2024 IEEE International Mediterranean Conference on Communications and Networking (MediCom) | 979-8-3503-0948-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/MediCom61057.2024.10621339

Explainable AI for Comparative Analysis of Intrusion Detection Models

Pap M. Corea[1A], Yongxin Liu[1a], Jian Wang[2b], Shuteng Niu[3c], Houbing Song[4d]
[1]Embry-Riddle Aeronautical University, FL 32114 USA, [2]University of Tennessee at Martin, TN 38237 USA
[3]Bowling Green State University, OH 43403 USA, [4]University of Maryland, Baltimore County, MD 21250 USA
[a]{moctarp@my.erau.edu, LIUY11@erau.edu}, [b]jwang186@utm.edu, [c]sniu@bgsu.edu, [d]songh@umbc.edu

***Abstract*—Explainable Artificial Intelligence (XAI) has become a widely discussed topic, the related technologies facilitate better understanding of conventional black-box models like Random Forest, Neural Networks and etc. However, domain-specific applications of XAI are still insufficient. To fill this gap, this research analyzes various machine learning models to the tasks of binary and multi-class classification for intrusion detection from network traffic on the same dataset using occlusion sensitivity. The models evaluated include Linear Regression, Logistic Regression, Linear Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, Decision Trees, and Multi-Layer Perceptrons (MLP). We trained all models to the accuracy of 90% on the UNSW-NB15 Dataset. We found that most classifiers leverage only less than three critical features to achieve such accuracies, indicating that effective feature engineering could actually be far more important for intrusion detection than applying complicated models. We also discover that Random Forest provides the best performance in terms of accuracy, time efficiency and robustness. Data and code available at https://github.com/pcwhy/XML-IntrusionDetection.git**

I. Introduction

Machine learning (ML) has emerged as a transformative tool in the field of intrusion detection, providing a robust approach to enhancing cybersecurity measures. By leveraging the ability to learn from and adapt to evolving data without explicit programming, ML techniques can effectively identify novel and sophisticated cyber threats. This adaptive capability is crucial in an environment where attackers continuously modify their strategies to evade detection. ML algorithms, including supervised, unsupervised, and reinforcement learning, analyze patterns and anomalies in vast datasets, enabling the prediction and detection of potential intrusions with high accuracy. As such, the application of ML in intrusion detection systems (IDS) represents a significant step forward in developing dynamic, responsive security strategies that can anticipate and mitigate threats in real-time, thus ensuring the integrity and confidentiality of information systems.

Despite the efficacy of machine learning in intrusion detection, the deployment of these technologies raises significant concerns, particularly regarding the opaque nature of certain ML models. Black-box models, such as deep neural networks, often lack transparency in their decision-making processes, making it challenging for cybersecurity professionals to interpret or trust the rationale behind specific detections or classifications [1]. This uncertainty can complicate compliance with regulatory standards that demand clear audit trails and explainability of security systems. Furthermore, the inability to interpret model decisions can hinder the identification and correction of biases in training data, potentially leading to unfair or ineffective security measures. Such limitations underscore the need for developing more interpretable machine learning models and methods that maintain high detection performance while providing greater transparency and accountability in their operations [2].

In the landscape of explainable AI (XAI), several methods stand out for their ability to render machine learning models more interpretable, especially in critical applications like intrusion detection. LIME (Local Interpretable Model-agnostic Explanations [3]) is another key technique that approximates the locally predictive behavior of the model around a specific instance, thus providing insights into the decision-making process. SHAP (SHapley Additive exPlanations [4]) assigns each feature an importance value for a particular prediction, integrating game theory to ensure consistency and accuracy in feature attribution. Grad-CAM (Gradient-weighted Class Activation Mapping [5]) uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting important regions for predictions. Some research even use GradCAM to analyze the potential vulnerabilities within deep neural networks [6], [7], but such method is only applicable to algbraically differentiable models. Occlusion Sensitivity [8] investigates the influence of different parts of input data on the output by systematically occluding sections of the data and observing the changes in output. This method is particularly useful for identifying which data segments are most critical for decision-making, offering clear visual explanations. Each of these methods offers a different approach to enhance transparency in ML models, but Occlusion Sensitivity is especially valuable for its direct and intuitive visualization capabilities.

