Interpretable ensemble deep-learning for intrusion detection: Enhancing detection performance and explainability

The authors propose a two-phase ensemble Deep Learning new method, called EED (Explainable Ensemble Deep learning) to address the need for accurate and explainable intrusion detection in networks. The method consists of two phases. The proposed method is of high interest and lies very important perspectives in the field of Deep Learning and its applications.

In the first phase, the authors propose an ensemble intrusion detection model using three Long Short-Term Memory (LSTM) models. The accuracy of attack identification is improved by aggregating the outputs of these deep learning classifiers with a meta-learner algorithm.

In the second phase, the authors focus on improving the interpretation and explanation of the detections tracked in Phase I. Based on the SHape Additive exPplanations (SHAP) capabilities, the authors highlighted the factors contributing to the identification and classification of attacks. These explanations provide a better understanding of detected attacks and allow to assist experts in developing effective response strategies to enhance network security.

Empirical experiments are conducted on the NSL-KDD dataset to demonstrate the effectiveness of the proposed method in terms of accuracy and explainability.

While the paper makes significant contributions, some limitations need to be addressed to enhance the quality of the paper:

1- Regarding the choice of Categorical Cross-Entropy (CCE) for the LSTM models, the authors should provide a clear justification for this

selection and provide any empirical evidence or prior work supporting its effectiveness in this context.

2- The paper should provide a justification for choosing SHAP (SHapley Additive exPLanations) over other explainable methods. The authors can discuss the specific advantages of SHAP in the context of intrusion detection. The authors must highlight why SHAP is the most suitable choice for the proposed method

3- The description of existing explainable intrusion detection methods in Related Works section is insufficient. The authors must provide a critical analysis of the existing works listed in Table 1, highlighting their strengths and weaknesses, and explaining how their proposed method complements or improves upon these approaches. This allows for the reader to better understand the utility of explanations in intrusion detection systems.

4- The description of SHAP (SHapley Additive exPLanations) should be enhanced by including a brief explanation of SHAP plots or visualizations. Including such visualizations will strengthen the paper's explanation of the SHAP method.

5- The formula description for the LSTM model can be improved to enhance clarity and understanding. (Formula 1 to 5). The authors should provide a more detailed explanation of the different components and variables in the equations, ensuring that readers can follow the mathematical formulation without ambiguity.

6- Consistency in notation should be maintained throughout the paper. The authors should use the same notation for intervals, whether it is [h_t−1, x_t−1] or [0..1]. This will avoid confusion and enhance the overall readability of the paper.

7- In the conclusion section, the authors may provide further details on potential future works

8- Figure 2 can be improved by numbering the arrows to indicate the sequential steps. This numbering will provide a clear visual flow and help readers understand the sequential nature of the process.