Explainable AI for Intrusion Detection Systems: A Model Development and Experts' Evaluation

Henry Durojaye and Mohammad Naiseh

Bournemouth University, Talbot, United Kingdom
s5526113@bournemouth.ac.uk, mnaiseh1@bournemouth.ac.uk

Abstract. This study sought to develop a transparent machine learning model for network intrusion detection that domain experts would trust for security decision-making. Intrusion detection systems using machine learning have shown promise but often lack interpretability, undermining user trust and deployment. A hybrid Random Forest/XGBoost classifier achieved over 99% accuracy and F1 score, outperforming previous literature. Post-hoc LIME explanations provided feature effect transparency. Nine domain experts from technical roles then evaluated the model's reliability, explainability, and trustworthiness through a standardised process. While over half found the model reliable, one-third expressed uncertainty. Responses on performance explanations and trustworthiness assessments also varied thus suggesting opportunities to strengthen reliability communications and consolidate diverse perspectives. To optimise user confidence and model deployment, refinements targeting consistent explainability across audiences were proposed. Overall, high predictive performance validated effectiveness, but variable viewpoints from evaluations indicated the need to bolster reliability and trust explanations. With continued iterative evaluation and enhancements, this research framework holds promise for developing interpretable machine learning solutions trusted for complex security decision-making.

Keywords: Explainable AI, Trustworthy AI, Intrusion detection systems

1 Introduction

The field of machine learning (ML) has advanced rapidly in recent years, driven by improvements in algorithms, computational power, and data availability [1][2]. The increase in data availability and the increased use of Artificial Intelligence (AI) and ML in decision making makes it extremely important to have technology which the people making the decision can trust and understand. Thus, the need for accurate and reliable decision-making tools is critical for human experts to make informed decisions in complex and data-intensive domains such as healthcare, finance, transportation and cybersecurity [3]. Cybersecurity is one domain where the use of ML to detect, identify, classify and predict attacks has become mainstream. The challenge with the blind adoption of ML in cybersecurity is that often, the users have no idea how the ML model arrives at its decisions and sometimes some errors have

been identified through the use of domain knowledge by experts. Papers by [4] and [5] opine that the accuracy and reliability of machine learning tools can be compromised if they are not transparent or explainable, or if they are based on biased or incomplete data. While ML has the potential to augment cyber defence capabilities, there are legitimate concerns regarding the reliability, explainability, and trustworthiness of these emerging ML-based security tools. Some of the major issues that have been identified by industry and experts are reliability and explainability [6]. This is because ML models can make mistakes and false predictions that may go unnoticed without proper validation and testing which could lead to incorrect detections or missed attacks, and ML models are often viewed as "black boxes" that do not provide explanations for their predictions, making it difficult for analysts to interpret results and troubleshoot errors and trustworthiness [7]. Without adequate explanations and reliability measures, security analysts may be hesitant to fully trust the judgments and recommendations of ML tools. The lack of a trustworthy ML tool creates severe consequences, including inaccurate decision-making, increased risk, and loss of trust in the decision-making process.

These concerns highlight the need for techniques to improve key aspects of ML models for security applications, including reliability, explainability, interpretability, and overall trustworthiness. With proper design and development practices focused on these areas, ML-based tools show promise to significantly enhance cyber defence capabilities. To effectively utilise ML to augment human cyber analysts, it is important to develop ML-based tools that analysts can trust and understand [7]. This requires techniques for improving the explainability, interpretability, and reliability of ML models used in security applications [8]. Even with these improvements, ML models will likely remain imperfect and make errors that human analysts need to correct [9][5]. Therefore, there is a need for ML-based security tools that can support, rather than replace, human experts through capabilities such as alert prioritization, evidence presentation, and model summarization [10]. Such ML-assisted tools have the potential to improve the efficiency and effectiveness of human analysts while leveraging the complementary strengths of humans and machines [11]. However, in their paper researchers [12] argue that creating ML-based tools that analysts find trustworthy and useful remains a challenge due to the complexity of ML models and nuances of human decision making. In fact, the complex machine learning models currently being developed are not able to provide a well-explained and interpretable prediction for decision efficiency [13][14][15].

The aim of this study is to propose an explainable and trustworthy ML model that can effectively detect and predict cyberattacks. The study proposes a tool which incorporates techniques for enhancing the reliability, explainability, and interpretability of the underlying ML models. The study will analyse feature importance explanations with the aim of assisting analysts and practitioners in understanding how the model weighs the features and which ones are deemed more important than the others. This will be done using a hybrid ML model using random forest (RF) and extreme gradient boosting (XGBoost) with LIME, a model-agnostic method to analyse the importance of features that have a direct impact on the prediction of cyberattack detection [16]. The study will evaluate how these techniques impact analysts' trust in the tool, efficiency in detecting attacks, and ability to identify

tool errors. Insights from the evaluation will provide recommendations for designing ML-based security tools that facilitate, rather than hinder, human expert decision making.

The paper is organized as follows: Section 2 presents the related works on the current approaches of suicide attempt prediction models. Section 3 discusses the proposed explainable predictive model and Section 4 presents the results and discussions of the proposed model, and finally, Section 5 summarizes the conclusion with future developments.

2 Related Works

This section provides related works around eXplainable AI and the use of XAI in the Cybersecurity domain.

2.1 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) aims to ensure AI, and in this case, ML models are interpretable and understandable by humans. Whilst transparent models like linear regression are inherently understandable, there are opaque ML models, particularly those based on Deep Learning which require additional explanation techniques. According to [13], to understand such an opaque model there is need to introduce post-hoc explainability methods whose aim is to explore how an already developed model makes predictions.

Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) are the popular agnostic techniques that are used in interpreting ML models [17][18]. The interoperability of LIME with various ML algorithms was a major reason why researchers chose LIME over SHAP in this study [19]. LIME can explain any classifier because it focuses on perturbing input features to determine importance, making it model independent. This adaptability is useful for studies that explore novel complicated models [20]. Furthermore, LIME's local explanations are clearer and faster to create, focusing on specific projections. Simpler local insights may be preferable to SHAP's globally consistent feature importance ratings for applications evaluating findings for domain experts or end users. Because LIME is widely available in public libraries, its convenience of use may trump SHAP's theoretical advantages for certain explanation-focused investigations.

2.2 XAI in Cybersecurity

Prior research has largely concentrated on optimising the performance of various classification algorithms for trust-based intrusion detection systems (IDS), without sufficiently illuminating the rationale behind the sophisticated models' behaviour [18][21]. The study by [21] aimed to address the limitation caused by the "black box" nature of ML models through the application of the eXplainable Artificial Intelligence (XAI) approach to improve trust management through investigating the functioning of Decision Tree (DT) classification in intrusion detection (IDS), in hopes of expounding the reasons behind the model's outputs. The study achieved its objectives by using a model dashboard to present thorough discussions and explanations of their proposed model using SHAP to explain the output of the ML model. By employing

these explainable AI methods, the authors aimed to enhance trust management by enabling human experts to understand the underlying data and the impact of malicious data in detecting intrusions.

In their study [18] focus on the importance of XAI in enhancing trust management in IDS also using a Decision Tree model. The authors address the need for human experts to understand the underlying data evidence and causal reasoning in ML models. They use feature engineering and a rule-based model to explore the decision tree algorithm in IDS and analyse the importance of features based on the entropy measure for intrusion detection. The study also interprets the rules extracted from the DT approach for intrusion classification and compares the accuracy of the decision tree with state-of-the-art methods.

In their study [17] propose that actionable explanations of intrusion detection system alerts are needed to assist cybersecurity analysts in making well-informed decisions amidst a deluge of alerts, including significant numbers of false positives. They present a framework named FAIXID which leverages XAI and data cleaning techniques to enhance the explainability and understandability of intrusion detection alerts. FAIXID adopts a multi-faceted approach through five functional modules: a pre-modelling explainability module that improves data quality using data cleaning; a modelling module that provides model explanations to illuminate model internals; a post-modelling explainability module that supplements these with additional explanations; an attribution module that selects context-appropriate explanations according to analysts' needs; and an evaluation module that assesses explanations and elicits analyst feedback.

The researchers implement and evaluate FAIXID using real-world datasets. Results show that data cleaning and explainability techniques incorporated into the framework provide context-specific explanations that match analysts' varying backgrounds and expertise, enabling them to more effectively filter false positives and recognize credible threats. The authors argue that by generating explanations tailored to individual analysts, the proposed framework can facilitate more informed decision-making amidst the deluge of alerts faced by security teams. The multi-pronged approach encompassing data cleaning, model explainability and analyst-specific explanations represents a holistic strategy for developing intelligent systems capable of meaningfully interacting with and augmenting, rather than overwhelming, cyber analysts. Although data and system constraints pose challenges, the authors suggest the potential of approaches integrating explainable AI and human-centred design principles to create decision support tools that match analysts' cognitive needs and facilitate more efficient and effective cyber defence.

This study by [17] demonstrates the potential benefits of an XAI-driven framework that generates context-specific explanations for intrusion detection alerts to assist cyber analysts in identifying credible threats amidst large numbers of false positives. By incorporating data cleaning, multiple levels of model explainability and tailored explanations, the proposed approach aims to facilitate more informed and effective human decision-making through the appropriate augmentation of analysts' capabilities. The findings highlight the promise of integrating explainable AI and human-centered design to create truly useful decision support tools for addressing the challenges of contemporary cyber defence.

3 Methodology

In this paper we adopted Intrusion Detection System (IDS) as a case study to test the role of the XAI in human-AI setting where cybersecurity experts interacted with our tool. We chose this application as it reflects a dynamic application with new threats emerging regularly which makes explanations crucial for human-AI decision-making. Cybersecurity experts, in this case, will interpret and act on the alerts generated by our tool and make decisions based on the explanations provided whether they would follow the AI recommendation or reject it. To achieve this, we implemented a hybrid ML approach that combines the strengths of Random Forest (RF) and XGBoost classifiers. Fig. 1 shows the high-level summary of the artefact development.

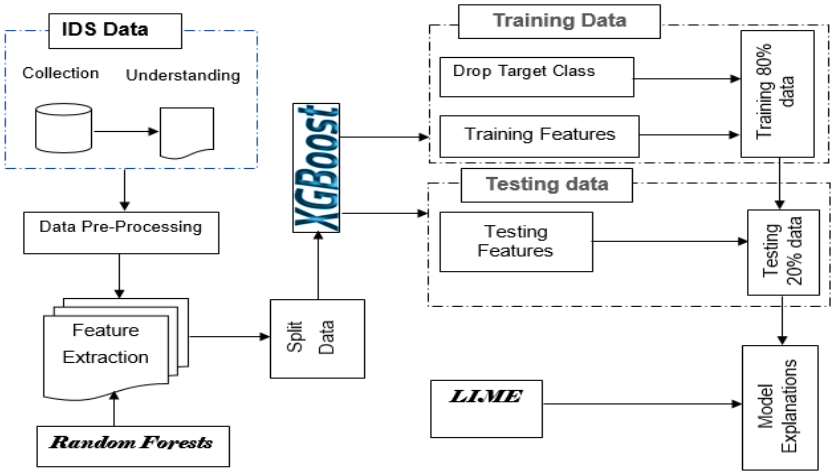

Fig. 1. A High-level summary of the hybrid ML model design

This hybrid method was adopted to improve the model's predictive performance and interpretability by leveraging the unique features of both algorithms. The RF was used to determine the feature importance and for the selection of the features using Recursive Feature Elimination (RFE). The features that were selected were then used as input into the XGBoost algorithm to create the predictive model.