In this paper, we utilize Occlusion Sensitivity to analyze the decision behavior of different machine learning models trained on the UNSW-NB15 Dataset [9], which captures network traffic traces of a hybrid of real modern normal activities and synthetic contemporary attacks. We compare the behavior of various classifiers and ovr findings are as follows:

- Most of our models only exploit less than three highly ranked features regardless if feature selection.
- The classifiers depends highly on time-dependent features, such as Source or Destination Time to Live (*ttl*),

and thus can generate highly domain-specific models.
- We show that
- Random Forest show the best robust by treating all input features equally, and therefore, it still maintains the best performance while top features are not available.

The remainder of this paper is organized as follows: A literature review of related work is presented in Section II. We present the methodology in Section III. Evaluation and discussion are presented in Section IV and conclusions in Section V.

II. Related Work

Statistical machine learning is characterized by its ability to model complex data through probabilistic approaches, enabling systems to make predictions or decisions based on data analysis, helping in tasks like classification, regression, and clustering under uncertainty. Statistical machine learning has played a pivotal role in advancing network intrusion detection systems (NIDS), offering diverse approaches for identifying and mitigating cyber threats. For instance, Barbara et al. [10] utilized data mining algorithms to develop the ADAM project, a real-time anomaly detection system. Another work, done by Tavallaee et al. [11], presented an improved KDD dataset for benchmarking intrusion detection algorithms. Similarly, Kruegel and Vigna [12] explored anomaly detection using sequences of system calls, enhancing the detection accuracy of host-based IDS. Additionally, Thaseen and Kumar [13] integrated SVM classifiers with feature reduction techniques to efficiently handle high-dimensional data in network traffic. Despite its wide adaption, Statistical learning models needs intensive human efforts in feature engineering and can be susceptible to overfitting, particularly when the data has high variance or the model is too complex, leading to poor generalization on new, unseen data.

Compared with Statistical Machine Learning, Deep learning has increasingly become a pivotal approach, providing robust mechanisms to detect sophisticated cyber threats. Deep learning models, primarily due to their ability to learn complex patterns without hard effort in feature engineering from large volumes of data, have shown significant promise in distinguishing between normal traffic and potential threats with high accuracy. Yin et al. [14], demonstrated their effectiveness in capturing spatial features within network traffic. Similarly, Kim et al. [15] employed Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, to analyze temporal features of traffic data for anomaly detection. Further, Javaid et al. [16] explored the use of Self-Taught Learning (STL), a hybrid model that combines deep learning with sparse coding to enhance feature learning in an unsupervised manner. While deep learning offers substantial improvements in network intrusion detection, it also presents several challenges. One major drawback is the requirement for vast amounts of labeled training data, which is expensive and time-consuming to gather in the cybersecurity domain. Additionally, deep learning models are often seen as "black boxes," providing limited interpretability regarding their decision-making processes, which can be a critical shortfall in security applications where understanding the rationale behind decisions is essential.

The growing interest in making machine learning models, especially those applied to network intrusion detection, more interpretable and trustworthy, has spurred the development of explainable artificial intelligence (XAI) approaches in this domain. For instance, Sauka et al. [17] developed an adversarial robust and explainable intrusion detection system using deep learning, emphasizing the enhancement of model transparency and robustness. Patil and colleagues [18] proposed a machine learning-based intrusion detection system that highlights the potential of explainability in security applications, focusing on demystifying the black-box nature of complex models. Keshk et al. [19] introduced an explainable deep learning framework specifically tailored for IoT networks, underscoring the critical need for clarity in automated security systems within such environments. Furthermore, Wang et al. [20] and Barnard et al. [21] have contributed significantly by integrating techniques like SHAP (SHapley Additive exPlanations) to elucidate the decision-making processes of their intrusion detection models, thus facilitating a better understanding and trust among network security personnel.

Explainable machine learning (XAI) models for network intrusion detection often face challenges such as increased computational complexity and potentially reduced performance due to the overhead of generating explanations. Additionally, while providing transparency, the explanations themselves may be too technical or abstract for non-specialist users, limiting their practical usefulness in real-world security applications where clear and actionable insights are required. Such limitations motivate us to use Explainable AI method to perform a comparative analysis on the behaviors of different machine learning-enabled intrusion detectors on the same dataset.

