HiFi-XAI: A Fidelity-Aware, LLM-Powered Framework for Trustworthy Intrusion Detection

Avinash Awasthi*, Pritam Vediya*, Hemant Miranka†, Ramesh Babu Battula*, Priyadarsi Nanda‡
*Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur 302017, India.
†The LNM Institute of Information Technology, Jaipur 302031, India.
‡University of Technology Sydney, Sydney, Australia.

***Abstract*—The increasing deployment of complex "black box" AI models in anomaly-based Intrusion Detection Systems (IDS) for future networks has created a critical trust gap, making human-interpretable explanations essential for cyber security analysts to act on alerts confidently. However, existing Explainable AI (XAI) methods like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) fail in this domain. They suffer from a severe fidelity problem by incorrectly assuming feature independence, which leads to untrustworthy explanations for inherently correlated network data for IDS. We propose HiFi-XAI, a novel framework that generates faithful and semantically rich explanations to address these challenges. HiFi-XAI introduces a model-agnostic Conditional Value Attribution Explanation (CVAE), a method based on probabilistic shapley values that models feature dependencies to ensure explanations are derived from plausible data distributions. These high-fidelity attributions are then translated into actionable, natural-language narratives by a fine-tuned Large Language Model (LLM). We validate our framework through all-aware scenario feature ablation studies on the CICIDS2017 and CICIOT2023 datasets. This demonstrates that CVAE consistently identifies more impactful features than SHAP and LIME across five anomaly-based IDS models. Furthermore, we deploy the HiFi-XAI to prove its practical feasibility and test it on a resource-constrained Raspberry Pi 4. Our work presents a complete, end-to-end solution for building trust in AI-driven IDS.**

***Index Terms*—Explainable AI, Intrusion Detection System, Network Security, Shapley Values, Large Language Models, Fidelity.**

I. Introduction

Network security development is a perpetual, adaptive challenge driven by increasingly sophisticated new threat vectors . To maintain operational integrity, the Intrusion Detection System (*IDS*) is the primary mechanism for identifying and keeping up with zero-day threats [1]. The signature-based IDS was limited in generalizing threat vectors, a constraint that mandated the transition to AI-driven solutions. Consequently, the field has transitioned to AI-driven solutions, leveraging Deep Learning (DL), Transformers, and Federated Learning to autonomously detect novel attacks across complex infrastructures like $6G$ networks and the Internet of Autonomous Things (*IoAT*).

While these advanced models excel at identifying subtle malicious patterns, their power comes at the cost of transparency. The internal complexity of Deep Neural Networks (*DNNs*) and Transformers creates opaque, "black box" models [2], posing a significant trust problem: a security analyst cannot confidently act on an alert without understanding the model's reasoning. This has spurred the rise of Explainable AI (*XAI*), which aims to bridge the gap between predictive accuracy and human interpretability [3].

Current XAI applications in cybersecurity heavily rely on post-hoc, model-agnostic techniques like *SHAP* (Shapley Additive exPlanations) [4] and *LIME* (Local Interpretable Model-agnostic Explanations) [5], [6]. These tools generate feature importance scores by simulating a feature's absence—often by replacing its value with random samples from the dataset. This process, however, is built on a critical and flawed premise: the *feature independence assumption*.

This assumption is fundamentally incorrect in network security, where traffic data is inherently Non-Independent and Identically Distributed (*Non-IID*) [7]. Features such as packet size, protocol, and flow duration are highly correlated [8]. By ignoring these dependencies, *SHAP* and *LIME* evaluate models on unrealistic, *out-of-distribution* (*OOD*) data points, producing low-fidelity explanations that misrepresent the model's true logic. This creates two distinct challenges:

1) *A Fidelity Deficit:* The explanations are not a faithful representation of the model's decision-making process.
2) *A Semantic Gap:* A list of feature importances is not an operational explanation. Analysts require a human-readable narrative that contextualizes an attack [9].

