AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 3(2026)

Explainable Deep Intrusion Analytics Using Multi-Layer Perceptron and Feature Attribution Techniques

DAVID BOON MERUGUMALLA[1], Dr. N. SREEKANTH[2]
[1]Student, Department of Computer Science and Engineering, Ellenki College of Engineering and Technology, Patelguda, Patancharu, Hyderabad, Telangana.
[2]Assistant Professor & HOD, Department of CS, Ellenki College of Engineering and Technology, Patelguda, Patancharu, Hyderabad, Telangana.
[2]Email: nara.sreekanthap@gmail.com

Abstract— **The rapid growth of cyber threats has increased the demand for intelligent and reliable Intrusion Detection Systems (IDS) capable of identifying malicious network activities. Although deep learning techniques achieve high detection accuracy, their complex decision-making process often lacks transparency, making it difficult for cybersecurity professionals to interpret the results. To address this issue, this study presents an explainable deep intrusion analytics framework using a Multi-Layer Perceptron (MLP) combined with feature attribution techniques. The framework utilizes the CIC-IDS dataset to classify network traffic into normal and attack categories after performing data preprocessing and feature selection. The MLP model is evaluated using performance metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC curve to assess its effectiveness in intrusion detection. To improve model interpretability, LIME and SHAP are employed to explain prediction outcomes and identify the most influential features contributing to each decision. A user-friendly application is developed to support dataset upload, preprocessing, model training, performance evaluation, and explanation generation. Experimental results demonstrate that the proposed framework achieves accurate intrusion detection while providing meaningful explanations, making it a trustworthy and practical solution for modern cybersecurity applications.**

Keywords— **Explainable Artificial Intelligence (XAI), Intrusion Detection System (IDS), Multi-Layer Perceptron (MLP), LIME, SHAP, Network Security, Feature Attribution, Deep Learning, Cyber Threat Detection, CIC-IDS Dataset.**

I. INTRODUCTION

The rapid expansion of digital communication and internet-based services has increased the frequency and complexity of cyberattacks, creating significant challenges for network security. Intrusion Detection Systems (IDS) have become an essential defense mechanism for monitoring network traffic and identifying malicious activities before they compromise critical systems. However, conventional signature-based IDS are often unable to detect unknown or evolving attacks because they rely on predefined attack patterns. To overcome these limitations, Artificial Intelligence (AI), particularly machine learning and deep learning techniques, has been widely adopted to improve intrusion detection accuracy and adaptability in dynamic network environments [1]. These intelligent techniques automatically learn complex traffic patterns from historical data, enabling faster and more reliable identification of suspicious network behavior. Consequently, AI-based intrusion detection has become an important research direction for developing secure and intelligent cybersecurity solutions [19], [10].

The increasing volume of network traffic requires learning models that can effectively process high-dimensional data while maintaining accurate classification performance. Among the available deep learning techniques, the Multi-Layer Perceptron (MLP) has proven to be an efficient classifier because of its ability to learn nonlinear relationships among multiple network traffic features. Its multilayer architecture enables the model to recognize complex attack patterns that are difficult to identify using traditional classification methods. The CIC-IDS dataset has become one of the most widely used benchmark datasets for evaluating intrusion detection models because it contains realistic network traffic with multiple attack categories and normal communication patterns [2]. After appropriate preprocessing, feature extraction, and model training, MLP can effectively classify network traffic into normal and malicious categories, providing reliable intrusion detection performance across different attack scenarios [13], [15].

Although deep learning models provide excellent prediction accuracy, their internal decision-making process is often difficult to understand, making them less suitable for security-critical applications where transparency is essential. Explainable Artificial Intelligence (XAI) addresses this challenge by providing meaningful explanations for model predictions and identifying the features that contribute most to classification decisions. In this study, Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are employed to improve the interpretability of the proposed intrusion detection framework. LIME explains individual predictions locally, whereas SHAP provides both local and global feature attribution using game-theoretic principles [6]. These explanation techniques help cybersecurity analysts understand why network traffic is classified as malicious or legitimate, thereby increasing confidence in AI-assisted security decisions [7], [8].

