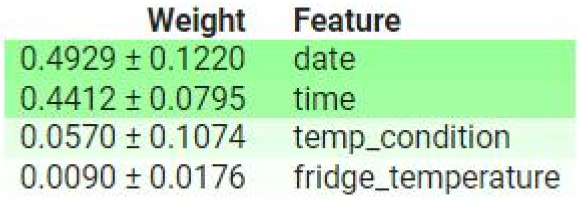


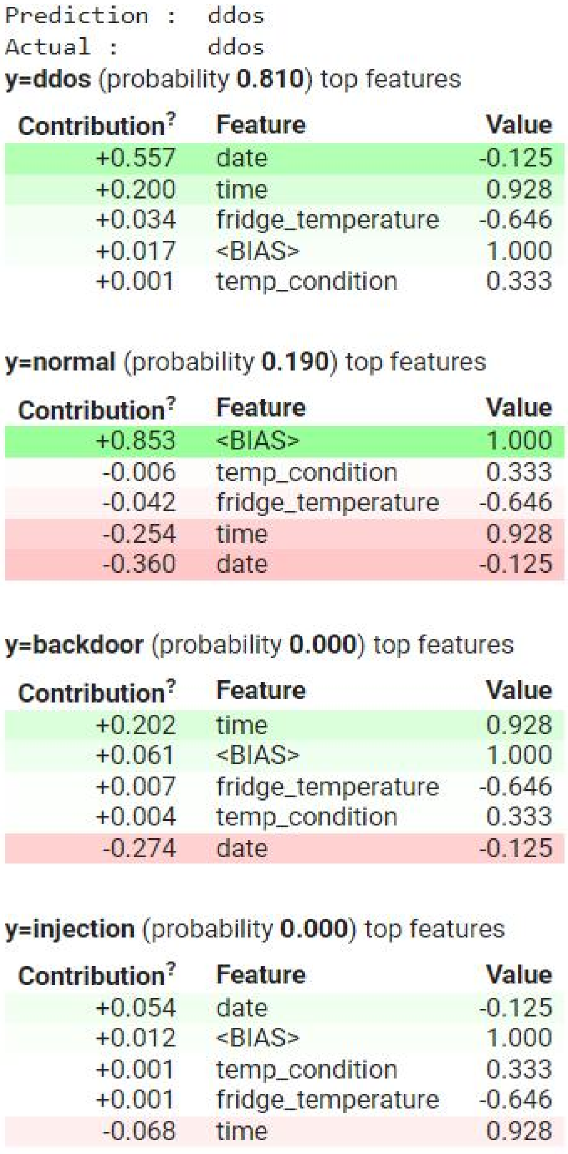
Fig. 7. SHAP Summary Plot for Fridge dataset using MLP



Weight	Feature
0.4929 ± 0.1220	date
0.4412 ± 0.0795	time
0.0570 ± 0.1074	temp_condition
0.0090 ± 0.0176	fridge_temperature

Fig. 8. Feature importance of Fridge dataset using ELI5

ELI5 for a single instance. This allows us to see how each feature contributes to the prediction of a certain input sample. The Random Forest model has accurately predicted the output for an input sample of the IoT fridge dataset. We have observed the top four targets for this prediction. The green color represents the most significant features of that prediction, while the red color represents the least important. According to this figure, the prediction probability of ‘ddos’ is 0.810, whereas the likelihood of ‘normal’ is 0.190. Other targets have a 0% chance of being predicted. The ‘date’ and ‘time’ are the two most important features with the contribution of (+0.557) and (+0.200) respectively for this particular prediction. The actual values of ‘date’ and ‘time’ are (-0.125) and (0.928) respectively. However, with the target class ‘normal’, a completely opposite scenario can be observed. These ‘time’ and ‘date’ features contribute the least to the target ‘normal’. Similarly, the feature ‘date’ contributes the least to the ‘backdoor’ attack, while ‘time’ contributes the least to the ‘injection’ attack.



```
Prediction :  ddos
Actual :      ddos
```

y=ddos (probability **0.810**) top features

Contribution?	Feature	Value
+0.557	date	-0.125
+0.200	time	0.928
+0.034	fridge_temperature	-0.646
+0.017	<BIAS>	1.000
+0.001	temp_condition	0.333

y=normal (probability **0.190**) top features

Contribution?	Feature	Value
+0.853	<BIAS>	1.000
-0.006	temp_condition	0.333
-0.042	fridge_temperature	-0.646
-0.254	time	0.928
-0.360	date	-0.125

y=backdoor (probability **0.000**) top features

Contribution?	Feature	Value
+0.202	time	0.928
+0.061	<BIAS>	1.000
+0.007	fridge_temperature	-0.646
+0.004	temp_condition	0.333
-0.274	date	-0.125

y=injection (probability **0.000**) top features

Contribution?	Feature	Value
+0.054	date	-0.125
+0.012	<BIAS>	1.000
+0.001	temp_condition	0.333
+0.001	fridge_temperature	-0.646
-0.068	time	0.928

