

Explanation Trust Score for Reliable LLM Incident Reporting in IoMT

Komal Ghafoor^{*†}, Furqan Rustam[†], Anca Delia Jurcut[†], Devis Bianchini^{*}

^{*}Department of Information Engineering, Università degli Studi di Brescia (Italy),

[†]School of Computer Science, University College Dublin (Ireland)

{komal.ghafoor, devis.Bianchini}@unibs.it, furqan.rustam@ucdconnect.ie, anca.jurcut@ucd.ie

Abstract—Explainable AI (XAI) techniques such as SHAP and LIME are increasingly used to justify intrusion detection decisions before being handed to downstream reporting systems, including Large Language Models (LLMs) that automatically generate incident narratives. Existing pipelines assume these post-hoc explanations are inherently reliable without verifying them first, a risk that is especially dangerous in Internet of Medical Things (IoMT) settings, where a misleading report can delay the response to a patient-safety threat. We propose a safety-verified alert-generation framework built around an Explanation Trust Score (ETS) that combines Prediction Confidence, Explanation Agreement (SHAP-LIME Jaccard similarity), Explanation Stability, and Prediction Robustness to determine whether an explanation is reliable enough for automatic reporting, cautious (hedged) reporting, or human review. Evaluated on labeled IoMT network-traffic instances using an XGBoost detection agent and an explicit IP-leakage diagnostic, the proposed framework shows that the detector is uniformly confident and robust (PC and PR ≈ 1.0), yet SHAP and LIME frequently disagree on the driving features (mean Jaccard agreement of 0.426 on a 300-instance evaluation subsample), confirming that prediction and explanation-side reliability are distinct signals. Applying the proposed routing policy, 69.0% of instances are classified as *Trusted Report* and 31.0% as *Cautious Report*, with no instances requiring *Human Review*.

Index Terms—Explainable AI, IoMT, Intrusion Detection, Machine Learning, Attack Detection

I. INTRODUCTION

The rapid expansion of the Internet of Medical Things (IoMT), including networked patient monitors, environmental sensors, and MQTT-based clinical telemetry devices, has created a safety-critical attack surface within healthcare infrastructure [1], [2]. Unlike conventional IT systems, delayed or incorrect responses to IoMT attacks can disrupt clinical operations and directly endanger patient safety. Consequently, IoMT intrusion detection must provide not only accurate predictions but also trustworthy evidence that analysts can interpret and validate [3]. Machine learning-based Intrusion Detection Systems (IDSs) have demonstrated strong performance in identifying malicious network traffic [4]. Existing studies have investigated ensemble learning, XGBoost, and deep reinforcement learning models for IoT and IoMT security [5]–[8]. However, high predictive performance does not necessarily guarantee that a model has learned generalizable attack behavior, particularly when benchmark datasets contain duplicated patterns, identifier leakage, or capture-specific artifacts [9].

Explainable Artificial Intelligence (XAI) refers to methods that make the predictions of complex machine-learning models more understandable to human users by identifying the factors that contribute to a particular decision. XAI has therefore become important for interpreting the decisions of black-box IDS models. SHapley Additive ExPlanations (SHAP) and Local Interpretable Model-Agnostic Explanations (LIME) are widely used to identify the network features influencing individual predictions [10]. Previous IoMT studies have applied both methods to explain intrusion-detection outcomes, but their results are typically presented independently rather than evaluated through a unified reliability measure [7]. In this context, an explanation refers to the set of feature-level attributions used to indicate which input characteristics contributed most strongly to a model prediction. Producing an explanation, however, does not establish that the explanation is trustworthy. Explanations may vary under small input changes, disagree across explanation methods, or remain vulnerable to manipulation even when the underlying prediction is confident [11]. Existing research has highlighted the need to assess explanation quality, robustness, and statistical reliability [12], [13], but these assessments are generally used for retrospective model evaluation rather than as operational controls for automated incident reporting.