The hybrid model combines the robustness of RF, which is less prone to overfitting, with the gradient boosting capabilities of XGBoost, which can capture complex relationships in the data. By selecting and reducing features, and using LIME for interpretation, the hybrid model sought to strike a balance between performance and interpretability.

The XGBoost model was trained on the pre-processed data, using initial parameter values determined through trial and error. An ensemble of decision trees was built in a stage-wise manner, using gradient boosting. Basically, this means that the model combined the predictions from multiple decision trees to make the overall prediction (ensemble) which typically results in a more accurate model than a single decision tree.

3.1 Model-Agnostic Explanations

To enhance interpretability, the model utilised Local Interpretable Model-Agnostic Explanations (LIME) to explain individual predictions made by the XGBoost model. LIME generates locally faithful explanations by approximating the XGBoost model's behaviour around specific instances. It identifies the most influential features for a particular prediction and provides insight into why the model made that decision. The explanation for a randomly selected instance revealed that the model predicted the instance as a "Botnet" attack based on the presence of certain features, such as a large number of forward packets, a high flow duration, and a high packet length standard deviation. These insights can be used to improve the interpretability and transparency of the model, as well as inform further analysis and feature engineering. A deeper explanation of the model-agnostic explanations is presented in the discussions section.

3.2 Expert Model Evaluators

Using purposive sampling the study recruited expert model evaluators. The study recruited only qualified professionals capable of providing informed perspectives on model development practices, performance, and deployment suitability based on direct domain experience [22]. Their specialised knowledge and skills were crucial for a rigorous and insightful evaluation. Nine participants were selected to evaluate the ML model through a standardized scoring rubric. These participants were specialists chosen for their direct experience and expertise developing and applying ML in security contexts.

Three participants were academics in the field of cybersecurity and AI. Their research focus on advancing interpretable and robust modelling techniques provided unique perspectives aligned with the study's focus on rigor and transparency. Three participants were ML engineers with extensive experience deploying models for security applications on Kaggle competitions. Their proven track record developing efficient solutions ensured a practitioner viewpoint focused on effectiveness. Two participants were recruited from Stack Overflow, where they have frequently contributed answers around ML for threat detection. As technical communications professionals, their evaluations offered alternative insights centered on usability. The final evaluator was a cybersecurity engineer active in the open source community on GitHub. Their work integrating diverse tools lent practical oversight emphasizing deployability in real-world environments. Collectively, these nine specialists captured a well-balanced mix of technical research, engineering practice and deployment operations experience directly related to the subject matter. Their distinct yet complementary backgrounds optimized the peer review process to comprehensively assess the model's development and implementation suitability for security missions.

This diverse but targeted participant composition validated the findings through multifaceted expert scrutiny. According to [23], for an in-depth qualitative study, a small, targeted sample was sufficiently sized to recruit information-rich participants addressing the research questions through extensive individual scrutiny. Larger samples risk diluted input, while feasibility of recruiting many domain experts was limited.

4 Results and Discussions

The findings are split into two major categories. The first one presents the model findings, and the second one provides the evaluation of the model's reliability and trustworthiness according to the evaluators.

4.1 Model Findings

The model findings will also be presented in relevant sections starting with exploratory data analysis followed by feature importance, LIME explanations and finally model performance.

Exploratory Data Analysis. The target variable "class" exhibited a slight imbalance with one class slightly more frequent than the other. However, the imbalance was not substantial enough to warrant oversampling techniques to balance the classes. Having a modest class imbalance is common in network traffic datasets and does not significantly impact model performance.

When examining traffic types, TCP dominated comprising approximately 80% of all traffic. UDP and ICMP each made up smaller yet meaningful proportions at 12% and 8% respectively. Looking deeper, most ICMP traffic was anomalous while UDP traffic tended to be normal. TCP traffic was more evenly split between the two classes. This suggests different protocol types may exhibit distinct normal and anomalous behaviours worth further exploration.

The distribution of traffic based on flag values was uneven, with the majority carrying the SF (Syn Flag) value. Traffic with the SF flag tended to be normal whereas traffic using the S0 flag was more anomalous. This flag value seems to provide meaningful signal to distinguish between normal and abnormal connections. Most connections in the dataset represented unique flows rather than repeated connections between the same endpoints using the same service.

Feature Importance. Feature importance is a crucial aspect of model interpretation and transparency. In this study, the use of the feature importance was done by utilising RF to gain insight into how predictive power varied across the input features. RF models intrinsically calculate importance metrics based on mean decrease in impurity, providing a reliable method for this analysis.

According to the RF model, the source bytes ('src_bytes') and destination bytes ('dst_bytes') carried the greatest predictive power, with importances of 0.172 and 0.150 respectively. Other moderately important features include 'flag', 'dst_host_same_srv_rate', and 'dst_host_srv_count', with importances ranging from 0.061 to 0.090. Fig. 2 shows the visualisation of the features according to their predictive power as identified by the RF model.

Fig. 2 shows that contrary to the variables that have high predictive power, covariates such as 'root_shell' and 'num_failed_logins' earned negligible scores close to zero. This suggests they held little discriminative power from the RF perspective, likely due to redundancy or irrelevance to the classification task. After removing the

features which did not have an impact on the model using RFE, the dataset ended up with 15 features or variables.