To address these dual challenges, we propose *HiFi-XAI*, a novel, fidelity-aware framework for generating trustworthy and semantically rich explanations for *IDS*. Our key contributions are summarized as follows:

- **Contribution 1**: We solve the fidelity problem by introducing *Conditional Value Attribution Explanation (CVAE)*, a method grounded in *Shapley values* that replaces the flawed independence assumption with a learned probabilistic model of feature dependencies, ensuring explanations are generated only from plausible data.
- **Contribution 2:** We solve the semantic interpretation problem by integrating *CVAE* with a fine-tuned *Large Language Model* (*LLM*) that translates high-fidelity mathematical feature weights into actionable, natural language insights for security operators.
- **Contribution 3:** We demonstrate the practical feasibility

of our approach by deploying and testing *CVAE* on a resource-constrained *IDS* testbed, making high-fidelity XAI practical for edge environments like $6G$ and *IoAT*.

II. Background and Related Work

The extensive investigations into explainable intrusion detection have produced a rich body of literature. However, a systematic review reveals critical research gaps that precisely motivate the core contributions of this work.

A. Standard Application of XAI to Advanced IDS Architectures

The most common approach involves the direct application of *SHAP* and *LIME* to complex *IDS* models in environments ranging from *IoT* and federated learning to *UAVs* and *IoMT* [10]–[15]. Even research focused on adversarial resilience or model efficiency for edge devices defaults to using these standard tools for explanation, inheriting their core vulnerabilities [16]–[18]. While these studies successfully demonstrate that *XAI* can provide a preliminary layer of transparency by generating feature importance scores, the methodological focus remains on applicability rather than *fidelity*. Concurrently, studies focused on making *IDS* models more resilient to adversarial attacks or more efficient for resource-constrained devices still apply *SHAP* to explain these improved models, thus inheriting the core methodological vulnerability [16]–[18]. **Research Gap:** The universal limitation in this body of work is the uncritical reliance on *SHAP* and *LIME*. By ignoring the inherent dependencies in *Non-IID* network data, the resulting explanations lack guaranteed fidelity and may not faithfully represent the model's logic.

Contribution 1: Our *HiFi-XAI* framework directly resolves this gap. By using *CVAE* to model the underlying data distribution, it generates high-fidelity, trustworthy explanations grounded in plausible feature correlations.trustworthiness.

B. Advanced Unsupervised, Rule-Extraction, and Semantic Methods

A more advanced line of research moves beyond standard post-hoc explanations by interpreting unsupervised models or extracting human-readable rules directly from the model. For instance, some works have proposed methods to explain *Self-Organizing Maps* or to translate the complex decision process of an *Autoencoder* into a set of *allow-list* rules [19], [20]. Other works use *XAI* as a debugging tool to discover structural problems and class imbalances within *IDS* datasets [21]. **Research Gap:** These methods, while innovative, still lack fidelity as they do not explicitly model data correlations. Furthermore, the *LLM*-based approaches suffer from a "garbage in, garbage out" problem: feeding a fluent *LLM* with low-fidelity explanations produces a well-written but misleading narrative.

Contribution 2: *HiFi-XAI* addresses both gaps. The *CVAE* stage first ensures high-fidelity attributions by modeling data dependencies. Only then does the *LLM*-powered engine translate these faithful results into a truly human-understandable narrative, ensuring the final explanation is both coherent and accurate.

C. Feasibility and Real-Time Deployment of IDS

Research focused on practical deployment typically optimizes the detection model for resource-constrained environments, such as those found in heterogeneous network deployments [22], [23]. These works discuss IDS systems but do not strongly deploy them to the resource-constrained devices to check their performance. However, this line of work generally neglects the computational cost of the subsequent interpretability layer. When they do include *XAI*, they rely on standard tools like *SHAP* or *LIME*, which can introduce significant latency and are not optimized for real-time inference, especially when generating computationally intensive, high-fidelity explanations. **Research Gap:** No existing framework simultaneously delivers both high explanation fidelity and the computational feasibility required for real-time deployment on lightweight edge devices.

Contribution 3: We address this by engineering *CVAE* for practical application. We demonstrate its feasibility on a real-time, resource-constrained *IDS* testbed, leveraging a high-performance *C++/CUDA* backend to make high-fidelity XAI a practical tool for demanding edge environments.