This limitation becomes more serious when XAI outputs are supplied to Large Language Models (LLMs), which are generative AI models capable of processing and producing human-like text based on the information provided to them. LLMs are increasingly used in Security Operations Centres for alert summarization, incident explanation, and natural-language report generation [14]. Existing reporting pipelines may provide SHAP explanations directly to an LLM without first verifying their agreement, stability, or reliability [15]. As a result, an unreliable explanation may be converted into a fluent and apparently authoritative report that misleads analysts during safety-critical incident response. The central research gap is therefore the absence of an operational mechanism for determining whether an explanation is sufficiently reliable for automated LLM-based reporting. To address this gap, this paper introduces an Explanation Trust Score (ETS) that combines prediction confidence, SHAP-LIME agreement, explanation stability, and prediction robustness. The resulting score routes each incident to a trusted report, a cautious report, or human analyst review, thereby introducing an explanation-aware trust

layer between intrusion detection and downstream incident reporting. The main contributions of this work are summarized as follows:

- We introduce an ETS that jointly evaluates prediction confidence, SHAP-LIME agreement, explanation stability, and prediction robustness to quantify the reliability of individual intrusion explanations.
- We propose a three-tier trust-based routing policy that assigns detected incidents to trusted reporting, cautious reporting, or human analyst review according to ETS.
- We validate the proposed framework on IoMT network-traffic instances, conduct an explicit identifier-leakage analysis, and demonstrate that highly confident and robust predictions may still exhibit substantial disagreement between explanation methods.

The remainder of this paper is organized as follows. Section II reviews the related work. Section III presents the proposed methodology. Section IV reports the experimental results. Section V discusses the findings and limitations. Finally, Section VI concludes the paper.

II. RELATED WORK

Beyond the background presented in Section I, a broader body of recent work bears directly on three aspects of this study: (i) explainable intrusion detection for IoMT networks, (ii) the quality and leakage-proneness of the benchmark datasets on which such methods are evaluated, and (iii) the emerging use of LLMs for security reporting and Security Operations Center (SOC) automation.

Explainable Intrusion Detection for IoMT Networks. Mohale and Obagbuwa [5] integrate XAI into an ML-based IDS evaluated on the UNSW-NB15 dataset, showing that interpretability can be added without materially sacrificing detection accuracy. However, like much of the existing literature, their evaluation treats explanation quality as a reporting feature rather than a gating signal. Focusing specifically on the IoMT domain, Alsolami *et al.* [6] benchmark Stacking, Bagging, and Boosting ensembles on the WUSTL-EHMS-2020 healthcare IDS dataset, finding that Stacking substantially outperforms individual base learners. This finding motivates our use of a strong tree-ensemble model as the detection agent.

Hosain and Çakmak [7] propose an XGBoost-based IDS for IoMT that, similar to our pipeline, applies both SHAP and LIME for global and local interpretability on the WUSTL-EHMS-2020 dataset. However, their SHAP and LIME outputs are reported independently as complementary evidence rather than being combined into a single quantitative agreement or trust metric, which is the gap addressed by ETS. Shaikh *et al.* [8] pursue a deep reinforcement learning architecture Deep Q-Network, Proximal Policy Optimization (CNN-LSTM with DQN/PPO) for IoMT intrusion detection on the CICIoMT2024 dataset, illustrating that detection performance continues to advance rapidly, while explanation reliability remains comparatively underexplored. At a broader methodological level, Khan *et al.* [11] systematically review XAI-based IDSs for Industry 5.0 environments together with adversarial attacks

against the explanations themselves, underscoring that an explanation’s trustworthiness cannot be assumed even when the underlying model performs well. This observation motivates treating explanation reliability as a first-class, measurable quantity rather than an afterthought.

Ogunseyi *et al.* [12] extend this discussion through a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guided systematic review of 129 studies on explainable deep learning IDSs for IoT. They propose a Unified Explainable IDS Evaluation Framework that models a three-way trade-off between detection performance, computational efficiency, and explanation quality. Although their framework considers multiple dimensions of explanation quality, it is intended as a post-hoc evaluation framework rather than an operational gate for downstream reporting. Finally, Zarkadis and Douligeris [13] apply SHAP together with rigorous statistical hypothesis testing (Westfall-Young permutation tests and Jensen-Shannon distance) to a tree-ensemble IDS for UAV networks. Their work demonstrates that statistically grounded validation of explanation-driven conclusions is both feasible and necessary, a methodological standard that our perturbation-restricted ES/PR computation seeks to achieve in the IoMT setting.