III. Methodology

We use Occlusion Sensitivity to analyze the behavior of different machine learning models on UNSW-NB15 dataset. We want to see if the trained machine learning model we use for IDS could unintentionally become biased towards specific features.

A. Data Preprocessing

The UNSW-NB15 Dataset [9] contains 175,341 entries across 45 distinct columns. We conduct the following data preprocessing steps:

- *Removal of incomplete records:* we remove records containing missing values resulting in a reduced dataset of 81,173 entries. An overview of intrusion category is given in Figure 1, the prevalence of 'Normal' traffic at 48.66%, followed by significant portions of 'Generic' at 24.01%, 'Fuzzers' at 19.94%, and smaller fractions for 'Backdoor', 'Analysis', 'Exploits', 'Reconnaissance', 'DoS', and 'Worms'.
- *Encoding categorical features:* we convert categorical features into one-hot encoding.

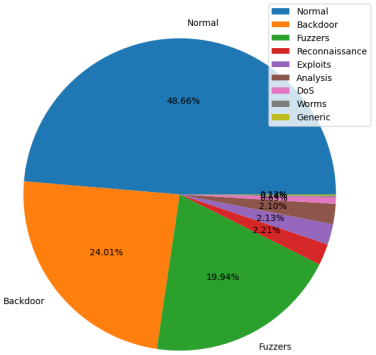

Fig. 1. Distribution of intrusion attack categories after data preprocessing.

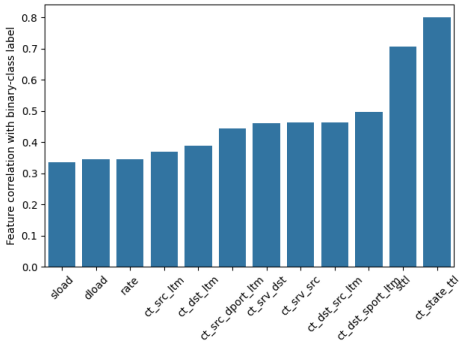

Fig. 3. Selected features for binary classifiers.

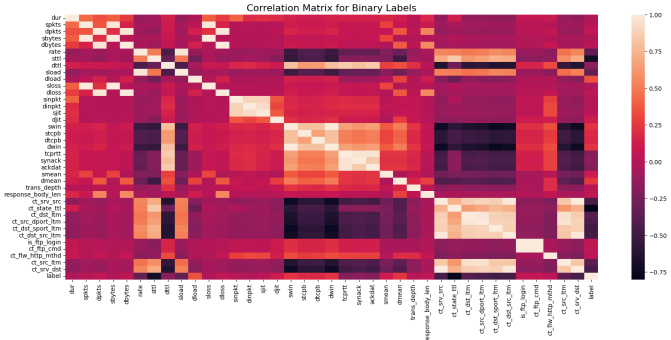

Fig. 2. Feature correlation matrix

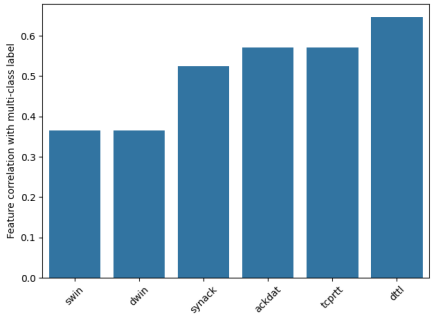

Fig. 4. Selected features for multi-class classifiers.

- *Scaling and Normalization:* We re-scale numerical values to the range of $[0, 1]$.
- *Feature Selection:* we remove features that have less than 0.3 of correlation with the classification label. The correlation matrix of features is given in Figure 2. The selected features for both binary and multi-class classifiers as well as their Correlation Coefficients are in Figures 3 and 4. To evaluate the impact of such feature selection criteria, we also train and analyze the models without feature selection as for comparison.
- *Data Synthetic and Model Training:* We divided the dataset randomly to compose training (80%) and test (20%) set, our stopping criteria for model training are either reach 90% of classification accuracy or improvement less than 1% after the latest epoch.