III. Proposed Methodology: The HiFi-XAI Framework

To address the dual challenges of fidelity and semantic interpretability in XAI for Intrusion Detection Systems (IDS), we propose *HiFi-XAI*, a novel three-stage framework. The Figure 1 illustrates that our framework is designed to be model-agnostic, generate high-fidelity feature attributions through a synthesis process, and finally translate these attributions into human-readable, actionable insights using a fine-tuned LLM.. This architecture is explicitly designed to be deployable on resource-constrained devices.

A. Model-Agnostic Interface

The foundation of our framework is its model-agnostic property. The Fidelity Engine does not require access to the internal architecture or parameters of the IDS model being explained. Instead, it treats the model as a black box, interacting solely through its prediction function, specifically the probabilistic output. This ensures that HiFi-XAI is universally compatible with any existing or future complex IDS model, from traditional machine learning classifiers like Random Forest and XGBoost to complex deep learning architectures like DNNs and Transformers. This interface is the critical first step, enabling the subsequent attribution analysis without sacrificing flexibility.

B. Phase 2: Global Attribution Synthesis via Conditional Value Attribution (CVA)

The core of our framework's trustworthiness lies in the global attribution synthesis, which implements a novel method we term Conditional Value Attribution (CVA). CVA is grounded in Shapley value theory but re-frames the calculation to respect the underlying data distribution, thereby solving the fidelity problem caused by the feature independence assumption.

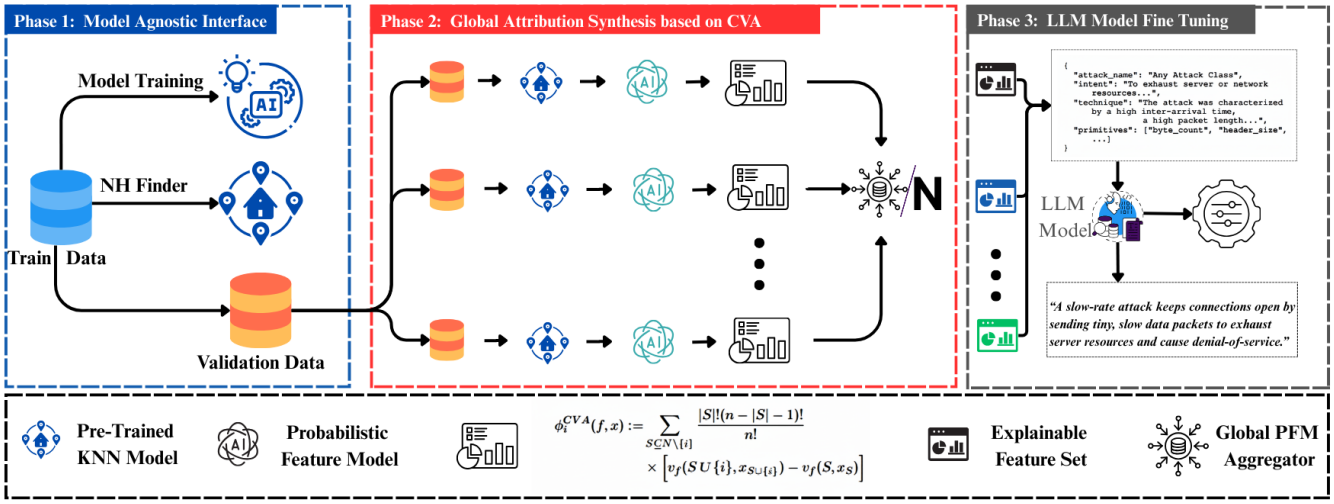


Fig. 1: The HiFi-XAI three-stage architecture. First, a model-agnostic interface finds local data neighborhoods using NH Finder. Second, a local PFM is trained on-the-fly to generate high-fidelity CVAE attributions. Finally, a fine-tuned LLM translates these attributions into a human-readable explanation.