Dataset Quality and Leakage in Intrusion Detection Benchmarks. A parallel line of work has scrutinized the benchmark datasets underlying claims of near-perfect IDS performance. Goldschmidt and Chudá [9] survey network intrusion datasets and catalogue recurring design flaws, including feature leakage, ambiguous ground truth, and unrealistic traffic diversity that can silently inflate reported detection performance. Their findings directly motivate our explicit IP-identifier leakage analysis in **Section III-B** rather than accepting near-ceiling accuracy at face value.

LLM-Based Security Reporting and SOC Automation. On the reporting side, Habibzadeh *et al.* [14] provide a comprehensive survey of LLM integration into Security Operations Centers, covering applications such as alert triage, log summarization, and natural-language incident explanation. They also identify alert fatigue and analyst trust as major unresolved challenges. An explanation-aware gating mechanism such as ETS is directly aimed at addressing these challenges by ensuring that automatically generated reports are presented with full confidence only after the underlying explanation has been verified as reliable. Taken together, these studies show that explainable IoMT intrusion detection, dataset-quality auditing, and LLM-based security reporting are each rapidly maturing research areas. However, consistent with the research gap identified in the literature, none of the reviewed studies combine explanation agreement, stability, robustness, and prediction confidence into a single actionable score that determines how an LLM should report an intrusion detection decision. Table I summarizes the main characteristics of the reviewed studies and compares them with the proposed framework.

TABLE I
COMPARISON OF RELATED WORK WITH THE PROPOSED FRAMEWORK

Ref	Year	XAI	IoMT	LLM	LA	ETS	DP
[6]	2024	✗	✓	✗	✗	✗	✗
[5]	2025	✓	✗	✗	✗	✗	✗
[7]	2025	✓	✓	✗	✗	✗	✗
[8]	2025	✗	✓	✗	✗	✗	✗
[9]	2025	✗	✗	✗	✓	✗	✗
[11]	2025	✓	✗	✗	✗	✗	✗
[12]	2026	✓	✗	✗	✗	✗	✗
[13]	2026	✗	✗	✗	✗	✗	✗
[14]	2026	✓	✗	✓	✗	✗	✗
This Work	✓	✓	✓	✓	✓	✓	✓

XAI: Explainable AI, IoMT: IoMT-specific IDS, LLM: LLM-oriented reporting, LA: Leakage Analysis, ETS: Explanation Trust Score, DP: Decision Policy.

III. PROPOSED METHODOLOGY

The proposed framework introduces an explanation-aware trust layer between the intrusion detection model and downstream LLM-assisted incident reporting. Rather than forwarding SHAP and LIME explanations directly to an LLM, the framework first evaluates the reliability of each prediction and its associated explanations using the proposed ETS. The ETS combines four complementary components: Prediction Confidence (PC), Explanation Agreement (EA), Explanation Stability (ES), and Prediction Robustness (PR). Based on the resulting ETS value, each detected incident is assigned to one of three handling tiers: *Trusted Report*, *Cautious Report*, or *Human Review*. High-trust instances are eligible for standard automated reporting, intermediate-trust instances are assigned to cautious reporting, and low-trust instances are deferred to a human analyst. The overall workflow is illustrated in Fig. 1.

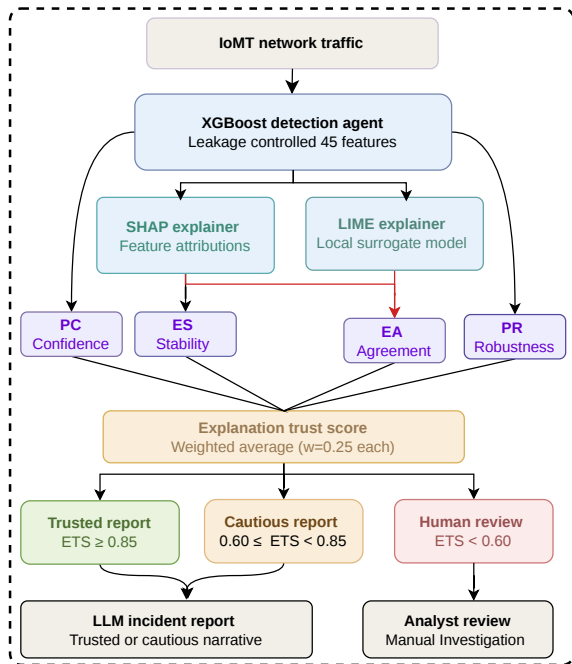


Fig. 1. Explanation-aware trust-gated pipeline for IoMT incident reporting.