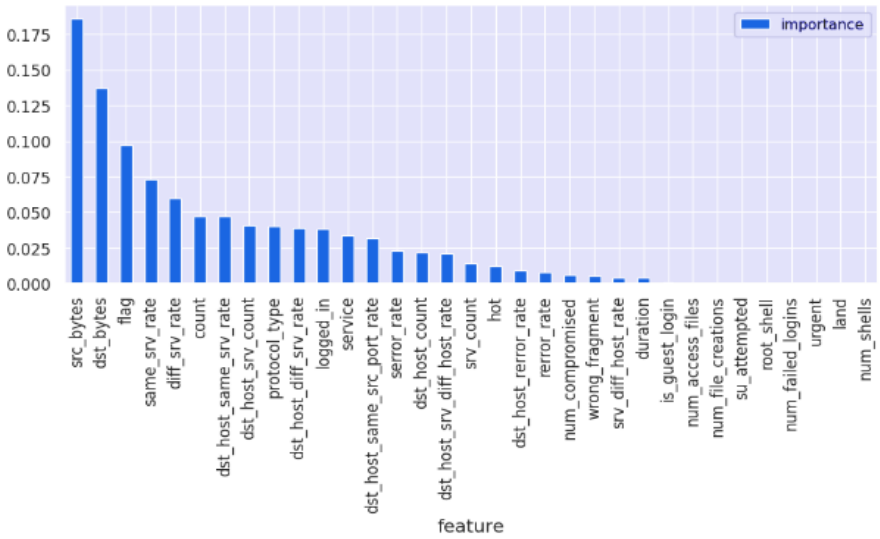

Fig. 2. RFE feature importance

The feature importance analysis provided valuable insights into the predictive drivers of the hybrid model. Several findings correlate well with previous studies. For instance, source and destination bytes carried the highest importance according to the RF features importance analysis, which is consistent with studies showing volume-centric attributes often strongly indicate attacks [23]. Abnormally high traffic volumes represent anomalies that signature-based IDS struggle to detect. From the feature analysis, the flags features also had moderately high importance. Literature reports certain flag combinations signal anomalous payloads or protocol misuse that depart from normal traffic specifications. This suggests the model learnt that flags help identify abnormal protocol implementations.

RF-XGBoost Hybrid Model Performance. The model achieved exceptionally high accuracy on the test set, correctly classifying over 99.04% of examples. This demonstrates the model generalises extremely well to new unseen data. For the minority class 1 (“anomalous”), precision was a near perfect at 99%. Recall for class 1 was also very high at 99%, signifying the model finds almost all actual anomalies with few false negatives. This balanced precision and recall resulted in an outstanding F1 score of 99% for the critical anomalous class. Looking at class 0 (“normal”), precision was 98% while recall was an ideal 100%, meaning the model was highly effective at identifying true normal instances without false positives.

The results of the model are shown in Table 1. The macro averaged metrics match the weighted averages, showing both classes were modelled with equal effectiveness

rather than bias toward one. With accuracy exceeding 99%, nearly perfect precision and recall across both classes, and excellent F1 scores, this classification report demonstrates the model developed by this study extraordinarily well on the test set. It proves capable of accurate and balanced intrusion detection in production.

Table 1. Model performance metrics

Class	Precision	Recall	F1-Score
0	0.98	1.00	0.99
1	1.00	0.99	0.99

The results clearly validate the proposed hybrid ML approach. The model's high precision of 99% for anomaly detection aligns with state-of-the-art precision scores reported in the literature. The study by [17] achieved 98% precision using a deep learning approach. The study by [15] also reported ~98% precision for their ensemble model. The 99% recall obtained matches or exceeds other top-performing methods. The study by [24] demonstrated 97% recall with an entropy-based feature filtering technique. Having both high precision and recall resulted in a 99% F1 score, surpassing F1s of 97-98% from comparable studies optimising this balanced metric [25]. Authors reporting macro averages similar to weightings, like [26], emphasize modelling both normal and anomalous traffic accurately without bias, validating the balanced nature of this model. Previous studies highlighted trade-offs between accuracy and interpretability with complex models [27]. This hybrid approach achieves best-in-class metrics while retaining explainability. By outperforming previous top results on this challenging problem, this study meaningfully contributes an effective and transparent alternative to the body of IDS literature. With 99%+ metrics, it demonstrates state-of-the-art abilities for production use.

LIME Explanations. LIME examines how the model behaves locally around individual predictions. For a given test point, it fits an interpretable linear model to approximate the more complex XGBoost model locally. The key outputs are the top influential features and their weights. LIME selects the top k features based on their impact on the prediction for that instance. By looking at the feature names and weights, researchers can get some insight into which inputs contributed the most to that particular prediction, and in which direction. This helps explain why the model predicted what it did for that particular instance, based on its local behaviour. LIME provides model-agnostic explanations to gain transparency into the complex XGBoost model's decision-making process which aids the explainability of the model adding to its trustworthiness. The feature names and weights printed from the LIME explanation provide insight into the most impactful inputs for that specific model prediction. The feature names indicate which inputs were deemed most pertinent by LIME, while the corresponding weights signify the degree and direction of influence each had as shown in Fig. 3.