1) Mathematical Preliminaries: Let the feature space be a set of n random variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ defined on a probability space where P is the actual joint probability measure over the features. Let $f : \mathbb{R}^n \to \mathbb{R}$ be the AI classifier learning model we aim to explain. For a specific instance $x \in \mathbb{R}^n$ for which we seek an explanation, a coalition is a subset of features $S \subseteq N$, where $N = \{1, \ldots, n\}$. A critical component of our framework is a *Probabilistic Feature Model (PFM)*, denoted as $\mathcal{M}$, which is a generative model trained to approximate the actual data distribution $P(\mathbf{X})$. In our framework, we specifically implement the PFM using a *Gaussian Mixture Model (GMM)* [24] with `n_components=5`. This deliberate choice prioritized computational efficiency and suitability for our dynamic, just-in-time explanation paradigm. Unlike more complex generative models like VAEs or GANs, a GMM is significantly lighter and faster to train, a critical requirement for the online phase where a new, ephemeral PFM is trained on-the-fly for each explanation using only a local data manifold identified by a k-NN index.

GMMs are highly effective at this task, as they are well-suited to approximating the local, multi-modal probability densities found within these small data neighborhoods. Furthermore, GMMs provide a stable and direct method for drawing samples, essential for the Monte Carlo approximation of the conditional expectation in our CVAE algorithm, as detailed in Algorithm 2. By instantiating a new, lightweight GMM for each local data subset, we ensure that the conditional sampling accurately reflects the specific feature dependencies relevant to the instance being explained, maximizing fidelity without incurring prohibitive computational costs.

2) Definition 1: The Conditional Expectation Value Function: Standard Shapley value approximations like SHAP calculate the value of a coalition $v(S)$ by assuming feature independence. This is the source of the fidelity flaw. We rectify this by defining a new value function conditioned on the actual data distribution learned by the PFM.

The value function, $v_f(S, x_S)$, for a coalition S with observed values x_S, is the expected output of the model f, where the expectation is taken over the features not in S (denoted $\bar{S}$), conditioned on the observed values x_S.

$$v_f(S, x_S) := \mathbb{E}_{X_{\bar{S}} \sim P(X_{\bar{S}} | X_S = x_S)}[f(x_S, X_{\bar{S}})] \tag{1}$$

This expectation is computed using the PFM, $\mathcal{M}$. Expressed as an integral, this is:

$$v_f(S, x_S) = \int_{x_{\bar{S}} \in \text{dom}(X_{\bar{S}})} f(x_S, x_{\bar{S}}) \cdot p(x_{\bar{S}} | x_S) dx_{\bar{S}}$$

where $p(x_{\bar{S}}|x_S)$ is the conditional probability density function derived from $\mathcal{M}$. This formulation ensures that the model is only evaluated over plausible, in-distribution data configurations.

3) Definition 2: Conditional Value Attribution (CVA): Using this new, high-fidelity value function, the Conditional Value Attribution (CVA) for a feature i at an instance x is the unique solution satisfying the classic Shapley axioms:

$$\phi_i^{CVA}(f, x) := \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \times \Big[v_f(S \cup \{i\}, x_{S \cup \{i\}}) - v_f(S, x_S) \Big] \tag{2}$$

This value, ϕ_i^{CVA}, represents the fair contribution of feature i's observed value, x_i, to the final prediction, averaged over all possible contexts of other features in a probabilistically coherent manner.

4) Algorithmic Implementation: This phase is the core of our fidelity-aware approach. We aggregate Conditional Value Attribution (CVA) scores from multiple instances from a validation set to derive a stable global feature importance ranking.

A critical innovation in our approach is the dynamic, on-the-fly instantiation of the Probabilistic Feature Model (PFM) for each instance.

Instead of relying on a single, global PFM, which may fail to capture localized data patterns, our method first identifies the local data manifold for the instance being explained using a pre-trained k-NN index. It then trains a new, ephemeral PFM exclusively on this local neighborhood. This "just-in-time" PFM generation ensures that each local explanation is calculated with the highest possible fidelity. These high-fidelity local attributions are then aggregated to form a robust global ranking.

The end-to-end process for generating this global importance list is detailed in Algorithm 1. This unified algorithm illustrates the workflow, from iterating through the validation sample to the final aggregation and normalization of the local CVA scores. It uses the value function approximation from Algorithm 1 as a subroutine.

