JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Transparent Threat Detection Using SHAP and LIME to build an Explainable Intrusion Detection System

[1]Dharaneesh Kuruba, [2]Gowtham R, [3]Monish V, [4]Vinayak Naik, [5]Dr. Deepthi V S
[1 2 3 4]Student, [5]Professor,
[1 2 3 4 5]Computer Science Engineering (Cyber Security),
[1 2 3 4 5]Dayananda Sagar College of Engineering, Bangalore, India

Abstract: The increasing rate of cyber threats requires the high level of accuracy in Intrusion Detection Systems (IDS). Nevertheless, the contemporary ML-driven IDS can be considered rather opaque black boxes that can be highly accurate but lose interpretability and credibility. This is a highly important impediment to quick decision-making in Security Operations Centres (SOCs), which is caused by the related fatigue in alerts. The present project reviews most recent works on Explainable Intrusion Detection Systems (XAI-IDS) that combine LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations). The purpose of this is to build a high-performance IDS framework relying on such models as XGBoost based on benchmark IDS datasets (e.g., CIC-IDS2017, UNSWNB15). We have shown that XAI can contribute substantially to the levels of analyst trust and that approaches based on XGBoost and SHAP give the best trade-offs between interpretability and accuracy. The proposed methodology is devoted to the effective XAI integration that will produce human-readable explanations of the predictions used in the IDS and will be checked with the help of the visualization and performance metrics to assure actionable results that will be offered to SOC analysts.

IndexTerms **- Explainable AI (XAI), Intrusion Detection Systems (IDS), Transparency, Cybersecurity, Machine Learning.**

I. INTRODUCTION

The progressive complexity and the increasing rate of online attacks have dictated the implementation of new sophisticated detection algorithms, which have resulted in the intensive use of Machine Learning (ML) models in Intrusion Detection Systems (IDS). Although these ML-based IDS are high accuracy and highly effective in recognizing complex and zero-day attacks, these supercomputers tend to be black boxes; that is, their decision-making is not transparent, and human operators cannot understand why they made these decisions. This lack of transparency and interpretability in itself poses a considerable set of operational complications in Security Operations Centres (SOCs). The suspicion of malicious activity that is not accompanied by an explanation when using an IDS, showing the lack of trust by the analysts, makes it harder to validate the alerts, and reduces the important response to an incident. Such an ongoing effort to comprehend the logic of the system is directly responsible to the high-level problem of alert fatigue, where analysts are likely to ignore or overinterpret important alerts, thus jeopardizing security. To address the mismatch in high performance and human understanding, the discipline of Explainable Artificial Intelligence (XAI) has been developed. This paper is devoted to building the effective model of Transparent Threat Detection and incorporating two prominent paradigm XAI-independent methods, SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-agnostic Explanations) into the system of high-performance IDS. The goal is to generate an IDS that is not only effective at identifying intrusion, but also, explains its predictions in a manner that is easily comprehensible by a human, thus enhancing the analyst confidence, alleviating alert fatigue and permitting the security response to be faster and more reliable.

II. LITERATURE SURVEY

The increasing sophistication of network-based attacks and intricacy has developed the emergent necessity of clear and explainable Intrusion Detection Systems (IDS). Recent papers have concentrated on the means of incorporating Explainable Artificial Intelligence (XAI) techniques especially SHAP and LIME in order to enhance interpretability, trust and efficiency in IDS.

A. Principles in Explainable IDS

The theoretical background of SHAP was provided by Lundberg and Lee [1], who proposed additive features attributions, and that of LIME was proposed by Ribeiro et al. [12] to serve model-agnostic local explanations. Based on them, Arreche et al. [2] synthesized SHAP and LIME in pipelines of IDS, enhancing the interpretability of datasets. Ensemble models with LIME in Patil et al. [25] were found to use high detection accuracy, but Wang et al. [11] and Alam et al. [5] demonstrated the applicability of SHAP in deep-models but noted scalability problems.

B. Public and Deep-Learning XAI Models

Autoencoders and CNNs with SHAP/LIME were utilized by Ahmed et al. [4] and Alam et al. [5] to find latent anomalies, and Zhou et al. [9] used GNNs to predict traffic topology. The stability of SHAP when used with ensemble models for predicting the future was verified in a study by Kim et al. [7] and Khan et al. [13], whereas Ryu et al. [19] suggested attention mechanisms that provide built-in transparency without post hoc overhead.

C. Dataset based Studies and Comparisons

Ahmad et al. [19] and Yadav and Sinha [22] used the LIME to the IoT data such as BoT-IoT and TON-IoT to provide lightweight edge explanations. Kumar and Bansal [17] and Zhou et al. [9] demonstrated that SHAP can reveal domain-specific anomalies of cloud and SCADA settings. Hermosilla et al. [21] compared SHAP and LIME to UNSWNB15, with global/local trade-offs being reported, whereas Coroama and Groza [14] emphasized inconsistent metrics of XAI evaluations.