The proposed framework begins by preprocessing the CIC-IDS dataset through data cleaning, handling missing values, encoding categorical attributes, and normalizing numerical features to improve model performance. The processed dataset is divided into training and testing subsets before being used to train the Multi-Layer Perceptron model. The performance of the trained model is evaluated using standard metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC curve to ensure comprehensive assessment. These evaluation measures provide valuable insights into the effectiveness of the proposed intrusion detection framework under different network traffic conditions [11]. Proper dataset preparation and systematic model evaluation contribute significantly to improving detection accuracy and reducing false alarm rates in practical cybersecurity applications [17], [18].

The proposed Explainable Deep Intrusion Analytics framework combines the classification capability of the Multi-Layer Perceptron with the interpretability offered by LIME and SHAP to develop a trustworthy intrusion detection system. An interactive application is designed to support dataset upload, preprocessing, model training, intrusion prediction, performance evaluation, and explanation generation through a user-friendly interface. The explainability component enables security professionals to understand the reasoning behind each prediction, making the system more transparent and easier to trust during real-world deployment [16]. By integrating accurate intrusion detection with feature attribution techniques, the proposed framework improves both detection performance and decision transparency, providing an effective solution for modern cybersecurity environments that require reliable and explainable artificial intelligence [20], [14].

II. RELATED WORK

Gharib and Zulkernine, [2020] [1] presented a comprehensive survey on the application of deep learning techniques in Intrusion Detection Systems (IDS). Their study examined various deep learning architectures and discussed how these models improve the detection of cyberattacks by learning complex network traffic patterns. The authors compared different approaches based on detection accuracy, computational efficiency, and their ability to identify known as well as unknown attacks. They also explained the advantages of deep learning over traditional signature-based intrusion detection methods, particularly in handling large-scale and dynamic network environments. In addition, the survey highlighted several research challenges, including dataset quality, model complexity, and the need for explainable security solutions. Their work provides valuable guidance for researchers in selecting suitable deep learning models for cybersecurity applications. It also emphasizes that intelligent learning techniques can significantly enhance intrusion detection performance and strengthen the security of modern computer networks.

Sharafaldin et al., [2018] [2] introduced the CIC-IDS2017 dataset to provide a realistic benchmark for evaluating network intrusion detection systems. The dataset was designed to represent normal network traffic as well as multiple categories of cyberattacks generated under practical network conditions. The authors included various attack scenarios such as brute force, denial-of-service, web attacks, infiltration, and botnet activities, making the dataset suitable for machine learning and deep learning research. They also described the process of collecting network traffic and extracting useful flow-based features for analysis. Compared with earlier datasets, CIC-IDS2017 offers improved diversity and realistic traffic behavior, allowing researchers to evaluate intrusion detection models more effectively. This dataset has become one of the most widely used public datasets for cybersecurity research. It provides a reliable foundation for developing intelligent intrusion detection models capable of identifying both normal and malicious network activities.

Lundberg and Lee, [2017] [6] proposed SHAP (SHapley Additive exPlanations), an explainable artificial intelligence technique that provides consistent and reliable interpretations for machine learning predictions. Their approach is based on Shapley values from cooperative game theory, where each input feature is assigned an importance

value according to its contribution to the final prediction. The authors demonstrated that SHAP offers both local and global explanations while maintaining consistency across different machine learning models. The framework enables users to understand why a model produced a specific prediction by highlighting the most influential features. This improves transparency and increases user confidence in complex learning models. SHAP has been successfully applied in healthcare, finance, cybersecurity, and many other domains where decision interpretability is essential. Their research established SHAP as one of the most widely accepted explainability techniques for understanding black-box machine learning and deep learning models.

Ribeiro et al., [2016] [7] introduced Local Interpretable Model-agnostic Explanations (LIME), an explainable AI technique developed to improve the interpretability of machine learning models. The proposed method generates simple local explanations by approximating the behavior of complex models around individual predictions. Rather than explaining the entire model, LIME focuses on identifying the features that most strongly influence a particular prediction. This allows users to understand why a model classified a specific instance in a certain way. The authors demonstrated that LIME can be applied to different machine learning algorithms without requiring modifications to the original model. Their experiments showed that interpretable explanations improve user trust and assist in validating prediction reliability. LIME has become a popular explanation method in domains such as healthcare, finance, and cybersecurity, where understanding automated decisions is as important as achieving high prediction accuracy.