A. Dataset Construction and Preprocessing. The proposed framework is evaluated using network traffic derived from the IoT-Flock framework.¹ Three CSV files are used to construct the experimental dataset: `Attack.csv`, containing malicious traffic, and `environmentMonitoring.csv` and `patientMonitoring.csv`, containing benign traffic. After duplicate removal, the three files are concatenated and randomly shuffled using a fixed random seed (`seed=42`). The resulting dataset contains 188,694 labeled packet-level instances, comprising 108,568 normal instances (57.54%) and 80,126 attack instances (42.46%), with 50 features extracted from the frame, TCP, IP, and MQTT protocol headers. Categorical protocol fields, including `tcp.flags`, `mqtt.hdrflags`, and `mqtt.conflags`, are ordinal-encoded, while missing values are imputed with zero. The dataset is then divided using stratified sampling into 80% training and 20% testing subsets. This produces 150,955 training instances and 37,739 testing instances.

B. Data-Leakage Diagnostic. Because the attack and benign samples originate from different traffic-generation scenarios, network identifiers may encode class-specific information rather than generalizable attack behavior. We therefore examine whether the IP-address fields introduce identifier-based data leakage before training the detection models. Specifically, we compare the sets of unique source and destination IP addresses, defined as $ip.src \cup ip.dst$, across the attack and benign classes. The attack class contains four unique IP addresses, whereas the benign class contains forty, with no overlap between the two sets. Consequently, the IP-address fields can act as direct proxies for the class label and may allow the classifier to distinguish attack traffic from benign traffic without learning generalizable protocol-level behavior.

To evaluate the influence of identifier-related features, two models are trained using identical hyperparameters and train-test partitions. The full-feature model retains all 50 features. The leakage-controlled model removes the five identifier-related fields `ip.src`, `ip.dst`, `tcp.srcport`, `tcp.dstport`, and `mqtt.clientid`, and is trained using the remaining 45 features. The performance of both models is compared, while the leakage-controlled model is used for the subsequent explanation and trust-scoring analyses.

C. XGBoost Detection Agent. An XGBoost classifier serves as the intrusion detection agent. The model is configured with `n_estimators = 300`, `max_depth = 6`, `learning_rate = 0.08`, `subsample = 0.8`, and `colsample_bytree = 0.8`. Two XGBoost variants were trained following the feature configurations defined in Section III-B. Since identifier fields can allow the classifier to memorize device or session-specific identities rather than learn generalizable traffic patterns, the leakage-controlled model is adopted as the primary detection agent for all subsequent explanation and trust-scoring analysis in this paper.

D. SHAP Explanation Agent. Feature attributions are computed using SHAP's `TreeExplainer`, which provides

¹www.kaggle.com/datasets/faisalmaalik/iot-healthcare-security-dataset

exact and computationally efficient Shapley-value estimation for tree-based ensemble models. SHAP values are computed for the entire held-out test set (37,739 instances). Global feature importance is then obtained from the mean absolute SHAP value across all test samples, computed for feature j as

$$I_j = \frac{1}{N} \sum_{i=1}^N |\phi_{ij}|, \quad (1)$$

where N is the number of test instances and ϕ_{ij} is the SHAP value of feature j for instance i . Features with larger mean absolute SHAP values have a greater overall influence on the model's predictions across the evaluation dataset.

E. LIME Explanation Agent. LIME is applied using `LimeTabularExplainer` with continuous-feature discretization enabled. Since LIME fits an independent local surrogate per instance and is therefore markedly more expensive than SHAP's vectorized `TreeExplainer`, explanations are generated for a random subsample of $N = 300$ test instances, drawn without replacement using the same fixed seed as the rest of the pipeline. This 300-instance subsample is reused for all subsequent ETS computations (EA, ES, PR, PC), ensuring the final trust score is never computed by mixing statistics drawn from differently sized samples.