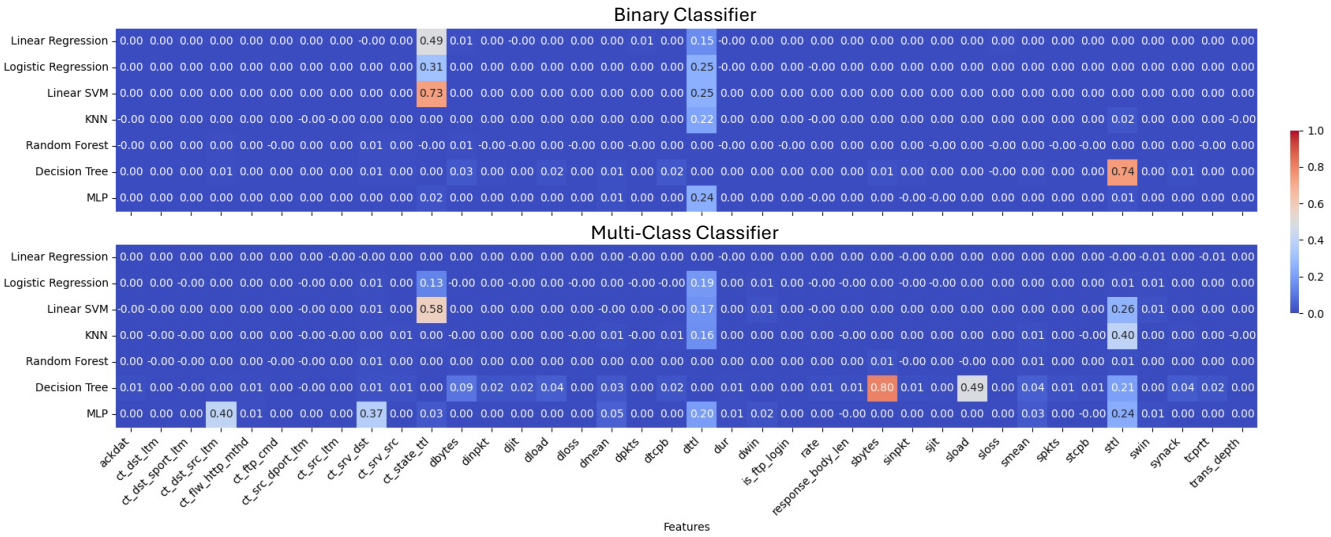

Fig. 5. Feature sensitivity of binary intrusion detection model classifiers trained with complete features.

B. Binary Classifier Analysis

The binary classification task distinguishes between normal network behavior (non-intrusive) and abnormal behavior (intrusive). The models employed for this task are imported directly from scikit-learn library with default configurations, they are: Linear Regression, Logistic Regression, Linear SVM, K-Nearest Neighbor, Random Forest, Decision Tree, and MLP.

C. Multi-Class Classification Analysis

The multi-class intrusion detection models utilize the same suite of models to predict various attack categories such as DoS, Exploits, Fuzzers, and others. Similar performance metrics have been calculated for the multi-class models to evaluate their effectiveness in distinguishing between the different attack categories. Additionally, occlusion sensitivity has been implemented to identify the most influential features for the predictions.

IV. Evaluation & Discussion

A. Binary Classifiers

The feature sensitivity with respect to classification accuracy degradation of binary classifiers are given in Figures 5 and 6. As depicted, most binary models are extremely sensitive to less than three top features, in particular, Decision Tree model is extremely sensitive to even single feature occulusion occlusion. Meanwhile, Multi-Layer Feed-Forward Neural Network and Random Forest models exploit more features than other models. We adjusted the L2 regularization coefficient of MLP model from 0.0001 to 0 and we did not observe significant differences. A possible explanation is that the neural network only leverages a few highly important features and thus develops a sparse internal structure which is not sensitive to the L2 regularization.