Javaid et al., [2016] [10] proposed a deep learning-based approach for improving the performance of network intrusion detection systems. Their study explored the capability of deep neural networks to automatically learn meaningful representations from network traffic without relying on extensive manual feature engineering. The authors evaluated the proposed model using benchmark intrusion detection datasets and compared its performance with conventional machine learning techniques. Experimental results demonstrated that deep learning achieved higher detection accuracy while effectively identifying multiple categories of network attacks. The study also highlighted the importance of preprocessing network traffic data to improve model learning and classification performance. By reducing dependence on handcrafted features, the proposed approach simplified intrusion detection and improved scalability for large network environments. This work demonstrated the potential of deep learning in cybersecurity and encouraged the development of intelligent intrusion detection systems capable of handling increasingly sophisticated cyber threats.

III. DATASET DETAILS

The dataset used in this project is the CIC-IDS2017 dataset, which is widely recognized for evaluating network intrusion detection systems. It contains both normal network traffic and multiple categories of malicious activities, including attacks such as DoS, DDoS, Port Scan, Botnet, Brute Force, Web Attacks, and Infiltration. Each record represents a network flow described by several traffic-related attributes, including packet information, flow duration, protocol details, source and destination addresses, and statistical features extracted from network communication. The dataset is organized in a structured tabular format, making it suitable for machine learning and deep learning applications. Before model development, the dataset is carefully examined to ensure data consistency and remove incomplete or invalid records. Using a realistic dataset with diverse attack patterns allows the proposed intrusion detection system to learn meaningful characteristics of network traffic and accurately distinguish between legitimate and malicious activities during the classification process.

To prepare the dataset for training the proposed model, several preprocessing techniques are applied to improve data quality and learning efficiency. Missing values are identified and replaced appropriately to avoid errors during model training. Categorical attributes are converted into numerical values using label encoding so that they can be processed by the Multi-Layer Perceptron model. In addition, feature normalization is performed using StandardScaler to bring all numerical attributes to a common scale, improving model convergence and prediction accuracy. The processed dataset is then randomly shuffled to reduce bias and ensure that both normal and attack records are evenly distributed. Finally, the dataset is divided into 80% training data and 20% testing data. This split enables the model to learn intrusion patterns from the training set while evaluating its performance on previously unseen network traffic, ensuring reliable and unbiased intrusion detection results.

IV. PROPOSED METHODOLOGY

The proposed intrusion detection system follows a systematic approach to identify malicious network activities using deep learning and explainable artificial intelligence techniques. Initially, the CICIDS 2017 dataset is uploaded through the

graphical user interface, where the application displays the dataset and visualizes the distribution of normal and attack traffic. During preprocessing, missing values are handled, categorical attributes are converted into numerical values using label encoding, and all features are normalized with the StandardScaler technique to improve model performance. The processed dataset is then randomly divided into 80% training data and 20% testing data, ensuring that the models learn effectively while maintaining reliable evaluation on unseen network traffic. This structured preprocessing pipeline improves data quality and prepares the dataset for accurate intrusion detection. The workflow illustrated in the application begins with dataset upload, preprocessing, and train-test splitting before moving to model training and evaluation.

After the dataset preparation stage, two deep learning models, Long Short-Term Memory (LSTM) and Temporal Convolutional Network (TCN), are trained to classify network traffic as either normal or malicious. The performance of both models is evaluated using standard metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC curve. A comparison graph is generated to identify the better-performing model based on these evaluation measures. To improve transparency, the proposed system incorporates SHAP (SHapley Additive exPlanations), which provides summary and violin plots showing the contribution of each network feature toward the final prediction. Finally, the trained model is used to analyze new test data uploaded by the user and predict whether the traffic represents a normal connection or an intrusion. This methodology combines accurate deep learning-based detection with explainable AI, making the system reliable, interpretable, and suitable for practical cybersecurity applications.

