ANALYSIS OF EXPLAINABILITY OF MACHINE LEARNING ALGORITHMS IN ATTACK DETECTION TASKS IN SOFTWARE-DEFINED NETWORKS

ABSTRACT

The doctoral dissertation presents a current perspective on contemporary challenges related to DDoS attack detection in Software Defined Networking environments, and also focuses on the important issue of the explainability of machine learning models.

The thesis of this doctoral dissertation is: **"Original proposals and implementations of explainable artificial intelligence criteria enable effective explanation of decisions made by machine learning models in the process of detecting distributed denial-of-service attacks in software-defined networks."**

Accordingly, this dissertation aims to demonstrate that the appropriate integration of classification algorithms, feature selection techniques, and explainability methods can not only effectively help detect DDoS attacks but also contribute to a deeper understanding of the crucial elements of the proposed SDN security system.

As part of the research, a series of experiments were conducted using advanced machine learning models, including, XGBoost, LightGBM, NGBoost, Decision Tree, and Random Forest, with an emphasis on their explainability. The key contributions of this research include:

- Development of an original feature selection method, SelectDVC, and the development of an original dataset, DVCDDoS2022.
- Comparative analysis of classification algorithms adapted to the task of DDoS attack detection in software-defined networks.
- Comparative analysis of network features and the impact of their reduction on model predictions.
- Development of original implementations of D, F, S metrics defined by A. Rosenfeld, and original proposals for their modification.
- Development of original implementations of monotonicity, fidelity, infidelity, and incompleteness metrics.
- Development of original Flip-Label attack algorithms and label-sanitization algorithms for sanitizing corrupted labels.

In total, eight original algorithms are presented in this work.

The doctoral dissertation is divided into ten chapters. It begins with an introduction that defines the thesis, aim, and scope of the research. The structure of the work is also presented, and the original solutions resulting from the research are discussed. Chapter 2 is devoted to comparing the architecture and application of traditional computer networks with software-defined networks. It also discusses the threats associated with DoS and DDoS attacks, as well as the role of explainable artificial intelligence and the Data Centric AI concept, which have been gaining importance recently. Finally, the issues of Adversarial Machine Learning and Data Poisoning are presented. Chapter 3 provides a literature review, analyzing previous research on DDoS attack detection using machine learning methods. Particular attention is paid to the application of explainable artificial intelligence methods, i.e., SHAP, in cybersecurity, and the impact of data poisoning attacks on the effectiveness of models. Chapter 4 describes the original research environment, which was built using Mininet, Ryu, and Scikit-learn tools. It also presents the proposed architecture of the DDoS attack detection system and details regarding the data collection process and the creation of the original DVCDDoS2022 dataset. Chapter 5 focuses on a detailed analysis of the DVCDDoS2022 dataset, justifying its creation in connection with the limitations of existing and publicly available datasets. The feature selection methods used, i.e., SelectKBest with the chi-square (chi2) test and the Anova test, the Recursive Feature Elimination (RFE) algorithm, Principal Component Analysis (PCA), and the AutoML tool called FeatureWiz, as well as the proposed original method called SelectDVC, are also presented. Chapter 6 is devoted to the evaluation of DDoS attack detection algorithms, with particular emphasis on the NGBoost algorithm. It compares it with other classification methods, analyzing the impact of feature selection on the performance and explainability of the models. The D metric, used to assess the trade-off between the effectiveness and explainability of models, was adapted and introduced. Chapter 7 focuses on the analysis of the explainability of classification models in the face of DDoS attacks, using the SHAP method and the F metric, examining the influence of individual features on model predictions and the effects of gradually removing features of low importance. Chapter 8 expands on the topic with an analysis of the impact of input data perturbations. Original implementations of monotonicity, fidelity, infidelity, and incompleteness metrics were introduced, which allow for the assessment of the stability of model explanations. Chapter 9 analyzes the resistance of detection models to data poisoning attacks and label manipulations. Original modifications of Flip-Label algorithms and label sanitization methods are presented, and the S metric for assessing the stability of models in the face of this type of threat is proposed. In the conclusion of the dissertation, a summary is made, including answers to the main research questions and the main conclusions formulated in the earlier parts of the work. Directions for further exploration in the undertaken issues are also indicated. The effectiveness of the proposed

original algorithms and concepts, as well as the importance of the analysis of model explainability in the field of cybersecurity of software-defined networks, are emphasized.

Naturally, the presented proposals and studies do not fully exhaust the topic, which remains open for further exploration. The thesis stated at the outset of this dissertation is thoroughly validated by the findings of this research.

This study has demonstrated that:

- The proposed feature selection methods, in combination with classification algorithms, achieve a high level of effectiveness in the identification of DDoS attacks, which confirms their usefulness in practical network applications.
- The implementation of explainable artificial intelligence techniques enabled in-depth analysis and interpretation of model decision-making processes, which contributes to increasing confidence in the solutions used.
- The models exhibited varying resistance to attacks and data perturbations, highlighting the need for further research to enhance their stability and reliability in real-world conditions.
- The existence of a trade-off between performance, the number of features, and the explainability of models was demonstrated, which indicates the validity of using a contextual approach and priorities in the process of feature selection and algorithm optimization.

In the future, it is planned to extend the research to include the implementation of a detection model in a real environment of an academic software-defined network. The next stage will be the development and implementation of attack mitigation and prevention mechanisms, integrated with the SDN controller, as well as further continuation of research related to the explainability of models. Additionally, future work will explore the evaluation of the detection environment across various datasets, alternative algorithms (e.g., transformer networks), their hybrid combinations, and real-world deployment scenarios.

14.03.2025