D. Explainability in SOC and SIEM

Ban et al. [8] constructed an AI-assisted SIEM that has interpretable SVM justification, and Takahashi et al. [15] showed that SHAP/LIME dashboards are superior to SOC decision making. Gupta [18] examined privacy-preserving, federated LIME, whereas Rastogi et al. [20] measured the effect of explainable alerts on reducing the cognitive load in analysts and enhancing the response efficiency. The research on mobile ad hoc network security indicates that blackhole attacks cause serious deterioration of AODV routing performance, at the same time, better multiphase detection methods can be used to decrease the number of packet losses and improve the general network reliability [31][32].

E. Systematic Reviews, Frameworks, and Domain Applications

Mohale and Obagbuwa [3] reviewed XAI-IDS techniques and observed scalability and real time limitations. Domain aware explanation, as shown by Islam et al. [23], enhances the generalizability of the model to various instances of intrusion detection. Zebin et al. [24] used XAI on DoH detection, and Patil [25] created hybrid visualization systems to simplify reasoning by an analyst. Standard benchmarks and metrics of quantitative assessment have been suggested by Coroama and Groza [14] and Kim and Cho [15].

F. Findings and General Shortcomings

SHAP over 25 studies has been able to explain globally stable tree and ensemble models [1][7][11][13], but LIME can give instance-specific insights to lightweight environments and IoT applications [6][12][19][22]. Hybrid pipelines, as well as attention-based approaches [19], provide a better explainability with no post-processing effort. Nevertheless, significant issues have still not been overcome SHAP computing cost [1][11], insufficient benchmark diversity [2][3][21], poor standardized quality indicators [14][15], and low adversarial resilience continue to be problematic, suggesting the need to develop scalable, real-time, and robust explainable IDS systems.

All these studies prove the idea that XAI methods when used with IDS can help overcome the performance and interpretability gap. They indicate the further directions of advancements: reducing the latency of SHAP and LIME, developing uniform evaluation metrics, making models resistant to adversarial attacks, and modifying explainability to privacy-oriented systems that involve human-in-the-loop checks.

III. RESEARCH GAPS

Explainable Intrusion Detection System (XAI-IDS) Literature has outlined some of the existing gaps that have constrained the use and scalability of explainable intrusion detection systems in real-world cybersecurity settings:

- **Experimental computation volumes**: The computational cost of the majority of SHAP-based frameworks is typically very high, which renders real-time or scale-based implementation in SOC operations impossible [1,5,11,16].

- **Absence of standardized evaluation metrics**: There is no agreement on how the quality, fidelity or human usefulness of explanations can be measured quantitatively and so there is a lack of consistency in the evaluation of the studies [3,14,15].

- **Weak human-centric validation**: There is weak engagement of analysts in the validation loop of many systems to make sure that explanations can be comprehended, acted upon and consistent with the actual investigative processes [8,20].

- **Adversarial vulnerability**: The existing XAI-IDS system is hardly tested on adversarial manipulation, and thus their interpretation and forecasting are vulnerable to adversarial attacks [16,18].
- **Dependency on datasets and generalization**: It becomes dependent on the old datasets, such as CICIDS2017 and NSL KDD, which restricts generalization to new attack patterns and dynamic traffic patterns [2,3,21]

The system offers a solution to these gaps by introducing a light weight, human friendly XAI-IDS, resistant to attacks, human-in-the-loop validation, and built in visualizations, which are aimed to execute efficiently in real time in SOC environments.