F. Explanation Trust Score (ETS). For each of the 300 aligned test instances, four components are computed: PC, EA, ES, and PR, measuring prediction confidence, SHAP-LIME consistency, explanation stability under perturbations, and prediction consistency under perturbations, respectively. The four components are formally defined as follows.

$$PC_i = \max_{c \in \mathcal{C}} P(y = c \mid x_i), \quad (2)$$

where $\mathcal{C}$ denotes the set of possible classes and $P(y = c \mid x_i)$ is the probability assigned by the XGBoost detector to class c for instance i . Thus, PC_i represents the confidence of the detector in its predicted class, with higher values indicating greater prediction confidence.

$$EA_i = \frac{|S_i \cap L_i|}{|S_i \cup L_i|}, \quad (3)$$

where S_i and L_i denote the top-10 feature sets identified by SHAP and LIME, respectively, for instance i . ES evaluates whether SHAP explanations remain consistent under small input perturbations. Gaussian noise ($\sigma = 0.01$) is injected into selected numerical protocol features (packet timing, packet length, TCP window size, keep-alive interval, and TTL), while ordinal-encoded categorical fields remain unchanged to avoid artificial category transitions. SHAP stability is measured over five perturbation trials using cosine similarity between the original and perturbed attribution vectors,

$$ES_i = \frac{\phi_i \cdot \phi'_i}{\|\phi_i\| \|\phi'_i\|}, \quad (4)$$

where ϕ_i and ϕ'_i are the original and perturbed SHAP attribution vectors, respectively. The final ES score is the

average cosine similarity across the five perturbation trials. PR measures whether the predicted class remains unchanged under the same perturbations used for ES,

$$PR_i = \frac{1}{K} \sum_{k=1}^K \mathbf{1}(\hat{y}_i = \hat{y}_i^{(k)}), \quad (5)$$

where $K = 5$, $\hat{y}_i$ is the original predicted class, $\hat{y}_i^{(k)}$ is the prediction after the k -th perturbation, and $\mathbf{1}(\cdot)$ returns 1 when the prediction remains unchanged and 0 otherwise. The overall Explanation Trust Score for each instance is computed as

$$ETS_i = w_1 EA_i + w_2 ES_i + w_3 PR_i + w_4 PC_i, \quad (6)$$

where $w_1 = w_2 = w_3 = w_4 = 0.25$.

G. Decision Policy. The ETS value from Eq. (6) assigns each test instance to one of three handling tiers. A trusted report is generated automatically when $ETS_i \geq 0.85$, a cautious report is produced when $0.60 \leq ETS_i < 0.85$, and the instance is referred for human review when $ETS_i < 0.60$. The thresholds were selected empirically to distinguish high, medium, and low-trust predictions while maintaining clear separation between automated reporting and human review.

IV. EXPERIMENTAL RESULTS

All experiments were conducted in Google Colab using a Python-based implementation of the proposed framework. The experimental workflow included data preprocessing, XGBoost training, SHAP and LIME explanation generation, perturbation analysis, and ETS computation.

A. Detection Performance. The leakage-controlled XGBoost detection agent achieved an accuracy, precision, recall, F1-score, and AUC of 1.000 on the held-out test set of 37,739 instances. The corresponding confusion matrix contained no false positives or false negatives, correctly classifying all 21,714 normal and 16,025 attack instances. The full-feature model achieved identical performance. Thus, under the current experimental split, removing the five identifier-related features, `ip.src`, `ip.dst`, `tcp.srcport`, `tcp.dstport`, and `mqtt.clientid`, did not change the observed detection performance. The leakage-controlled model was therefore adopted for the subsequent explanation and trust-scoring analyses in order to reduce reliance on direct identifier information.

B. Global SHAP Feature Importance. Global feature importance was computed as the mean absolute SHAP value over all 37,739 instances in the held-out test set. Table II reports the five highest-ranked features, showing that TCP header length, packet timing, checksum information, and MQTT header flags had the greatest influence on the model's predictions.