Algorithm 1 Global CVA Calculation via Local PFM Aggregation

Input: Model f, k-NN Index $\mathcal{I}$, Training Data X_{train}, Validation Sample X_{val}, coalition samples K
Output: Global CVA values Φ_{global}
Initialize global accumulator $\Phi_{\text{acc}} \leftarrow \{0\}^n$
Let $M \leftarrow |X_{\text{val}}|$ be the number of validation instances
for each instance x_j in X_{val} **do**
    Find neighbor indices N_{x_j} for x_j using k-NN Index $\mathcal{I}$ on X_{train}
    Create local data subset $X_{\text{local}} \leftarrow X_{\text{train}}[N_{x_j}]$
    Train an ephemeral local PFM, M_{local}, on X_{local}
    Initialize local attributions $\phi_{\text{local}} \leftarrow \{0\}^n$
    for $k = 1$ to K **do**
        Randomly choose a permutation π of feature indices $\{1, \ldots, n\}$
        Randomly choose a split point $p \in \{0, \ldots, n-1\}$
        Let $S \leftarrow \{\pi(1), \ldots, \pi(p)\}$ and $i \leftarrow \pi(p+1)$
        $v_{S\cup\{i\}} \leftarrow \text{ApproximateValueFunction}(f, x_j, S \cup \{i\}, M_{\text{local}})$
        $v_S \leftarrow \text{ApproximateValueFunction}(f, x_j, S, M_{\text{local}})$
        $\phi_{\text{local},i} \leftarrow \phi_{\text{local},i} + (v_{S\cup\{i\}} - v_S)$
    end for
    $\phi_{\text{local}} \leftarrow \phi_{\text{local}}/K$
    $\Phi_{\text{acc}} \leftarrow \Phi_{\text{acc}} + |\phi_{\text{local}}|$
end for
$\Phi_{\text{global}} \leftarrow \Phi_{\text{acc}}/M$
return Φ_{global}

C. Phase 3: Semantic Engine and LLM Fine-Tuning

The Fidelity Engine outputs a numerically precise, high-fidelity list of feature attributions (ϕ^{CVA}). While trustworthy, this output is not yet a human-friendly explanation. The Semantic Engine bridges this final gap by translating the CVA output into a natural language narrative.

Algorithm 2 Value Function Approximation via Local PFM

Input: Model f, instance x, coalition S, Local PFM M_{local}, MC samples M_s
Output: Approximated value v_S
    ▷ *Approximate $p(x_{\overline{S}}|x_S)$ by sampling from the local PFM*
$X_{\text{bg}} \leftarrow M_{\text{local}}.\text{sample}(M_s)$
    ▷ *Construct full data points $(x_S, x_{\overline{S}})$ for the integral*
$X_{\text{final}} \leftarrow \text{combine}(x_S, X_{\text{bg}})$
▷ *Return the Monte Carlo approximation of the integral in Eq. 1*
return $\frac{1}{M_s}\sum_{m=1}^{M_s} f(X_{\text{final}}^{(m)})$

1) LLM Fine-Tuning: To ensure the LLM understands the specific domain of network intrusion detection, we fine-tune a pre-trained model on a curated dataset of security incident descriptions. This dataset was not manually created but was systematically generated through a data-driven pipeline to ensure it faithfully reflects the underlying network traffic patterns based on semantics and CVAE features.

The process began by training high-performance expert models on the full CICIDS2017 [25] and CICIOT2023 [26] datasets. For each correctly identified attack instance in the test set, we employed CVAE to determine the top-k most influential features driving the model's decision. These numerical feature importances were then translated into features are were added in JSON object containing the attack name, a mapped attack intent, a human-readable description of the technical characteristics, and a list of abstract feature primitives. The primitives contain CVAE derived top features. An example data point is shown below based on semantic features and data we used for training our LLM.

```
"attack_name": "Any Attack Class",
"intent": "To exhaust server or network
  resources...",
"technique": "The attack was characterized
  by a high inter-arrival time, a high
  packet length...",
"primitives": ["byte_count", "header_size",
  ...]
```

Listing 1: Example Semantic Data Point

This automated process yielded a comprehensive semantic corpus containing over 50,000 unique security incident descriptions. We utilized a random sample of 10,000 records from this corpus for the fine-tuning processing a bigger datasets was not possible based on our existing computation we used in our research. This data-driven approach teaches the LLM to connect low-level feature primitives to high-level attack concepts and intents, grounded in the statistical reality of the original datasets. The dataset will be made available to the public.