TABLE I: Summary of key literature and relevance

No	Paper Title	Authors	Problem Addressed	Methodology	Key Findings	Limitations	Relevance to Present Work
1	A Unified Approach to Model Interpretation	Lundberg, S.M.; Lee S.-I.	Lack of a unified, theoretically sound method to explain predictions from any machine learning model.	Foundational SHAP Framework (SHapley Additive exPlanations); mathematical proofs; introduced Kernel SHAP and Deep SHAP.	Established SHAP as the unique additive feature attribution method satisfying desired properties (Local Accuracy, Missingness, Consistency).	High computational complexity for exact Shapley value calculation in large datasets/models.	Foundational. It is the basis for SHAP, the most used XAI method in IDS literature ([2], [5], [7], [16], etc.).
2	XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network IDS	Arreche, O., et al.	Low trust and lack of actionable insights in traditional black-box IDS models.	Proposed an end-to-end XAI pipeline; employed multiple ML/DL models; used SHAP and LIME for post-hoc explanation and feature selection.	The XAI pipeline improved overall performance and provided clear explanations; feature selection reduced training time significantly.	Generalizability is dependent on the specific training data distribution (CICIDS-2017, NSL-KDD).	Proposes a contemporary, comparative XAI framework directly applicable to network IDS enhancement.
3	A Systematic Review on the Integration of Explainable AI in Intrusion Detection Systems	Mohale, V.Z.; Obagbuwa I.C.	Lack of a structured overview of XAI adoption, evaluation, and future trends in IDS.	Systematic Review (meta-analysis and synthesis) of recent research papers on XAI-IDS.	Identified common XAI techniques (SHAP, LIME) used in IDS and categorized research gaps in evaluation and deployment.	It is a conceptual review; does not offer new experimental results or quantitative validation.	Provides a comprehensive, up-to-date map of the XAI-IDS landscape, defining current trends and challenges.
4	Hybrid Deep-Learning and Explainable AI for Anomaly Based Intrusion Detection	Ahmed, H., et al.	Difficulty in interpreting complex anomaly-based deep learning decisions in IDS.	Hybrid Deep Learning Model (e.g., Autoencoders or CNNs) for detection, integrated with a post-hoc XAI method (SHAP/LIME).	Achieved superior anomaly detection accuracy while providing feature contribution scores for each anomaly alert.	The "hybrid" nature adds complexity and computational over-head compared to single models.	Shows how to successfully pair high-performance DL with XAI for anomaly detection, a key IDS task.
5	Explainable Deep Intrusion Detection Model Using SHAP Values and Feature Attribution	Alam, M. R., et al.	Opacity of deep learning (DL) models hinders their deployment in mission-critical IDS.	Developed a Deep Learning Model and utilized SHAP for feature attribution and interpretation.	SHAP values effectively revealed which network features drove the DL model's detection decisions.	SHAP computation adds overhead, making real-time explanation in high-speed networks challenging.	Validates SHAP's effectiveness specifically for explaining Deep Learning-based IDS decisions using benchmark datasets.
6	LIME-based Explainable Network Intrusion Detection for IoT Devices	Bhattacharya, A., & Das, P.	Resource constraints and the need for localized trust in IDS for IoT environments.	Applied standard Machine Learning models (ML) and used LIME for local model explanation.	LIME successfully provided timely and locally relevant explanations, which is crucial for edge/IoT de- vices.	LIME's local approximation can be unstable and inaccurate if the local decision boundary is highly non- linear.	Addresses a key sub-domain: XAI for resource-constrained IoT networks, prioritizing the LIME method.
7	SHAP-based Interpretation of Network Anomalies Using Ensemble Learning	Kim, K. S., et al.	Interpreting the combined, opaque decisions of high-performing Ensemble Learning models.	Implemented Ensemble ML (e.g., Random Forest) and used SHAP to interpret anomaly predictions.	SHAP successfully interpreted the ensemble model's output, offering high accuracy and clear, counterfactual anomaly explanations.	Ensemble model complexity coupled with SHAP leads to significant training and explanation latency.	Demonstrates a practical solution for interpreting the often-used Ensemble IDS black-box architecture.