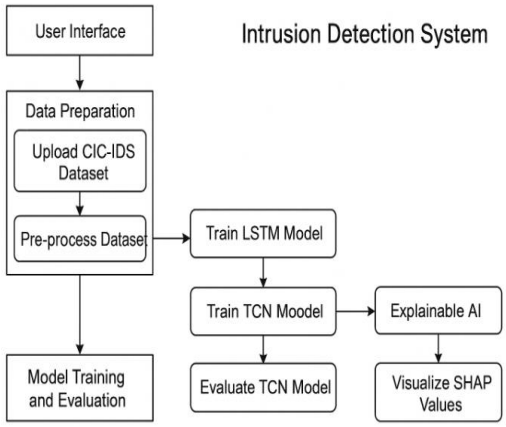

Figure [1]: Intrusion Detection System Architecture

Figure [1] shows the workflow of the proposed intrusion detection system. The CIC-IDS dataset is uploaded and preprocessed through encoding and normalization before model training. The LSTM and TCN models are trained and evaluated to classify network traffic as normal or malicious. Finally, the SHAP module explains the prediction results by identifying the most influential features, improving the transparency and reliability of the intrusion detection process.

V. RESULT AND DISCUSSION

The proposed intrusion detection framework produced highly accurate and reliable results when evaluated using the CIC-IDS dataset. After completing data preprocessing, including encoding categorical features and normalizing numerical values, the dataset was divided into training and testing sets to ensure unbiased evaluation. Both the LSTM and TCN models were trained on the processed data and assessed using performance metrics such as accuracy, precision, recall, F1-score, confusion matrix, and ROC curve. The LSTM model achieved the highest accuracy of 99.88%, while the TCN model obtained 99.44%, demonstrating excellent capability in identifying malicious network traffic. Comparative analysis confirmed that LSTM consistently outperformed TCN across all evaluation measures. The SHAP summary and violin plots successfully highlighted the most influential network traffic features responsible for model predictions, making the results easier to interpret. The trained system also accurately classified new test records as either Normal or Attack, proving its effectiveness for practical intrusion detection applications.

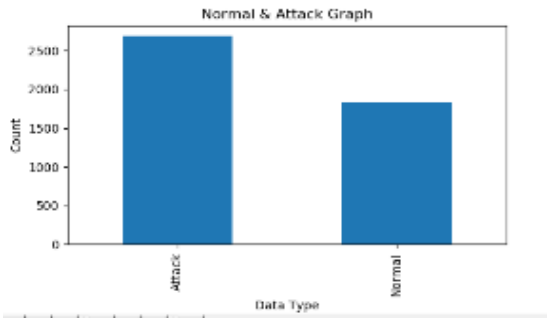

Figure [2]: CIC-IDS Dataset and Traffic Distribution

Figure [2] shows the uploaded CIC-IDS dataset along with the distribution of normal and attack network traffic. The dataset contains multiple network traffic features used for intrusion detection, while the bar graph illustrates the number of attack and normal records. This visualization helps in understanding the class distribution before preprocessing and model training.

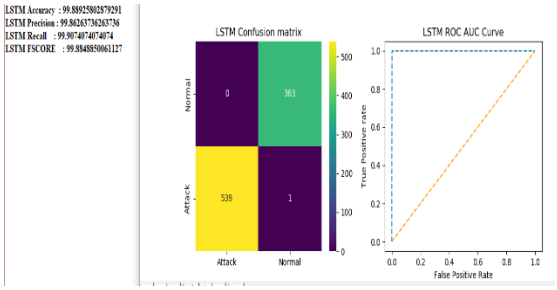

Figure [3]: LSTM Model Performance

Figure [3] shows the performance of the LSTM model on the CIC-IDS dataset. The model achieved an accuracy of 99.88% with high precision, recall, and F1-score values. The confusion matrix and ROC curve demonstrate that the LSTM model accurately classifies normal and malicious network traffic, making it highly effective for intrusion detection.