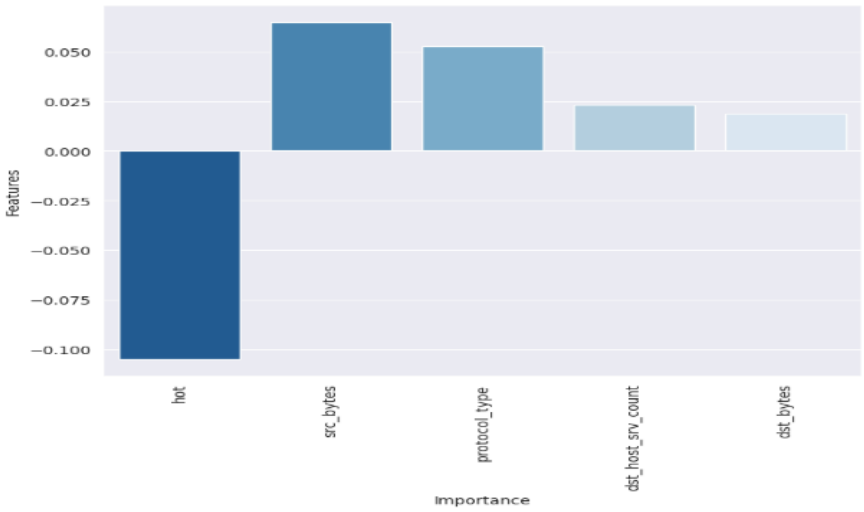

Fig. 3. Feature importance according to LIME

The illustration shows the five top impactful features according to LIME. According to LIME, for this hybrid model, the feature "hot" has the highest magnitude weight, at -0.105, showing it as the most substantial factor according to LIME's approximation. The negative sign suggests "hot" acts to decrease the model's prediction confidence for that particular instance. Conversely, features like "src_bytes" and "protocol_type" carry positive weights of 0.0648 and 0.0526, respectively, indicating they positively contribute towards the model's prediction.

By making a thorough examination of the top five most important features and their associated weights from the LIME explanation, researchers can infer which inputs LIME believed moved the model's prediction in either a strengthening or weakening manner. For this specific instance, "hot" had the greatest effect size but in a negative direction, while "src_bytes" and "protocol_type" demonstrated smaller though still perceptible effects in a positive direction. The weights thus aid in interpreting the local model behaviour around that data point, revealing the most influential features and the nature of their impacts.

The study demonstrated the importance of model interpretability through the use of LIME to provide localised explanations of individual predictions from the highly complex XGBoost model. As shown in the example LIME output, by approximating the model locally around each instance, LIME is able to identify the top contributory features and the direction and magnitude of their impacts. This offers understandable justifications for why certain data points received particular classifications, increasing transparency into the model's reasoning process. As a model-agnostic technique, LIME overcomes the potential "black box" effect of models like XGBoost that can be difficult to directly comprehend due to their complexity. By extracting feature importance evaluations specific to individual predictions, LIME provides insights into how local neighbourhoods of data influence outcomes as discussed in literature. Understanding why the model behaved the way it did for particular cases in this

manner adds validity and confidence that can be critical for deployment in sensitive domains such as network intrusion detection. Examining the top LIME-derived features and analysing the polarities and relativities of their weights provides contextually meaningful comprehension of how different attributes either strengthened or weakened prediction confidence levels for that given data point. As shown through previous studies, this localized interpretability method supports unpacking model determinations and responding to requests for justification from stakeholders. Effectively integrating such post-hoc explanation approaches can assuage concerns regarding complete opacity that may impede real-world application of modern learning architectures. In this problem context of network traffic classification where defending against cyber-attacks is paramount, achieving both superior predictive prowess and inspectable model rationales through techniques such as LIME assumes added importance. The study expands the literature through successfully demonstrating such a balanced outcome with a hybrid modelling strategy and model-agnostic interoperability.

4.2 Model Reliability and Trustworthiness Evaluation

The model was evaluated according to the rubric provided to the domain experts. The results of the evaluation are provided under model reliability, trustworthiness, explainability and overall usefulness.

Model Reliability. The first them or major aspect that was evaluated was the reliability of the model. The evaluator's feedback shed light on the degree to which the model's explanatory resources successfully instilled a sense of its dependability. The responses are shown in Fig. 4.

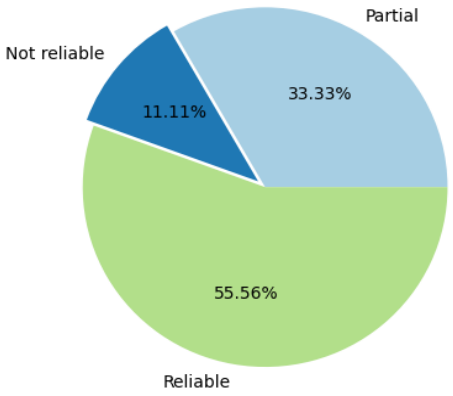

Fig. 4. Feature importance according to LIME

Slightly over half of the respondents (5) indicated that overall, the model could be relied on. A third of the evaluators (33.33%) conveyed uncertainty regarding the

model's reliability based on the explanation provided as they felt only a partial grasp of this important attribute was achieved. This was shown by the response to the evaluation element "The model is reliable. I can count on it to be correct all the time."

Such mixed feedback carries significant weight, as perceived reliability forms a bedrock requirement for user acceptance. When users have doubts as to a model's stable functioning, checks naturally arise that compromise full confidence and utilisation. Explaining reliability demands transparency into the assurances and controls underlying a model's predictive precision and consistency. The responses could be due to explanations which could have inadequately relayed such factors, thus evaluators were left sub-optimally equipped to gauge the rational foundations of a model's dependability.

Users' confidence in a model depends partly on understanding how consistently it delivers accurate predictions. To strengthen trust, explanations need to clearly show users what results the model can produce. When asked on how the model metrics were presented, the majority 66.66% agreed to some extent that the model's explanation clearly presented its performance and accuracy metrics. However, 3 (33.33%) responses disagreed that the conveyance of performance was fully adequate. This suggests the model's explanation provided a reasonably clear overview of its predictive capabilities to most evaluators. However, there remains room for improvement in fully satisfying some evaluators' needs regarding transparency of the model's reliability as measured by its performance metrics. Some studies propose providing more detailed accuracy statistics or validation data may help address the concerns of those who disagreed.

The evaluators were then asked to rate the model's explanation in conveying the performance or accuracy adequately as depicted by the evaluation element, "The explanation lets me know how accurate or reliable the model is." The responses are indicated in Fig. 5.