8	Breaking Alert Fatigue: AI-Assisted SIEM Framework Using Explainable SVM	Ban, T., et al.	Excessive false positives (alert fatigue) in SIEM systems and lack of context for security analysts.	Designed an Instance-weighted SVM (IWSVM) within an AI- assisted SIEM framework with visualization (ATG).	XAI visualization and IWSVM significantly reduced alert fatigue and improved the speed and confidence of security triage.	Limited reproducibility since the experiments were run on proprietary enterprise SIEM alert data.	Directly addresses the operational pain point of security analysts using XAI to enhance trust and efficiency.
9	Interpretable Cyber Threat Detection Using SHAP and Graph Neural Networks	Zhou, L., et al.	Traditional IDS models fail to capture complex network dependency relationships for threat detection.	Combined Graph Neural Networks (GNN) to model network entities with SHAP for local explanation.	GNN-SHAP provided high-fidelity explanations that accounted for the network's topology and relationship-based features.	GNN models are inherently computationally intensive, and adding SHAP further increases resource demands.	Innovative use of GNNs for relationship-based detect ion, making its interpretability particularly valuable.
10	Securing SCADA Systems: IDS with Shapley Analysis	Gaj, P., et al.	Need for high accuracy and trans- parent decision-making in critical SCADA/ICS environments.	Applied ML models (DT, RF, MLP) and utilized SHAP for protocol-based feature attribution.	Achieved high accuracy in the Modbus protocol dataset, with SHAP providing clear, modular at- tack insights.	The evaluation lacked a deep learning model comparison, which are increasingly used in SCADA security.	Highly relevant to Industrial IoT/SCADA security, a domain where interpretability is mandated for safety.
11	Explainable ML Framework for IDS	Wang, M., et al.	Difficulty in implementing a general, explainable framework for complex DL models in IDS.	Proposed a general ML framework using a DNN classifier and integrating SHAP for explanation generation.	Successfully provided both local and global explanations for the DNN's output on network traffic.	SHAP's inherent slow computation limits its use in high-throughput, real-time network security systems.	An influential early paper that explored the feasibility and limitations of DNN-SHAP integration for IDS.
No	**Paper Title**	**Authors**	**Problem Addressed**	**Methodology**	**Key Findings**	**Limitations**	**Relevance to Present Work**
12	"Why Should I Trust You?" Explaining Any Classifier	Ribeiro, M.T.; Singh S.; Guestrin C.	Black-box models offer high performance but no insight, limiting trust and deployment.	Foundational LIME Framework (Local Interpretable Model-agnostic Explanations); uses a simple surrogate model to explain local predictions.	Introduced LIME as a fast, model- agnostic local explanation method; proved its effectiveness on diverse data types (text, images, tabular).	Local explanations can be unstable across small perturbations; the linear approximation may be poor for highly non-linear regions.	Foundational. The basis for LIME, the second most used XAI method in IDS for its speed and simplicity.
13	Explainable Ensemble IDS Based on Feature Attribution	Khan, H., et al.	Black-box nature of ensemble models prevents security teams from understanding combined detection logic.	Employed Ensemble ML techniques and used a Feature Attribution method (e.g., permutation/Shapley-based).	Provided feature-level attribution for ensemble decisions, allowing security experts to audit the combined model's logic.	The exact XAI method used (specific type of feature attribution) is less generalizable than SHAP or LIME.	Directly addresses the interpretability of Ensemble Models, which are ubiquitous in commercial IDS solutions.
14	Evaluation Metrics for XAI Techniques	Coroama, L.; Groza A.	Lack of standardized, rigorous metrics for quantitatively evaluating the quality of an XAI explanation itself.	Theoretical Review and taxonomy; synthesized existing fidelity, stability, and	Proposed a taxonomy for XAI evaluation metrics; highlighted the crucial theory–practice gap in XAI adoption.	Purely theoretical; does not provide empirical results or test new XAI- IDS	Crucial for research. Defines how XAI-IDS proposals should be measured beyond just