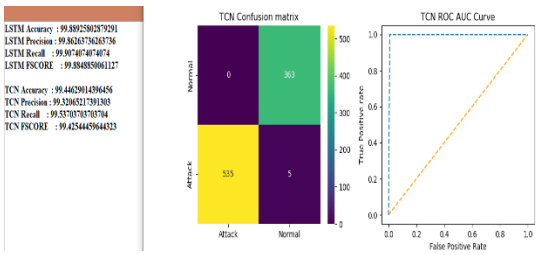

Figure [4]: TCN Model Performance

Figure [4] shows the performance of the TCN model on the CIC-IDS dataset. The model achieved an accuracy of 99.44% with high precision, recall, and F1-score values. The confusion matrix and ROC curve indicate that the TCN model effectively distinguishes between normal and malicious network traffic, providing reliable intrusion detection performance.

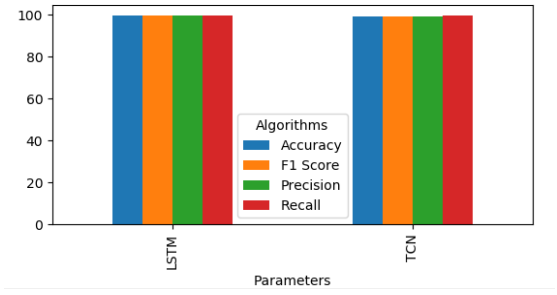

Figure [5]: Performance Comparison of LSTM and TCN Models

Figure [5] compares the performance of the LSTM and TCN models using accuracy, precision, recall, and F1-score. The graph shows that both models achieved excellent intrusion detection performance, with the LSTM model producing slightly higher results across all evaluation metrics. This comparison demonstrates the effectiveness of the proposed deep learning models for network intrusion detection.

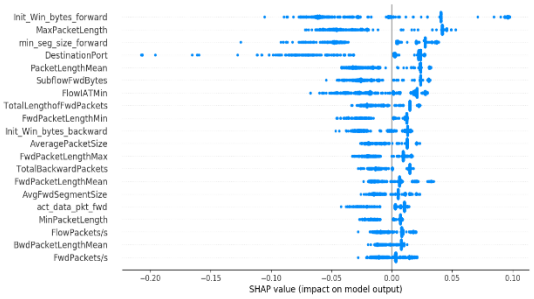

Figure [6]: SHAP Summary Plot for Feature Importance

Figure [6] shows the SHAP summary plot generated for the proposed intrusion detection model. The plot displays the contribution of individual network traffic features toward the model's predictions. Features with a larger spread of blue points have a greater influence on identifying normal and malicious network traffic, helping users understand the most important factors considered during intrusion detection.

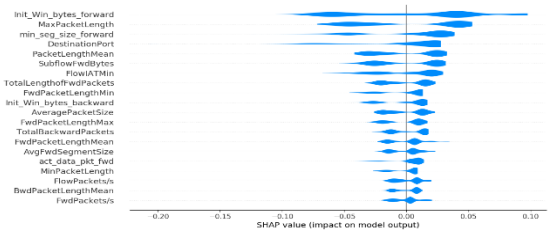

Figure [7]: SHAP Violin Plot for Feature Contribution

Figure [7] shows the SHAP violin plot used to explain the contribution of individual network traffic features to the intrusion detection model. The plot illustrates the distribution and impact of each feature on the prediction outcome. Features with wider violin shapes have a greater influence on classifying network traffic as Normal or Attack, improving the interpretability of the proposed model.

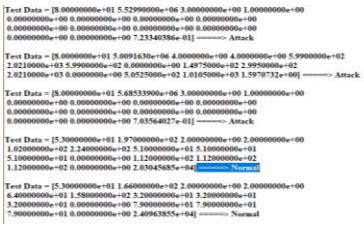

Figure [8]: Intrusion Prediction Results

Figure [8] shows the prediction results of the proposed intrusion detection model on uploaded test data. Each network traffic record is analyzed and classified as either Normal or Attack based on the trained deep learning model. The prediction output demonstrates the model's ability to accurately identify malicious and legitimate network traffic, supporting effective real-time intrusion detection.