Fig. 9. ELI5 explanation of a prediction using Random Forest in Fridge dataset

IV. Conclusion

Most IoT devices are less secure and the attacks on those devices threaten data security and negatively impact people's lives and customer trust. Although machine learning and deep learning models may identify and classify various network attacks, they seem to users as black box models due to a lack of interpretability and explainability. Subsequently, in our earlier research, we analyzed several ML and DL models to identify and classify various IoT network attacks whereas, for binary classification, we obtained an accuracy of more than 96% for all models. In this article, we have extended our work by applying three Explainable AI (XAI) methods: LIME, SHAP, and ELI5 to explain how these black box models are working under the hood, why one prediction is made instead of another, and the factors which influence to make any prediction. LIME provides a local explanation for every given prediction, but SHAP provides both a local and global explanation using SHAP values. Furthermore, ELI5 color-codes the most significant features to the least significant features both locally and globally. These explanations will increase human understanding of those models in classifying those attacks as well as it will improve customer trust. We couldn't use real-time data for our experiment due to device limits. In the future, we will attempt to evaluate the classifiers using real-time data. We will also try to incorporate an alarm system in case any of the attacks are detected.

V. Reference

[1] B. Jovanovic, "Internet of Things statistics for 2022 - Taking Things Apart," May 2022. [Online]. Available: https://dataprot.net/statistics/iot-statistics/. [Accessed 10 September 2022].

[2] J. P, J. Shareen, A. Ramdas and H. A P, " Intrusion Detection System for IOT Botnet Attacks Using Deep Learning," *SN Computer Science,* vol. 2, no. 3, p. 205, 2021.

[3] W. Zhou, Y. Jia, A. Peng, Y. Zhang and P. Liu, "The Effect of IoT New Features on Security and Privacy: New Threats, Existing Solutions, and Challenges Yet to Be Solved," *IEEE Internet of Things Journal,* vol. 6, no. 2, pp. 1606-1616, 2019.

[4] A. Tasnim, N. Hossain, N. Parvin, S. Tabassum, R. Rahman and M. I. Hossain, "Experimental Analysis of Classification for Different Internet of Things (IoT) Network Attacks Using Machine Learning and Deep learning," in *2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 2022.

[5] M. Gashi, M. Vuković, N. Jekic, S. Thalmann, A. Holzinger, C. Jean-Quartier and F. Jeanquartier, "State-of-the-Art Explainability Methods with Focus on Visual Analytics Showcased by Glioma Classification," *BioMedInformatics,* vol. 2, no. 1, pp. 139-158, 2022.

[6] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access,* vol. 6, pp. 52138-52160, 2018.

[7] D. Castelvecchi, "Can we open the black box of AI?," *Nature,* vol. 538, no. 7623, pp. 20-23, 2016.

[8] F. Königstorfer and S. Thalmann, "Software documentation is not enough! Requirements for the documentation of AI," *Digital Policy, Regulation and Governance,* vol. 23, no. 5, pp. 475-488, 2021.

[9] N. G., "How many iot devices are there in 2021?[all you need to know]," 12 August 2022. [Online]. Available: https://techjury.net/blog/how-many-iot-devices-are-there/#gref. [Accessed 2022].

[10] D. L. Marino, C. S. Wickramasinghe and M. Manic, "An Adversarial

Approach for Explainable AI in Intrusion Detection Systems," in *IECON 2018 - 44th Annual Conference of the IEEE Industrial Electronics Society*, 2018.

[11] J. Adams and H. Hagras, "A Type-2 Fuzzy Logic Approach to Explainable AI for regulatory compliance, fair customer outcomes and market stability in the Global Financial Sector," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2020.

[12] M. H. S. Pannu and A. Malhi, "Deep learning-based explainable target classification for synthetic aperture radar images," in *2020 13th International Conference on Human System Interaction (HSI)*, 2020.

[13] R. K. Muna, H. T. Maliha and M. Hasan, "Demystifying machine learning models for IOT attack detection with explainable AI," Brac University, Dhaka, 2021.

[14] M. Kuzlu, U. Cali, V. Sharma and O. Guler, "Gaining Insight Into Solar Photovoltaic Power Generation Forecasting Utilizing Explainable Artificial Intelligence Tools," *IEEE Access,* vol. 8, pp. 187814-187823, 2020.

[15] M. T. Ribeiro, S. Singh and C. Guestrin, "Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[16] M. T. Ribeiro, S. Singh and C. Guestrin, "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction," 12 August 2016. [Online]. Available: https://www.oreilly.com/content/introduction-to-local-interpretable-model-agnostic-explanations-lime/.

[17] T. Peltola, "Local interpretable model-agnostic explanations of Bayesian predictive models via kullback-leibler projections," arXiv:1810.02678, 2018. [Online]. Available: https://arxiv.org/abs/1810.02678.

[18] R. Saluja, A. Malhi, S. Knapič, K. Främling and C. Cavdar, "Towards a rigorous evaluation of explainability for multivariate time series," arXiv preprint arXiv:2104.04075 , 2021.

[19] N. Prentzas, A. Nicolaides, E. Kyriacou, A. Kakas and C. Pattichis, "Integrating machine learning with symbolic reasoning to build an explainable AI model for stroke prediction," in *In 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, 2019.