				complexity metrics.		implementations.	IDS accuracy.
15	Benchmarking Explainable AI Models for Intrusion Detection: A Comparative Study	Kim, S., & Cho, Y.	Absence of a comparative study that rigorously evaluates different XAI models on both security performance and explainability metrics.	Formal Benchmarking Study; used multiple XAI models and multiple datasets; evaluated against fidelity, stability, and detection metrics.	Provided clear, quantitative evidence of the trade-offs between XAI models (e.g., speed vs. fidelity).	The "best" XAI method remains context-dependent based on the specific trade-off prioritized by the user.	A valuable, recent benchmarking study that informs the choice of XAI model for new IDS applications.
16	Introspective IDS Through Explainable AI	Guvenc Paltun, B.; Fuladi R.	Detecting and explaining real- time adversarial attacks and novel threats in dynamic networks.	Developed an Autoencoder-based IDS and used global/local SHAP for introspective analysis of detection failures.	Demonstrated the ability to detect and explain evasive adversarial samples in real time within an SDN testbed.	Limited external validation against a broader, more diverse set of advanced persistent threats (APTs).	Addresses the advanced challenge of real-time XAI in adversarial defense, a state-of-the-art topic.
17	XAI-Augmented IDS for Cloud Environments: An Interpretable Deep Learning Framework	Kumar, D., & Bansal, A.	Security in dynamic, multi-tenant Cloud environments requires auditable, interpretable threat detection.	Proposed an Interpretable Deep Learning Framework specifically tailored for cloud network traffic analysis.	Showed that XAI-augmented DL models can maintain high accuracy while providing audit logs that satisfy regulatory requirements.	Computational overhead and de-pendency on the cloud provider's API for data collection are practical constraints.	Focuses on a high-growth domain, Cloud Security, where transparency and auditability are paramount.
18	Interpretable Feature Attribution for Federated IDS Models	Gupta, P.	Lack of transparency and trust in models trained on decentralized, private data via Federated Learning (FL).	Applied a Feature Attribution method to an FL-trained IDS model without accessing raw private data.	Successfully provi ded interpretability for a model trained on distributed data, maintaining data privacy guarantees.	Explanation quality relies on ap-proximations needed to circumvent the data privacy restrictions of the FL setup.	Highly relevant for decentralized and privacy-preserving security systems using Federated Learning.
19	Attention-based Explainable Intrusion Detection Using Deep Neural Networks	Ryu, J., et al.	Post-hoc XAI methods (SHAP/LIME) are too slow for fast IDS systems.	Designed a Deep Neural Network with an Attention Mechanism for intrusion detection.	The attention weights serve as an intrinsic (ante-hoc), fast explanation of feature importance, integrated directly into the model.	Attention weights are not mathematically guaranteed to satisfy the fairness and consistency properties of methods like SHAP.	Pioneers ante-hoc (inherently interpretable) DL models, offering a solution to the real-time speed limitation.
20	Measuring Security and Cognitive Impact in AI-Driven SOCs	Rastogi, N., et al.	Need to quantify the human-factors benefits (trust, efficiency) of XAI in Security Operations Centers (SOCs).	Used a Mixed-Methods approach (quantitative metrics and qualitative analyst data) within the COR-TEX framework.	Found that XAI significantly reduced analyst triage time and cognitive load, leading to faster incident resolution.	Limited to a simulated SOC environment, which may not perfectly replicate the pressure of a real-world environment.	Focuses on the socio-technical aspect of XAI, providing empirical evidence of value to human operators.
21	SHAP vs. LIME for Forensic IDS Analysis	Hermosilla, P., et al.	Uncertainty over which XAI method is best for post-incident forensic and auditing purposes.	Direct Comparative Study of SHAP and LIME using highly accurate models (XGBoost, TabNet).	Validated the superior explicability and consistency of SHAP for forensic analysis compared to LIME's instability.	Forensic analysis requires highly stable explanations, but both methods incur significant computational cost for this task.	Essential for security teams performing post-mortem analysis and auditing of past intrusion events.
22	SHAP and LIME for Trans- parent Intrusion Detection in IoT Environments	Yadav, J., & Sinha, P.	The dual challenges of low- resource operation and high-stakes decision-making in IoT IDS.	Applied and compared SHAP and LIME for transparency in ML-driven IDS within an IoT context.	Provided detailed comparative in- sights on the trade-offs between SHAP (consistency) and LIME (speed) for IoT devices.	The resource overhead for running SHAP on some extremely low-power IoT devices remains prohibitive.	A targeted study on the comparison of the two dominant XAI methods (SHAP and LIME) for IoT security.
23	Domain Knowledge Aided Explainable AI for IDS	Islam, S.R., et al.	Explanations must be meaningful to security experts, which requires incorporating cybersecurity domain knowledge.	Proposed infusing Domain Knowledge (CIA principles) into the model and explanation generation via pre- and post-processing.	Domain knowledge improved the relevance and actionability of explanations and boosted the generalizability of the IDS model.	The formalization of domain knowledge and its universal application across all XAI methods remains a challenge.	Highlights the need for actionable, domain-aware explanations rather than purely mathematical feature scores.
24	XAI for DNS over HTTPS Attack Detection	Zebin, T., et al.	Evasive attacks like DNS over HTTPS (DoH) require transparent detection to aid security analysts.	Used a Stacked Random Forest model and applied SHAP to analyze features related to the DoH protocol.	SHAP identified key features (e.g., payload size, domain entropy) re- sponsible for DoH attack classification with high precision.	The model is specialized and limited to differentiating between legitimate and malicious DoH traffic, not general network attacks.	Relevant for tackling modern, evasive attack vectors where protocol level feature importance is key.
25	Explainable AI for Intrusion Detection System	Patil, S., et al.	Traditional ensemble models pro- vide high accuracy but no insight into their voting process.	Used a Voting Ensemble of classifiers and utilized LIME to explain the collective decision of the ensemble.	Achieved a high 96.25% accuracy; LIME provided a local, accessible explanation of the ensemble's black-box output.	The use of SMOTE may introduce overfitting, and the results are limited to a single dataset (CICIDS- 2017).	A good example of how to apply LIME as a lightweight post-hoc method to explain highly accurate ensemble models.