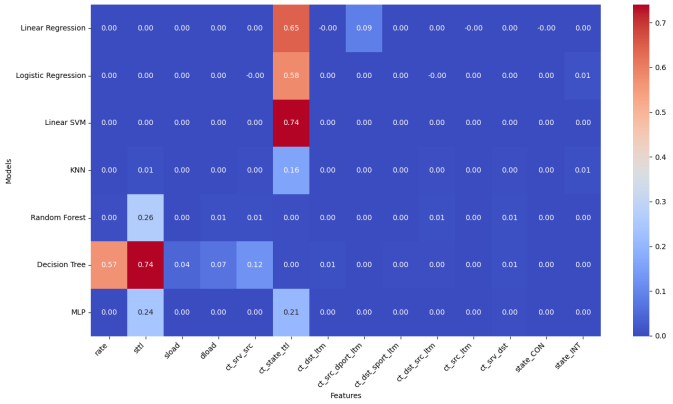

Fig. 6. Feature sensitivity of binary intrusion detection model classifiers trained with selected features.

We masked the Top-2 features of the binary classifiers to analyze the performance degradation of classifiers, depicted in Figure 7. We found that only Random Forest and K-NN classifiers maintain the most insignificant performance degradation.

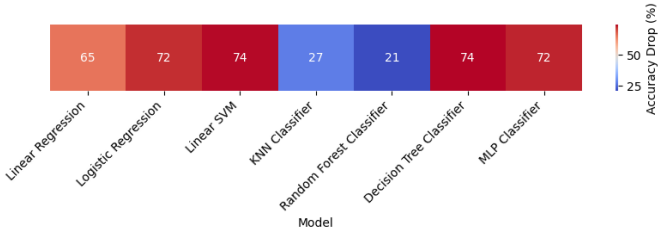

Fig. 7. Accuracy degradation after masking the Top-2 features.

B. Multi-Class Classifiers

The feature sensitivity results of multi-class classifiers also indicates that models utilize more features than in binary classifier when there's no feature selection procedure. Conversely, when feature selection is performed, most classifiers indicates the *ttl*-related features are more critical for classification accuracy. Interestingly, Random Forest is not sensitive to single feature masking in both binary and multi-class scenarios, similar to binary classification scenario, Decision Tree model is extremely sensitive to even single feature occlusion.

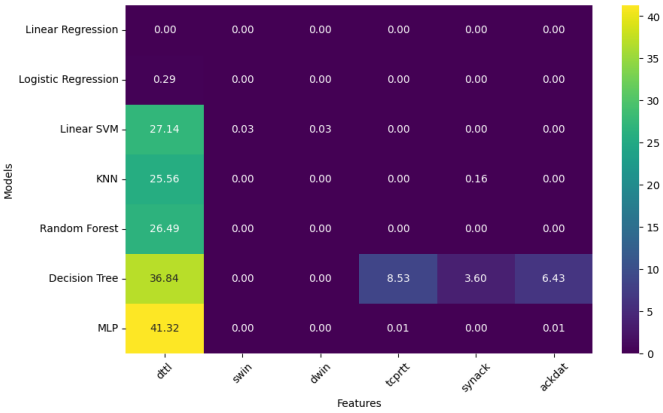

Fig. 8. Feature sensitivity w.r.t. classification accuracy of multi-class intrusion detection model.

Interestingly, if we mask out the top-2 most important features, as depicted in Figure 9, the models performance degradation may not be as significant as in binary classifiers as in Figure 7. We still find that Random Forest still has the best robustness when the top-2 features are mask-out. Moreover, we found that the classifiers rely highly on the time-dependent features, such as *sttl* and *dttl*, simply means that the all these models may face challenges or become useless if they are ported to different application scenarios.

To compare the feature sensitivity of the models, we re-train all the models with the top-3 *TTL*-related features removed and derive the feature sensitivities in Figure 10. Compare with Figure 5, the models utilize more features while the Decision Tree model still has a strong bias towards specific features.

C. Model Overhead Comparison

We compare the time consumption of deriving all the models considering full feature set as in Figure 11. Our experiment