2) Explanation Generation: Once the CVA has identified the top features for a given instance, these features are formatted into a prompt and sent to the fine-tuned LLM. The LLM,

now equipped with domain-specific knowledge, generates a concise, human-readable explanation that describes the likely attack, its intent, and the technical evidence supporting the conclusion, grounded in the high-fidelity features provided by the CVA. The end-to-end process is summarized in Algorithm 3.

Algorithm 3 The HiFi-XAI End-to-End Framework

Input: IDS Model f, PFM $\mathcal{M}$, Fine-tuned LLM $\mathcal{L}$, instance x
Output: Natural language explanation E_{text}
// Phase 2: Global Attribution Synthesis via CVA
$\phi^{CVA} \leftarrow \text{CalculateCVA}(f, x, \mathcal{M})$
$F_{top-k} \leftarrow \text{GetTopKFeatures}(\phi^{CVA}, k = 10)$
// Phase 3: Semantic Engine and LLM Fine-Tuning
$P_{prompt} \leftarrow \text{FormatAsPrompt}(F_{top-k})$
$E_{text} \leftarrow \mathcal{L}.\text{generate}(P_{prompt})$
return E_{text}

IV. Deployment on Resource-Constrained Devices

A key goal of our work is to make high-fidelity XAI practical for real-time IDS on edge devices, which includes a two-phase deployment model that maximizes offline computation to enable lightweight online explanation. Our approach using dynamic local PFMs is uniquely suited for this task.

A. Deployment Model: Offline Preparation and Online Explanation

The offline phase can be trained on high end systems and online phase will be executed on resource constrained devices. In the *Offline Phase*, executed on a powerful server, we first train the primary IDS model (f) and a k-Nearest Neighbors (k-NN) index ($\mathcal{I}$) on the full training data. The resulting deployment artifacts includes the trained model, the k-NN index, and a representative data sample (X_{train})—are then packaged for the device.

In the *Online Phase*, executed on the resource-constrained device, this lightweight package is loaded. When an explanation for a prediction $f(x)$ is needed, a streamlined CVA process begins. The pre-trained k-NN index instantly identifies the local data neighborhood for x, upon which a new, ephemeral Probabilistic Feature Model (PFM) is trained on-the-fly. The CVA algorithm then performs fast sampling queries against this temporary, perfectly tailored PFM to generate the high-fidelity explanation.

B. Advantage Over SHAP and LIME for Deployment

This two-phase model provides a significant advantage over other model-agnostic XAI methods like SHAP's KernelExplainer and LIME. These methods require deploying a large, general-purpose background dataset on the device.

In contrast, our approach deploys a k-NN index and a training data sample. This package is not only often smaller, but it also provides higher fidelity. By dynamically instantiating a PFM on the most relevant local data manifold for each explanation, we avoid the fidelity-performance trade-off, achieving both speed and accuracy in a fundamentally better-suited form for lightweight, real-time IDS deployment.

V. Results and Discussion

To validate our HiFi-XAI framework's fidelity, semantic richness, and practical feasibility, we conducted a series of targeted experimental evaluations and compared it with state-of-the-art XAI algorithms. All models were trained on a high-performance workstation running Ubuntu 24.04 with an NVIDIA RTX 3070 (8GB VRAM) and 64GB of RAM. The experiments utilized two benchmark datasets: CICIDS2017 and CICIOT2023.