As shown in Fig. 2, `tcp.hdr_len` and `frame.time_relative` exhibit the widest SHAP-value distributions and therefore have the strongest influence on the model predictions. From a network-traffic perspective, the importance of `tcp.hdr_len` may indicate that differences in TCP header structure or the use of optional header fields are associated with distinct traffic-generation patterns

TABLE II
TOP-5 GLOBAL FEATURES BY MEAN ABSOLUTE SHAP VALUE

Rank	Feature	Mean SHAP
1	tcp.hdr_len	4.106
2	frame.time_relative	2.673
3	tcp.checksum	1.449
4	mqtt.hdrflags	1.118
5	tcp.time_delta	1.020

between the normal and attack classes. Similarly, the high importance of `frame.time_relative` suggests that temporal characteristics, such as differences in packet arrival patterns over the duration of a traffic capture, contribute substantially to the model’s discrimination between classes.

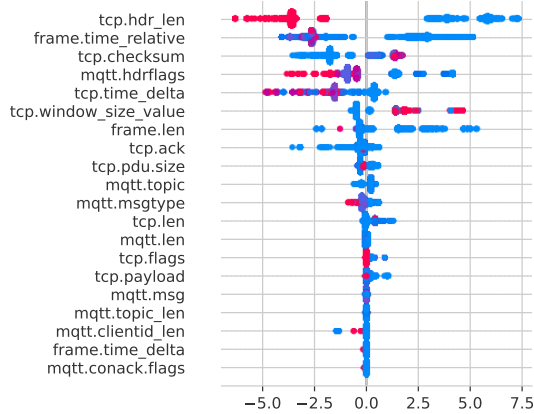


Fig. 2. Global SHAP summary plot for the leakage-controlled XGBoost.

The importance of `tcp.checksum` may reflect systematic differences in packet construction or processing across the traffic-generation scenarios, rather than indicating that the checksum itself is inherently an indicator of malicious activity. Likewise, `mqtt.hdrflags` may capture differences in the use of MQTT control-message flags between normal and attack traffic, while `tcp.time_delta` may reflect variations in packet inter-arrival timing and communication rates. These observations provide a plausible interpretation of the model’s behaviour; however, SHAP importance indicates the contribution of a feature to the model’s predictions and should not be interpreted as establishing a causal relationship with malicious activity. The remaining features have SHAP values concentrated closer to zero and therefore exert comparatively limited influence on the model predictions.

C. LIME Feature Frequency. Figure 3 shows how frequently each feature appears among the top-10 LIME attributions across the 300-instance evaluation subsample. The resulting frequency ranking provides a qualitative comparison with the global SHAP ranking reported in Table II before the instance-level SHAP-LIME agreement analysis.

D. Explanation Trust Score Components. The four ETS components were evaluated over the aligned 300-instance subsample. PC and PR remained near 1.0, while ES was

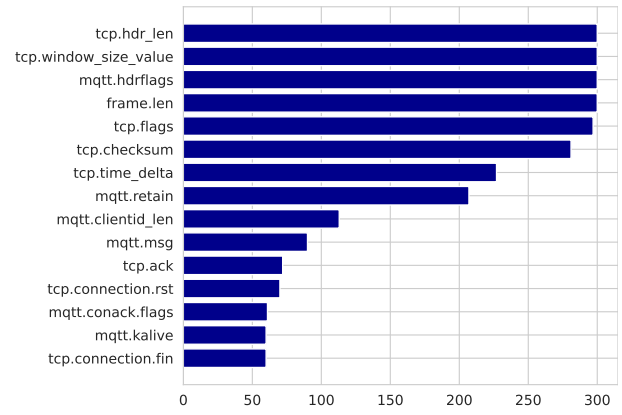


Fig. 3. Top-10 LIME feature frequency across 300 test instances.

also high, indicating confident predictions that remained stable under the applied perturbations, as summarized in Table III. In contrast, EA had a mean of 0.4263, indicating only moderate agreement between SHAP and LIME. As illustrated in Fig. 4, most agreement scores are concentrated around 0.43, although substantial instance-level variation remains. These results demonstrate that prediction confidence and robustness alone do not capture explanation reliability.