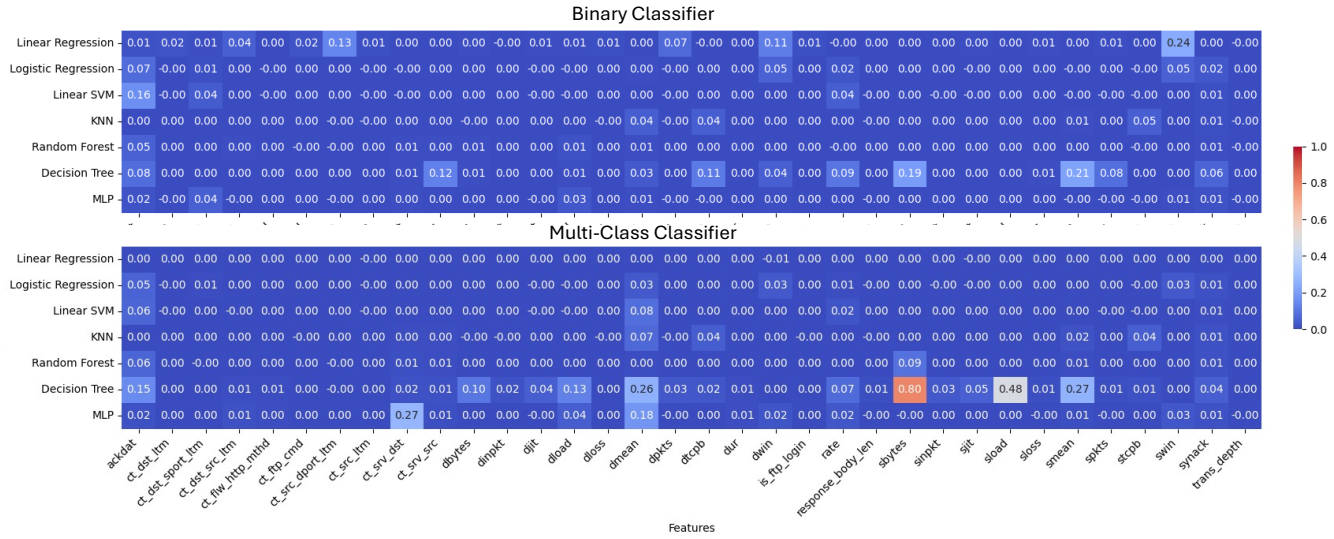

Fig. 10. Feature sensitivity of binary intrusion detection model classifiers trained with complete features.

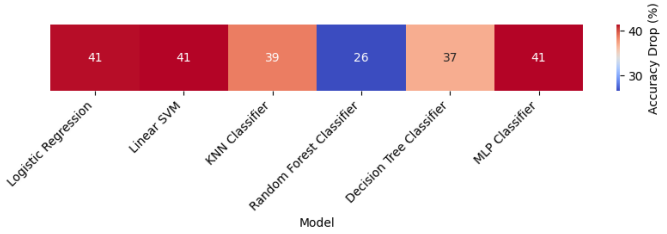

Fig. 9. Accuracy degradation after masking the Top-2 features

is done in standard Google Colab environment with Intel(R) Xeon(R) CPU at 2.20GHz and 12.7GB of RAM, as depicted, Random Forest becomes the best model for the UNSW-NB15 dataset by fully utilizing the features and providing the best efficiency.

V. Conclusion

This paper utilizes Occlusion Sensitivity method for a comparative study on feature importance of various machine learning models for network intrusion detection on UNSW-NB15 dataset. We found that most machine learning models, including Neural Network model exploit few critical features to make decisions and users have to mask out critical features to let model focus on other useful features. In the meantime, Random Forest is the only model that treat all input features equally. Our further experiment also reveal that Random Forest is more veratile than neural network models such as MLP by proving similar performance with better robustness and significantly less training time. Our finding also indicates that explainable AI-guided feature engineering could be a promising approach for deriving robust model while maintain uncompromising performances.

Our future direction includes improving airspace ATC workload assessment by considering metrics beyond delayed and

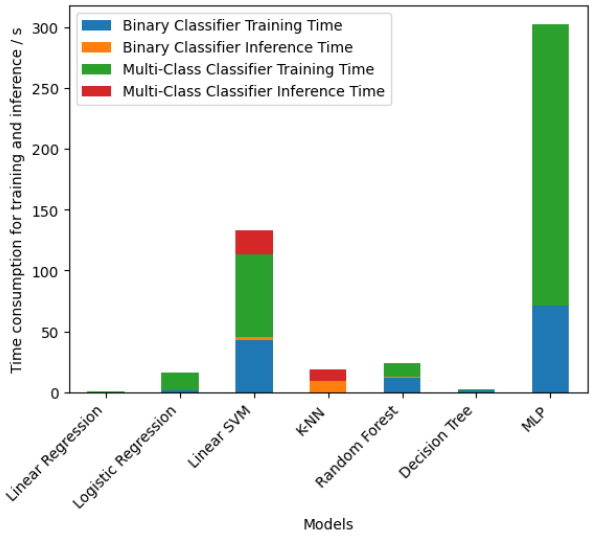

Fig. 11. Comparison of model overhead.

total flights. Additionally, we plan to explore neural networks' potential in generating comprehensive airspace configuration plans.