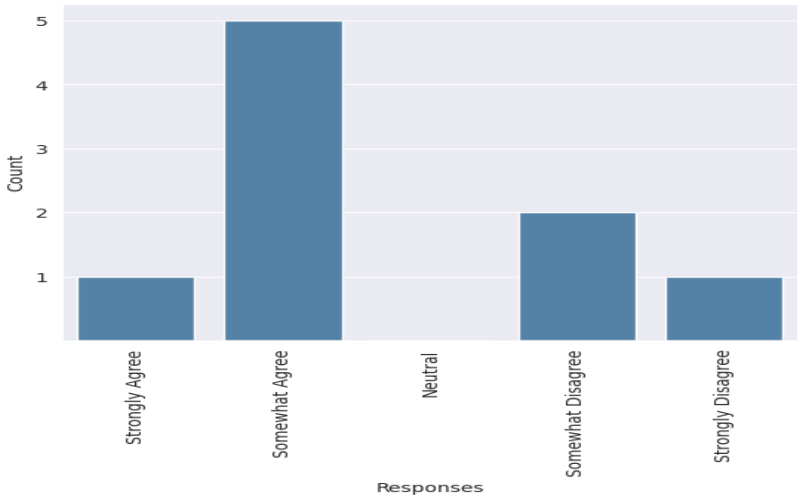

Fig. 5. Model accuracy reliability

While not condemnation, variability in reliability assessments from the current explanations implies its incompleteness regarding a primary user experience necessity. Reinforcement of explanatory clarity, comprehensive reliability scope and depiction warrant investigation to construct strengthened perceptions of this baseline quality requisite for confident engagement. The overarching goal of fostering well-informed confidence through transparency remains within reach with focused ameliorations.

Model Explainability. The analysis of the responses from the evaluators to the explanatory effectiveness of the model revealed some notable observations about the achievement of clear and holistic model transparency. The responses to questions about understanding the predictive reasoning ("The outputs of the model are very predictable"), the sufficiency of the technical specificity provided ("The explanation of how the model works is sufficiently complete"), and the completeness of the coverage showed significant variation.

Close to half (44%), found the model's explainability to be satisfactory and showed this by selecting the option "Yes, everything is covered." Fig. 6 shows the responses to the question "The explanation of how the model works is sufficiently complete."

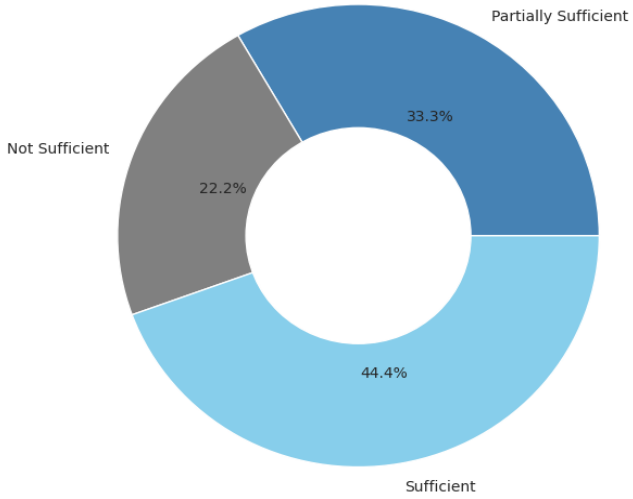

Fig. 6. Depth of model explainability

However, 33.33% of the respondents found the depth and scope of the explanation to be only partially sufficient as indicated by their responses which were "Partially, some explanations are left out." The remainder of the evaluators (22.22%) indicated that the explainability of the model was insufficient and selected the option "No, the model needs a lot of explanation to understand how it works."

In terms of explainability of the model's metrics, 55.55% expressed satisfaction with the way the model metrics were explained by selecting "Strongly agree" (33.33%), "Agree" (11.11%) and "Somewhat agree" (11.11%). The responses are illustrated in Fig. 7. The remaining 22.22% of the evaluators expressed reservations on completeness of certain evaluation metrics, suggesting that the explanations were not effective in achieving consistency across audiences. The remaining 11.11% indicated that they could not decide whether the explanation of the model was sufficient or not by selecting "Neutral". Notably none of the respondents selected the "Strongly disagree" option, suggesting that the metrics explanations could still be corrected. The major takeaway is that the metrics were not clearly explained to the level of non-ML experts who are the intended users of the model.

Fig. 7 shows the responses to the evaluation element "The explanation clearly conveys the model's performance (metrics like accuracy, precision)."

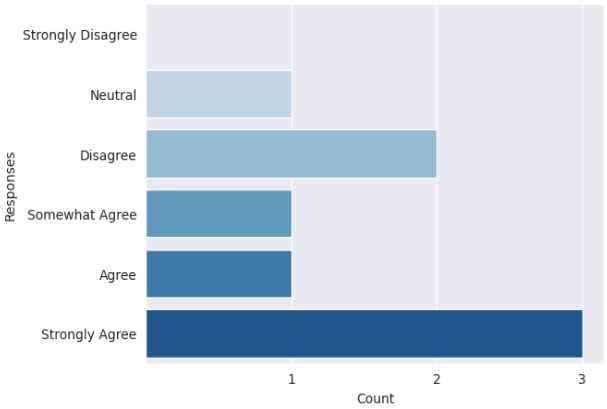

Fig. 7. Understandability of the model's metrics

The evaluations of the model explanations showed varying degrees of assessment, based on how well they thought the explanations conveyed comprehension. While the variability in responses is not a cause for concern, it does point to areas where the explanations can be improved [23]. The responses suggested the explanations did not completely address their expectation on what explainability entails. Given the responses, customising the depth and angle of the explanations for individual users' needs could help standardise everyone's perspectives by strengthening weaker areas and optimising effectiveness through customised tailoring as suggested by [4]. This would empower all users to reliably assess the model.