IV. METHODOLOGY

The project is also designed with a systematic approach to produce an accurate and interpretable explainable intrusion detection system (XAI-IDS). The methodology can be broken into five modules, each of which deals with a pivotal point of the system pipeline, such as creating realistic datasets and training high-performance predictive models, including ex- plainability, performance, and the presentation of the insights in visual ways. This framework is modular in nature to make sure that the system is not only efficient in detecting intrusions but it is also transparent in its operation and makes sense to cybersecurity analysts.

A. Dataset Loading & Preprocessing

This module chooses benchmark datasets of IDS, such as CICIDS2017, UNSW-NB15, and NSL-KDD. It takes care of data cleaning, normalization, encoding categorical features, dealing with the imbalance of classes (e.g., using SMOTE) and feature selection aimed at focusing on the most relevant attributes.

B. ML Model Training

This module trains the high-performance models like Random Forest, XGBoost, and DNN, and hyperparameters are also tuned to make the model resilient to known and zero- day attacks. Standard metrics (Accuracy, Precision, Recall, F1- score, AUC-ROC) are used to evaluate the quality of models whilst hyperparameters are optimized to make them resistant to both known and zero-day attacks.

C. Explainability Module (SHAP/LIME)

SHAP gives the global feature importance, whereas LIME gives the local explanation of individual predictions. These methods assist the analysts to know the reason why cases are either considered malicious or benign.

D. Performance & Interpretability Evaluator

This module can evaluate the performance and interpretability of different models based on standard IDS metrics as well as explainability-oriented indicators such as Fidelity, Stability, and Sparsity to make the results reliable and com- prehensible.

E. Visualization and Reporting

Dashboard shows real-time alerts that have SHAP/LIME explanations of these alerts and significant features that lead to each alert. It provides a simple, visual representation of the complex model decisions and allows the easy exportation of reports that indicate the performance and trends of the major features.

V. SIGNIFICANCE OF PROPOSED SYSTEM

The proposed Explainable Intrusion Detection System (XAI-IDS) provides the essential advantages over the traditional IDS because it provides the accurate threat detection with the transparent and practical explanations. It is designed with efficiency, interpretability, and usability with regards to cybersecurity analysts. The crucial details of the system are:

A. Real-Time Threat Detection with Transparency

Gives real-time explanations of flagged network activities that can be easily understood to enable the analyst to make quick and informed decisions.

B. Human-Centric Validation

Makes the descriptions of models practical and interpretable to gain greater confidence with AI-based warnings.

C. Adversarial Resilience

Uncorrupted prediction and explanation even on evasion attacks

D. Adaptive and Scalable Design

The modular design can be easily configured with new datasets, models or explainability techniques, enabling the system to be flexible to changing network environments.

E. Visualization and Reporting

Converts the complex decisions of the model in to simple easy-to-understand visuals and reports, simplifies the tracking of trends and monitoring performance and disseminating in- sights with stakeholders.

System Architecture: Architecture diagram (as mentioned in Fig. 1) will depict the general processes of the proposed Explainable Intrusion Detection System (XAI-IDS). It involves two key steps, including offline training and live threat analysis. During the offline stage, network datasets, such as CIC-IDS2017, UNSW-NB15 are preprocessed and trained with the XGBoost, and the resulting trained model is stored to be deployed. When the user activity is being tracked in real time, the live stage starts, the incoming events are processed by the model, and it defines whether they are benign or malicious. In case of threat detection, SHAP and LIME produce human-readable explanations, which are then shown on the dashboard to be countered by security analysts.