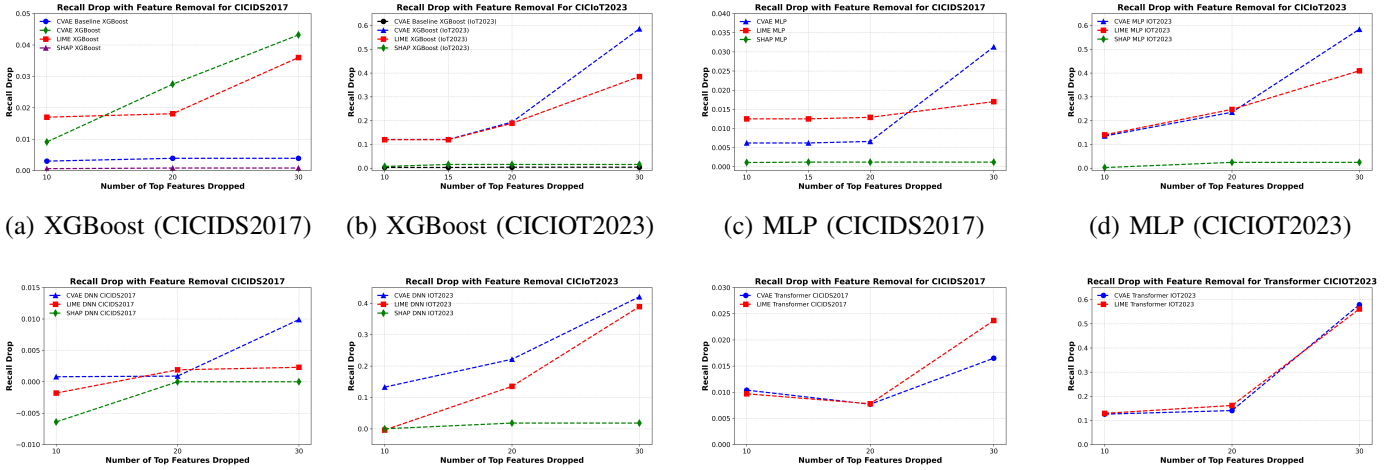
A. Measuring Fidelity: The Explainable Trust Score (ETS)

To quantitatively measure the fidelity and trustworthiness of an XAI method for IDS, we introduce the *Explainable Trust Score (ETS)*. This metric is specifically designed to quantify how effectively an XAI technique identifies features critical to a model's predictions of *malicious activity*. The rationale is that a high-fidelity XAI method will identify features whose removal causes a significant drop in the model's performance on attack classes. It explains the decision of any black box model.

The ETS is calculated based on the drop in *macro-averaged recall* for attack classes after ablating the top-k features identified by the XAI method. We chose macro-recall for two key reasons:

1) *Focus on Attack Detection:* In a security context, the primary goal is to correctly identify all attacks (True Positives) while minimizing missed attacks (False Negatives). Recall is the most direct measure of this capability.
2) *Class-Agnostic Importance:* Using a macro average ensures that every attack class contributes equally to the final score without any internal bias or class imbalance issues. This prevents the metric from being skewed by high-volume, easy-to-detect attacks and ensures the model's ability to identify rare but critical threats is also measured.

The number of features to drop, k, is a variable that can vary proportionally to the dataset's dimensionality and features to ensure a fair comparison. Dropping more than 50% can sometimes lead to a higher accuracy drop, but we compare this by keeping the same scenario for all state-of-the-art models. For our experiments on CICIDS2017 (71 features) and CICIOT2023 (46 features), we used k values of 10, 20, and 30, representing a progressive removal of the most impactful features. CICIDS2017 consistently achieved the highest accuracy established in the Related Work section, which is less sensitive to features, and CICIOT2023, a real-time dataset with 33 attack classes, is sensitive to features. So we considered both datasets for robust evaluation and to avoid sensitive bias. A higher ETS, as calculated in Table I, signifies a more trustworthy explanation, as it indicates that the XAI method has successfully pinpointed features whose



(a) XGBoost (CICIDS2017) (b) XGBoost (CICIOT2023) (c) MLP (CICIDS2017) (d) MLP (CICIOT2023)

(e) DNN (CICIDS2017) (f) DNN (CICIOT2023) (g) Transformer (CICIDS2017) (h) Transformer (CICIOT2023)

Fig. 2: Impact of Feature Ablation on model recall, comparing CICIDS2017 and CICIOT2023 across different models. The steeper drop in recall for CVAE indicates its superior fidelity in identifying critical features.

absence critically degrades the model's core security mission: to detect attacks.