Acknowledgment

This research was supported by the Center for Advanced Transportation Mobility (CATM), USDOT Grant No. 69A3551747125, 270128BB(AWD00237), the U.S. National Science Foundation under Grant No.2231629, 2142514 and Grant No.2309760 and the USDOT Tier-1 University Transportation Center (UTC) Transportation Cybersecurity Center for Advanced Research and Education (CYBER-CARE) (Grant No. 69A3552348332).

REFERENCES

[1] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[5] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[6] W. Tan, J. Renkhoff, A. Velasquez, Z. Wang, L. Li, J. Wang, S. Niu, F. Yang, Y. Liu, and H. Song, "Noisecam: Explainable ai for the boundary between noise and adversarial attacks," *arXiv preprint arXiv:2303.06151*, 2023.

[7] J. Renkhoff, W. Tan, A. Velasquez, W. Y. Wang, Y. Liu, J. Wang, S. Niu, L. B. Fazlic, G. Dartmann, and H. Song, "Exploring adversarial attacks on neural networks: An explainable approach," in *2022 IEEE International Performance, Computing, and Communications Conference (IPCCC)*. IEEE, 2022, pp. 41–42.

[8] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Springer, 2014, pp. 818–833.

[9] N. Moustafa and J. Slay, "Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 military communications and information systems conference (MilCIS)*. IEEE, 2015, pp. 1–6.

[10] D. Barbará, J. Couto, S. Jajodia, and N. Wu, "Adam: a testbed for exploring the use of data mining in intrusion detection," *ACM Sigmod Record*, vol. 30, no. 4, pp. 15–24, 2001.

[11] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*. Ieee, 2009, pp. 1–6.

[12] C. Kruegel and G. Vigna, "Anomaly detection of web-based attacks," in *Proceedings of the 10th ACM conference on Computer and communications security*, 2003, pp. 251–261.

[13] I. S. Thaseen and C. A. Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class svm," *Journal of King Saud University-Computer and Information Sciences*, vol. 29, no. 4, pp. 462–472, 2017.

[14] C. Yin, Y. Zhu, J. Fei, and X. He, "A deep learning approach for intrusion detection using recurrent neural networks," *Ieee Access*, vol. 5, pp. 21 954–21 961, 2017.

[15] J. Kim, J. Kim, H. L. T. Thu, and H. Kim, "Long short term memory recurrent neural network classifier for intrusion detection," in *2016 international conference on platform technology and service (PlatCon)*. IEEE, 2016, pp. 1–5.

[16] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, 2016, pp. 21–26.

[17] K. Sauka, G.-Y. Shin, D.-W. Kim, and M.-M. Han, "Adversarial robust and explainable network intrusion detection systems based on deep learning," *Applied Sciences*, vol. 12, no. 13, p. 6451, 2022.

[18] S. Patil, V. Varadarajan, S. M. Mazhar, A. Sahibzada, N. Ahmed, O. Sinha, S. Kumar, K. Shaw, and K. Kotecha, "Explainable artificial intelligence for intrusion detection system," *Electronics*, vol. 11, no. 19, p. 3079, 2022.

[19] M. Keshk, N. Koroniotis, N. Pham, N. Moustafa, B. Turnbull, and A. Y. Zomaya, "An explainable deep learning-enabled intrusion detection framework in iot networks," *Information Sciences*, vol. 639, p. 119000, 2023.

[20] M. Wang, K. Zheng, Y. Yang, and X. Wang, "An explainable machine learning framework for intrusion detection systems," *IEEE Access*, vol. 8, pp. 73 127–73 141, 2020.

[21] P. Barnard, N. Marchetti, and L. A. DaSilva, "Robust network intrusion detection through explainable artificial intelligence (xai)," *IEEE Networking Letters*, vol. 4, no. 3, pp. 167–171, 2022.