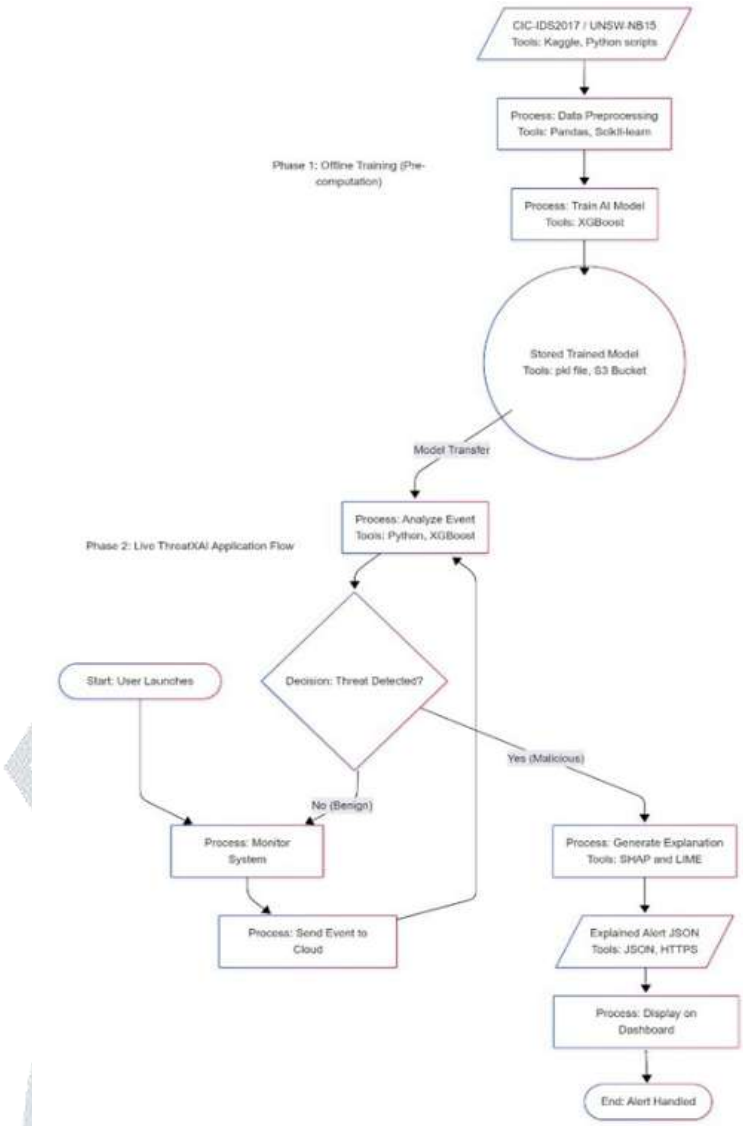

Fig. 1. Proposed Architecture of Explainable Intrusion Detection System (XAI-IDS) integrating SHAP and LIME for model interpretability.

VI. RESEARCH OPPORTUNITIES AND FUTURE WORK

Although the current project outlined major progressions, there are still a number of research areas that could be explored in the future towards a better understandable intrusion detection system (X-IDS):

A. Coherent XAI Framework of Cybersecurity

The tools available, most of which explain certain types of models, are designed independently. The next step of work might be creating a single framework that will be able to harness numerous XAI methods and choose the most appropriate method based on the type of machine learning model and type of data, simplifying the interpretability process overall across deployments of the IDS (Doshi-Velez and Kim, 2017).

B. Explainability in Real Time in Live Networks

Although the existing XAI techniques offer post hoc explanations, they cannot be applied in real time intrusion detection because of the computational complexity. A study might investigate the existence of light or rough XAI methods that can produce near-real-time explanations that can be used in the analysis of mass network traffic without necessarily compromising the interpretability (Zhang et al., 2022).

C. Introducing Threat Intelligence and XAI-IDS to Each Other

Systems recommended may be used to integrate SHAP and LIME explanations with threat intelligence feeds to assist in prioritizing and evaluating model responses according to established attack patterns. This hybrid method would be able to improve automated threat classification and allow analysts to trace attack campaigns in a more efficient way.

D. Human Concentrated Evaluations of Explanations

The existing XAI evaluation measures are mainly quantitative. One of the new directions that are going to be incorporated is the use of human-centered evaluation frameworks, in which the understandability and usefulness of explanations can be evaluated by cybersecurity experts. This will assist in fine- tuning instances of explanation and make them most practical in SOCs (Bhatt et al., 2020).

E. Explainability Preserving System-to-System Privacy

As more people worry about their data privacy, it is possible to expand explainable IDS to federated learning systems, where it is possible to both trains distributed models and provide explainability, without the need to move data centrally. This may enhance the data confidentiality and interpretability (Liu et al., 2021).

F. Explainability With Reinforcement Learning

The other possible area of research is to utilize reinforcement learning to dynamically adjust explanation strategies based on the feedback of analysts. The system might learn what sort of explanation (e.g. feature-based or instance-based) would give the most value to specific types of attacks or conditions.

VII. CONCLUSION

This paper introduces a holistic model in developing an explainable intrusion detection system with the combination of SHAP and LIME. Its methodology provides a clear flow, i.e. data preparation, model training, interpretability assessment, and visualization, which provide machine learning- driven security systems to be more transparent and actionable. This method increases the effectiveness of intrusion detection because the analysts are empowered to interpret the reasoning behind the models predictions, thereby causing them to have more trust and accountability in the processes. The next steps are expected to be on the further evolution of explainable AI and its scalability and real-time applicability to cybersecurity, which will allow the next generation of IDS that is not only able to detect threats accurately but also provide intelligent justification to its actions. The combination of explainability and lifelong learning, federated intelligence and human-in-the-loop mechanisms is the direction towards reliable, adaptive, and more resilient security systems that can adapt to the changing cyber threats.

VIII. ACKNOWLEDGMENT

The authors would also like to say their sincere thanks to Dr. Deepthi V. S. at the Department of Computer Science and Engineering (Cybersecurity) at Dayananda Sagar College of Engineering, who has guided, advised, and supported them throughout the research period with the invaluable guidance, insightful advice, and untenting support.