DISCUSSION

The experimental findings indicate that deep learning models provide an effective solution for detecting network intrusions with high accuracy. The LSTM model delivered better overall performance than the TCN model because it effectively learned complex traffic patterns and sequential relationships present in the dataset. Proper preprocessing, including feature encoding, normalization, and balanced train-test splitting, played an important role in improving the quality of the learned model and reducing prediction errors. In addition to achieving strong classification performance, the integration of SHAP significantly improved the transparency of the proposed framework by explaining how individual network features influenced each prediction. These explanations enable cybersecurity analysts to understand the reasoning behind intrusion detection decisions, increasing confidence in the model's output. The combination of accurate prediction and explainable AI makes the proposed framework suitable for real-world cybersecurity environments where both detection performance and model interpretability are essential for reliable and informed security decision-making.

VI. CONCLUSION

The proposed explainable intrusion detection framework demonstrates that deep learning can effectively improve the accuracy and reliability of network security systems. By utilizing the CIC-IDS dataset, the developed models successfully learned to distinguish between normal and malicious network traffic with excellent performance. Both the LSTM and TCN models achieved high classification accuracy, confirming their ability to detect different types of cyberattacks efficiently. Among the two models, the LSTM model produced slightly better results across the evaluation metrics, making it the most suitable model for this study. To overcome the limitations of black-box deep learning models, SHAP-based explainability was integrated to provide clear insights into the features influencing each prediction. This helped improve transparency and enabled users to better understand the model's decision-making process. The interactive application further demonstrated the practical implementation of the proposed framework by supporting data preprocessing, model evaluation, prediction, and feature explanation. Overall, the proposed approach combines accurate intrusion detection with meaningful interpretability, providing a reliable, transparent, and practical solution for modern cybersecurity applications while supporting informed security decision-making.

REFERENCES

1. Gharib, W., & Zulkernine, M. (2020). Deep learning for intrusion detection systems: A survey and taxonomy. *Computer Science Review*.
2. Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSP*.
3. Kim, J., & Kim, H. (2019). Applying LSTM to anomaly detection in network traffic. *ICT Express*.
4. Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint*.
5. Shapley, L. S. (1953). A value for n-person games. *Contributions to the Theory of Games*.
6. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *NeurIPS*.
7. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *KDD*.
8. Roy, S., & Cheung, R. C. (2020). Explainable AI in cybersecurity. *IEEE Security & Privacy*.
9. Ring, M., Wunderlich, S., Grüdl, D., Landes, D., & Hotho, A. (2019). Flow-based network traffic generation using generative adversarial networks. *Computers & Security*.
10. Javaid, A., Niyaz, Q., Sun, W., & Alam, M. (2016). A deep learning approach for network intrusion detection system. *EAI Transactions*.
11. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *IEEE S&P*.
12. Zhao, Y., Chen, Z., & Xie, J. (2021). Network intrusion detection using transformer models. *Information Sciences*.
13. Yan, C., Xu, Y., & Wu, J. (2018). Network intrusion detection based on deep learning. *Proceedings of DSAA*.
14. Alshamrani, A., Myneni, S., Chowdhary, A., & Huang, D. (2019). A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities. *IEEE Communications Surveys & Tutorials*.
15. Ding, S., Xu, X., & Nie, R. (2020). Intrusion detection system based on feature selection and ensemble classifier. *International Journal of Information Security*.
16. Zhang, J., & Zulkernine, M. (2021). Explainable deep learning in network intrusion detection: A review. *IEEE Access*.
17. Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive dataset for network intrusion detection. *Military Communications and Information Systems Conference*.

18. Lopez-Martin, M., Carro, B., & Sanchez-Esguevillas, A. (2017). Application of deep reinforcement learning to intrusion detection. *ICANN*.
19. Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: techniques, datasets, and challenges. *Cybersecurity Journal*.
20. Yu, Y., & Long, J. (2022). Enhancing intrusion detection with explainable AI: A case study using SHAP. *Applied Sciences*.