TABLE III
ETS COMPONENT STATISTICS ($n = 300$)

Statistic	PC	EA	ES	PR	ETS
Mean	0.9999	0.4263	0.9861	1.0000	0.8531
Median	1.0000	0.4286	0.9968	1.0000	0.8564
Std. Dev.	0.0008	0.0708	0.0431	0.0000	0.0214
Minimum	0.9869	0.2500	0.6451	1.0000	0.7678
Maximum	1.0000	0.6667	1.0000	1.0000	0.9167

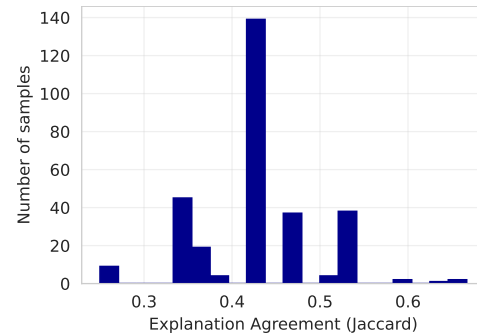


Fig. 4. SHAP-LIME agreement across 300 test instances.

E. Decision Routing Outcomes. Applying the decision policy yields the routing distribution. Of the 300 evaluated instances, 207 (69.0%) were assigned to *Trusted Report* and 93 (31.0%) to *Cautious Report*; no instance required *Human Review*. Figure 5 shows the ETS distribution relative to the trusted-report threshold of 0.85 and the human-review threshold of 0.60. All scores remained above 0.60, while instances below 0.85 were routed to cautious reporting, primarily re-

flecting lower SHAP-LIME agreement despite high prediction confidence and robustness.

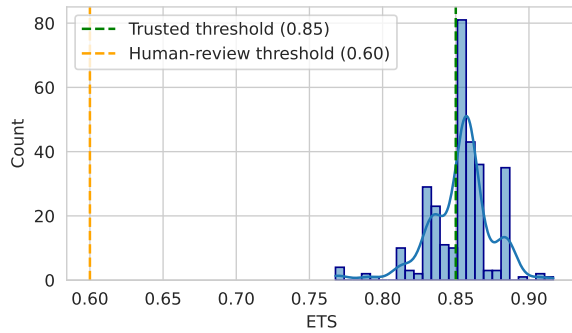


Fig. 5. ETS distribution with decision thresholds.

V. DISCUSSION & LIMITATIONS

The leakage-controlled XGBoost detector achieved perfect performance on the current held-out test set, showing that the remaining protocol-level features retained sufficient discriminatory information after removing the selected identifier fields. However, this result should be interpreted cautiously because a random packet-level split may preserve similarities between training and test samples. More importantly, the ETS results reveal a clear distinction between prediction-side and explanation-side reliability. Although PC and PR were near 1.0, SHAP and LIME showed only moderate agreement (mean EA = 0.4263). Consequently, 69.0% of instances were routed to Trusted Report and 31.0% to Cautious Report, demonstrating that confident predictions do not necessarily produce consistent explanations.

Several limitations remain. PC measures model confidence but does not assess probability calibration, while PR evaluates class consistency under only five small perturbations and may therefore overestimate robustness. EA relies on top-10 Jaccard overlap and ignores attribution magnitude, ranking, and direction; moreover, disagreement between SHAP and LIME does not necessarily imply that either explanation is incorrect. ES evaluates only SHAP stability and depends on a fixed perturbation magnitude whose suitability may vary across features. At the aggregate level, ETS uses equal component weights and empirically selected thresholds without calibration or sensitivity analysis. The evaluation is also limited to 300 instances, a single dataset, and a random train-test split, and the downstream LLM reporting stage is not implemented or assessed. Further, no evaluated instance fell below the 0.60 threshold; therefore, the Human Review branch was not empirically exercised in the current evaluation.

VI. CONCLUSION & FUTURE WORK

This paper introduced the ETS, which combines PC, SHAP-LIME agreement, ES, and PR to assess explanation reliability in an IoMT intrusion detection pipeline. Using a leakage-controlled XGBoost model, the results showed that highly

confident predictions that remained stable under the applied perturbations can still exhibit only moderate explanation agreement, with a mean Jaccard score of 0.426. The ETS-based policy routed 69.0% of instances to Trusted Report and 31.0% to Cautious Report, demonstrating the value of separating prediction reliability from explanation reliability. Future work will focus on device and session-held-out evaluation, probability calibration, feature-aware perturbations, data-driven ETS weighting, threshold sensitivity, and end-to-end evaluation of LLM-assisted report quality across the decision tiers.