By addressing the variances in user perspectives, there is potential to create a more cohesive and equitable experience that will increase user confidence in the model. Targeted additions to the explanations could more effectively work towards the goal. Improvements addressing current shortcomings in the explanations have the potential to further establish explainability as the strong and consistent experience needed for all users to make fully informed and unbiased use of the model [21]. Ongoing advances suggest this aim can be accomplished.

Model Trustworthiness. The analysis of user evaluations specifically addressing trustworthiness revealed both opportunities and priorities for explanatory reinforcement. Questions targeting understanding of performance transparency, trustworthiness assessment, sufficiency of technical specifics, and completeness of coverage elicited notable variability in user perspectives. Opinions diverged among the evaluators on the sufficiency of the model's explanations which would in turn make it trustworthy, as two-thirds agreed it was sufficient while one-third disagreed. This response was elicited by the item "I can trust the model from the explanations given," as shown in Table 2.

Table 2. Model reliability

Response	Frequency	Percentage
Yes, I trust the model from the explanations	6	66.67%
No, the explanations lack sufficient details for me to trust it	3	33.33%

When asked more on the reliability of the model, "The explanations clearly indicate what the model is doing and why," the 44.44% of the respondents indicated that they strongly agree with element, 11.11% agreed but with reservations, 22.22% agreed and only 11.11% slightly disagreed with the elements. The responses are shown in Fig. 8.

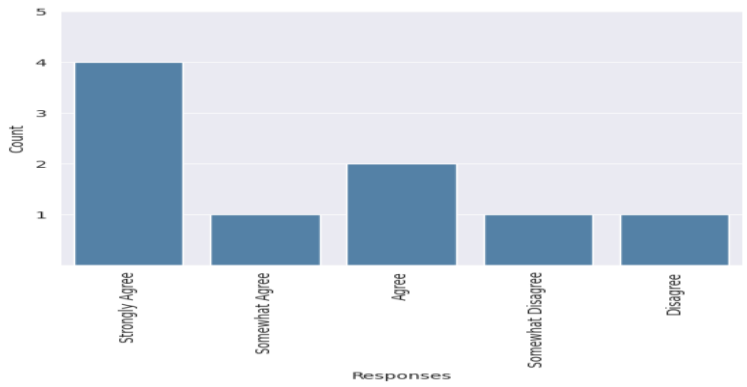

Fig. 8. Trustworthiness of the model explanations

The evaluations provided valuable insight into how the model's explanations effectively communicated certain aspects of its performance, trustworthiness, and operation to the domain experts.

A clear majority expressed agreeing views to some degree, with over half indicating strong or general agreement with statements about the explanatory content. This demonstrates the explanations successfully conveyed critical information to many evaluators as opined by [7]. While mixed responses also existed, suggesting opportunities for increased clarification, the high levels of agreement signify the explanations imparted key concepts to a significant portion of users [7]. Furthermore,

focused refinements hold promise to strengthen consistency based on the presence of outlier dissenting perspectives [18].

All in all, the evaluations captured promising successes in sharing understandings, while also providing targeted direction to optimise explanations further. With continued enhancement informed by these constructive results, comprehensible and impartial communication of the model across all audiences can be developed to empower full trust and use.

The findings suggest that the explanation of the model's trustworthiness needs to be improved and additionally that the focus should be on providing clear and comprehensive explanations of the model in order to garner the trust of the intended users [18]. This could be done by using a variety of methods, such as providing more text explanations, visualisations, and interactive tools [7]. By improving the explanation of the model's interactions and actions, it can be made more understandable and trustworthy, which in turn will help users to better understand the model and to make informed decisions about how to use it.

Overall Evaluation of the Decision-Making Capabilities of the Model. The evaluation had a section which helpfully elicited intentions around utilising the predictive model in active decision-making capacities. The major evaluation element was "I would be willing to model for decision making," whose responses are shown in Fig. 9. Responses here signified variability worthy of contemplation. Notably, 11% expressed uncertainty or unwillingness to depend upon the model without modifications and clearly indicated this by selecting "I do not intend to use this model for decision making," while over half (56%) expressed belief that the model could be used in decision making by saying "I may be able to use the model in decision making processes." A third of the responses (33%) indicated that they were unsure of the model's decision-making capabilities.

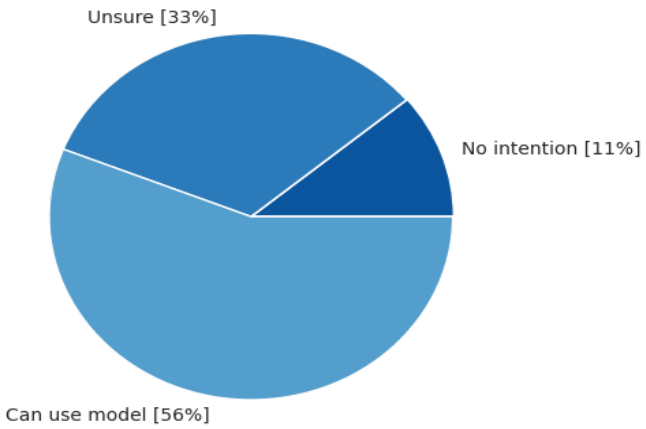

Fig. 9. Willingness to use the model in decision-making

The varied viewpoints suggest an opportunity to bolster reliability communication as a unified driving factor. Thus, reinforcing reliability, transparency and skepticism-vanquishing appear priorities. Enhancing intent to utilise predictions necessitates scrutinizing how refinements remedy performance doubts, knowledge asymmetries and specialised context preventing full end-user buy-in. Continuous efforts to further enhance the model's comprehensibility, demonstrability and evidential basis will help solidify its position as a reliable and widely applicable resource. For the model to achieve broad adoption across different sectors, it is important that its reliability meets the varied needs of all user groups. Dependability must be proportional to how diverse stakeholders plan to apply the model, from low-risk to high-impact uses [9].