REFERENCES

[1] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Model Interpretation," *NIPS Conference*, 2017.
[2] O. Arreche, *et al.*, "XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network IDS," *Applied Sciences*, 2024.
[3] V. Z. Mohale and I. C. Obagbuwa, "A Systematic Review on the Integration of Explainable AI in Intrusion Detection Systems," *Frontiers in AI*, 2025.
[4] H. Ahmed, *et al.*, "Hybrid Deep-Learning and Explainable AI for Anomaly-Based Intrusion Detection," *IEEE Access*, 2022.
[5] M. R. Alam, *et al.*, "Explainable Deep Intrusion Detection Model Using SHAP Values and Feature Attribution," *Sensors*, 2022.
[6] A. Bhattacharya and P. Das, "LIME-based Explainable Network Intrusion Detection for IoT Devices," *Computers & Security*, 2023.
[7] K. S. Kim, *et al.*, "SHAP-based Interpretation of Network Anomalies Using Ensemble Learning," *Applied Intelligence*, 2023.
[8] T. Ban, *et al.*, "Breaking Alert Fatigue: AI-Assisted SIEM Framework Using Explainable SVM," *Applied Sciences*, 2023.
[9] L. Zhou, *et al.*, "Interpretable Cyber Threat Detection Using SHAP and Graph Neural Networks," *Expert Systems with Applications*, 2023.
[10] P. Gaj, B. B. Akgu¨n, and S. Ustebay, "Securing SCADA Systems: IDS with Shapley Analysis," *Protocol-based IDS*, 2024.
[11] M. Wang, K. Zheng, Y. Yang, and X. Wang, "Explainable ML Framework for IDS," *IEEE Access*, 2020.
[12] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?" Explaining Any Classifier," *KDD Conference*, 2016.
[13] H. Khan, *et al.*, "Explainable Ensemble IDS Based on Feature Attribution," *IEEE Access*, 2024.
[14] L. Coroama and A. Groza, "Evaluation Metrics for XAI Techniques," *CCIS Chapter*, 2022.
[15] S. Kim and Y. Cho, "Benchmarking Explainable AI Models for Intrusion Detection: A Comparative Study," *IEEE Access*, 2025.
[16] B. Guvenc Paltun and R. Fuladi, "Introspective IDS Through Explainable AI," *Adversarial Defence + Real-time XAI IDS*, 2024.
[17] D. Kumar and A. Bansal, "XAI-Augmented IDS for Cloud Environments: An Interpretable Deep Learning Framework," *IEEE Trans. Cloud Comp.*, 2024.
[18] P. Gupta, "Interpretable Feature Attribution for Federated IDS Models," *IEEE Trans. Net. & Ser. Manag.*, 2025.
[19] J. Ryu, *et al.*, "Attention-based Explainable Intrusion Detection Using Deep Neural Networks," *Neural Comp. and Appl.*, 2024.
[20] N. Rastogi, *et al.*, "Measuring Security and Cognitive Impact in AI- Driven SOCs," *ACM CCS*, 2025.
[21] P. Hermosilla, *et al.*, "SHAP vs. LIME for Forensic IDS Analysis," *Applied Sciences*, 2025.
[22] J. Yadav and P. Sinha, "SHAP and LIME for Transparent Intrusion Detection in IoT Environments," *Internet of Things*, 2024.
[23] S. R. Islam, *et al.*, "Domain Knowledge Aided Explainable AI for IDS," *AAAI Symposium*, 2020.

[24] T. Zebin, *et al.*, "XAI for DNS over HTTPS Attack Detection," *IEEE TIFS*, 2022.
[25] S. Patil, *et al.*, "Explainable AI for Intrusion Detection System," *Electronics*, 2022.
[26] M. Tajuddin and A. N. Harigol, "Intrusion Detection of Imbalanced Network Traffic Based on Machine Learning and Deep Learning Techniques," *International Journal of Innovative Research in Computer and Communication Engineering*, 2022.
[27] M. Tajuddin and M. T. P. Manasa, "Cancer Detection and Classification Using Random Forest, Neural Network, and XGBoost Algorithms of Machine Learning," *Journal of Information Systems Engineering and Management*, 2024.
[28] C. V. Balaji, A. Bharadwaj, A. Denny, A. M. Pamadi, S. Christa, and S. Padmavathi, "A Survey of Adversarial Attack and Defense Methods for Deep Learning Models," *Data Science Exploration in Artificial Intelligence*, 2025.
[29] V. S. Vagdevi and V. S. Deepthi, "RCO-Based Multilayer IDS in MANETs with Collective Fluctuation Measure," *Design Engineering (Toronto)*, 2021.
[30] J. M. Avinash, P. B. Bhat, B. A. Shri Prasada, and V. S. Deepthi, "Review on Detection of Phishing Attacks Using Machine Learning Algorithms," *Journal of Emerging Technologies and Innovative Research*, 2021.
[31] V. S. Deepthi and V. S. Vagdevi, "Behaviour Analysis and Detection of Blackhole Attacker Node under Reactive Routing Protocol in MANETs," in *Proc. 2018 International Conference on Networking, Embedded and Wireless Systems (ICNEWS)*, 2018.
[32] V. S. Deepthi and S. Vagdevi, "Multiphase Detection and Evaluation of AODV for Malicious Behaviour of a Node in MANETs," in *Proc. 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, 2018.