REFERENCES

- [1] M. M. Ozcelik, I. Kok, and S. Ozdemir, "A survey on internet of medical things (iomt): Enabling technologies, security and explainability issues, challenges, and future directions," *Expert Systems*, vol. 42, no. 5, p. e70010, 2025.
- [2] A. Ghubaish, T. Salman, M. Zolanvari, D. Unal, A. Al-Ali, and R. Jain, "Recent advances in the internet-of-medical-things (iomt) systems security," *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8707–8718, 2020.
- [3] J. Doménech, I. V. Martin-Faus, S. Mhiri, and J. Pegueroles, "Ensuring patient safety in iomt: A systematic literature review of behavior-based intrusion detection systems," *Internet of Things*, vol. 28, p. 101420, 2024.
- [4] A. H. Ali, M. Charfeddine, B. Ammar, B. B. Hamed, F. Albalwy, A. Alqarafi, and A. Hussain, "Unveiling machine learning strategies and considerations in intrusion detection systems: a comprehensive survey," *Frontiers in Computer Science*, vol. 6, p. 1387354, 2024.
- [5] V. Z. Mohale and I. C. Obagbuwa, "Evaluating machine learning-based intrusion detection systems with explainable ai: enhancing transparency and interpretability," *Frontiers in Computer Science*, vol. 7, p. 1520741, 2025.
- [6] T. Alsolami, B. Alsharif, and M. Ilyas, "Enhancing cybersecurity in healthcare: Evaluating ensemble learning models for intrusion detection in the internet of medical things," *Sensors*, vol. 24, no. 18, p. 5937, 2024.
- [7] Y. Hosain and M. Çakmak, "Xai-xgboost: an innovative explainable intrusion detection approach for securing internet of medical things systems," *Scientific Reports*, vol. 15, no. 1, p. 22278, 2025.
- [8] J. A. Shaikh, C. Wang, M. W. U. Sima, M. Arshad, M. Owais, D. S. Hassan, R. Alkanhel, and M. S. A. Muthanna, "A deep reinforcement learning-based robust intrusion detection system for securing iomt healthcare networks," *Frontiers in Medicine*, vol. 12, p. 1524286, 2025.
- [9] P. Goldschmidt and D. Chudá, "Network intrusion datasets: A survey, limitations, and recommendations," *Computers & Security*, vol. 156, p. 104510, 2025.
- [10] A. M. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva, S. E. Petersen, K. Lekadir, and G. Menegaz, "A perspective on explainable artificial intelligence methods: Shap and lime," *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400304, 2025.
- [11] N. Khan, K. Ahmad, A. Al Tamimi, M. M. Alani, A. Bermak, and I. Khalil, "Explainable ai-based intrusion detection systems for industry 5.0 and adversarial xai: A systematic review," *Information*, vol. 16, no. 12, p. 1036, 2025.
- [12] T. B. Ogunseyi, G. Thiyagarajan, H. He, V. Bist, and Z. Du, "Performance analysis of explainable deep learning-based intrusion detection systems for iot networks: A systematic review," *Sensors*, vol. 26, no. 2, p. 363, 2026.
- [13] I.-C. Zarkadis and C. Douligeris, "Xai and statistical analysis for reliable intrusion detection in the uavids-2025 dataset: From tree to hybrid and tabular dnn ensembles," *arXiv preprint arXiv:2605.13922*, 2026.
- [14] A. Habibzadeh, F. Feyzi, and R. E. Atani, "Large language models for security operations centers: A comprehensive survey," *Journal of Electrical and Computer Engineering*, vol. 2026, no. 1, p. 3383674, 2026.
- [15] C.-H. Chou, A. Sudheer, and Y. Park, "An llm-based agentic network traffic incident-report approach towards explainable-ai network defense," *Journal of Sensor and Actuator Networks*, vol. 15, no. 2, p. 32, 2026.