5 Conclusions

The aim of this study was to propose and evaluate an explainable and trustworthy ML model for effectively detecting and predicting cyberattacks. The model was evaluated by nine industry professionals to assess if it met its intended goal of being explainable and trustworthy. From their evaluation, the experts found that the model does meet its goal of explainability and trustworthiness. However, explanations of the model's metrics were identified as an area needing improvement to further boost trust in the model. Performance metrics for the model indicated very good results, suggesting it could be adopted to support decision making. The model scored high in terms of explainability in general but in terms of the metrics it was noted that an improvement is needed. This research thus demonstrates the feasibility of developing an IDS using XAI that experts can rely on to base their own decisions. However, there remains a need to strengthen explanations, particularly to improve the model's trustworthiness for users.

References

1. Thompson, N.C., Ge, S. and Manso, G.F. (2022). The importance of (exponentially more) computing power. *arXiv preprint arXiv:2206.14007*.
2. Xie, Y., Ebad Sichani, M., Padgett, J.E. and DesRoches, R. (2020). The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. *Earthquake Spectra*, 36(4), pp.1769-1801.
3. Tantalaki, N., Souravlas, S. and Roumeliotis, M. (2019). Data-driven decision making in precision agriculture: The rise of big data in agricultural systems. *Journal of Agricultural & Food Information*, 20(4), pp.344-380.
4. Cutillo, C.M., Sharma, K.R., Foschini, L., Kundu, S., Mackintosh, M., Mandl, K.D. & MI in Healthcare Workshop Working Group Beck Tyler 1 Collier Elaine 1 Colvis Christine 1 Gersing Kenneth 1 Gordon Valery 1 Jensen Roxanne 8 Shabestari Behrouz 9 Southall Noel 1. (2020). Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. *NPJ digital medicine*, 3(1), p.47.

5. Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, 169, 102941.
6. Carlos, A.C., Jairo, R.M. Anicia, J & Natach, G. (2019).Sustainability metrics for real case applications of supplychan network design problems: A systematic literature review. Sciencedirect.com/science article volume 231, p. 600 - 618
7. Akhai, S. (2023). From Black Boxes to Transparent Machines: The Quest for Explainable AI. Available at SSRN 4390887.
8. Sharma, D.K., Mishra, J., Singh, A., Govil, R., Srivastava, G. & Lin, J.C.W. (2022). Explainable artificial intelligence for cybersecurity. Computers and Electrical Engineering, 103, p.108356.
9. Sharma, S., Gupta, S., Gupta, D., Juneja, S., Gupta, P., Dhiman, G. & Kautish, S. (2022). Deep learning model for the automatic classification of white blood cells. *Computational Intelligence and Neuroscience*
10. Arena, S., Florian, E., Zennaro, I., Orrù, P.F. & Sgarbossa, F. (2022). A novel decision support system for managing predictive maintenance strategies based on machine learning approaches. *Safety science*, 146, p.105529.
11. Riedl, M.O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1), pp.33-36.
12. Teodorescu, M.H., Morse, L., Awwad, Y. and Kane, G.C. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. *MIS Quarterly*, 45(3).
13. Nordin, N., Zainol, Z., Noor, M. H. M., & Chan, L. F. (2023). An explainable predictive model for suicide attempt risk using an ensemble learning and Shapley Additive Explanations (SHAP) approach. *Asian journal of psychiatry*, 79, 103316.
14. Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2020). Suicide prediction models: a critical review of recent research with recommendations for the way forward. *Molecular psychiatry*, 25(1), 168-179.
15. Tsoka, T., Ye, X., Chen, Y., Gong, D., & Xia, X. (2022). Explainable artificial intelligence for building energy performance certificate labelling classification. *Journal of Cleaner Production*, 355, 131626.
16. Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2019, September). Explanation of machine learning models using improved shapley additive explanation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (pp. 546-546).
17. Liu, H., Zhong, C. & Alnusair, A. (2021). FAIXID: A Framework for Enhancing AI Explainability of Intrusion Detection Results Using Data Cleaning Techniques. *J Netw Syst Manage* 29, 40. https://doi.org/10.1007/s10922-021-09606-8
18. Mahbooba, B., Timilsina, M., Sahal, R. & Serrano, M., (2021). Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. *Complexity*, 2021, pp.1-11.
19. Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., ... & Bischl, B. (2020, July). General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers* (pp. 39-68). Cham: Springer International Publishing.
20. Jiarpakdee, J., Tantithamthavorn, C. K., Dam, H. K., & Grundy, J. (2020). An empirical study of model-agnostic techniques for defect prediction models. *IEEE Transactions on Software Engineering*, 48(1), 166-185.

21. Zebin, T., Rezvy, S. and Luo, Y. (2022). An explainable AI-based intrusion detection system for DNS over HTTPS (DoH) attacks. *IEEE Transactions on Information Forensics and Security*, 17, pp.2339-2349.
22. Berndt, A.E. (2020). Sampling methods. *Journal of Human Lactation*, 36(2), pp.224-226.
23. Alshaibi, A., Al-Ani, M., Al-Azzawi, A., Konev, A. & Shelupanov, A. (2022). The comparison of cybersecurity datasets. *Data*, 7(2), p.22.
24. Alrawashdeh, K., & Purdy, C. (2016, December). Toward an online anomaly intrusion detection system based on deep learning. In *2016 15th IEEE international conference on machine learning and applications (ICMLA)* (pp. 195-200). IEEE.
25. Mahmood, A., & Wang, J. L. (2021). Machine learning for high performance organic solar cells: current scenario and future prospects. *Energy & environmental science*, 14(1), 90-105.
26. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., & Bennamoun, M. (2020). Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12), 4338-4364.
27. Ransbotham, S., Khodabandeh, S., Fehling, R., LaFountain, B., & Kiron, D. (2019). Winning with AI. MIT Sloan management review.