B. State-of-the-Art Comparison of XAI Fidelity

Table I compares our XAI algorithm, CVAE method, against state-of-the-art IDS and XAI techniques such as SHAP and LIME, with their inherent library functionalities such as DeepLIFT, DeepSHAP, TreeSHAP from recent literature. We replicated the architectural principles described in these works to ensure a fair and direct comparison. Since the source code and specific training environments were unavailable, the baseline recall and subsequent ablation study results are derived from our implementations on a standardized platform. This approach allows us to isolate the performance of the XAI methods themselves.

C. Fidelity Evaluation: Feature Ablation Study

To empirically test this and justify our Contribution 1, we performed a feature ablation study. The logic is straightforward: if an XAI method correctly identifies the most important features, removing them should cause a more significant drop in the model's predictive performance (in this case, attack recall) than removing features identified by a less faithful method.

The results, shown in Fig. 2, provide clear and consistent evidence of *CVAE's superior fidelity*. Across all five tested models (Random Forest, XGBoost, MLP, DNN, and Transformer) and both datasets, the performance drop was sharpest when removing features ranked by CVAE. In contrast, removing the top features identified by SHAP and LIME resulted in a much smaller, more gradual degradation of model recall. The LIME is not mathematically and theoretically justified, and local explanations are generated and added to produce global explanations. However, in the case of IDS, it performs better than SHAP and can also explain the transformer output, where SHAP fails to analyze. This outcome directly validates our hypothesis. Consequently, they fail to identify the genuinely critical features of the model's reasoning. While LIME occasionally performs slightly better than SHAP, its theoretical foundation relies on local linear approximations, less robust than the Shapley value theory that underpins SHAP and our CVAE method. CVAE, by explicitly modeling conditional feature dependencies, provides a more faithful and reliable representation of the model's logic.

D. Computational Analysis on Edge Devices

To validate Contribution3, we deployed our framework on a *Raspberry Pi 4* and measured the computational overhead required for the explanation. The results, summarized in TABLE II, demonstrate the feasibility of our approach.

The most critical finding is the *extremely low memory footprint* during the real-time explanation phase (14.31-16.81 MB). While the total explanation time is computationally intensive, the high-fidelity CVAE calculation accounts for over 99% of this duration, with all preparatory steps completing in just over one second. The crucial achievement is not sub-second latency, but rather making high-fidelity, conditional-aware explanations *feasible on a device with a minimal memory footprint*, a task previously confined to powerful servers.

E. Semantic Engine Evaluation

The final phase of our framework translates the high-fidelity CVA scores and sends them to primitives used for LLM fine-tuning, and justifies our contribution to Contribution 2. We fine-tuned lightweight, open-source LLM *Gemma 2B* [31]from Google to ensure this component remains suitable for resource-constrained environments. During inference, the top-k features from CVAE are formatted into a structured prompt, enabling the LLM to generate concise and contextually aware explanations. Two representative outputs generated during evaluation are illustrated below.

TABLE I: State-of-the-Art Comparison of XAI Fidelity using Explainable Trust Score (ETS). A higher ETS signifies a more faithful explanation. Our proposed CVAE method consistently outperforms others by a significant margin.

Study	Year	ML Model	XAI Method	Dataset	Baseline Recall	Recall Drop (Top 30)	ETS
CICIDS2017 Dataset							
HIFI-XAI	**2025**	**XGBoost**	**CVAE**	**CICIDS2017**	**0.9819**	**0.0432**	**0.144**
HIFI-XAI	**2025**	**Random Forest**	**CVAE**	**CICIDS2017**	**0.9925**	**0.0735**	**0.245**
HIFI-XAI	**2025**	**Transformer**	**CVAE**	**CICIDS2017**	**0.9499**	**0.0165**	**0.055**
Alnahdi, A. et al. [27]	2024	MLP	LIME	CICIDS2017	0.9542	0.0170	0.056
Attique, D. et al. [28]	2024	DNN	SHAP	CICIDS2017	0.9994	0.0000	0.000
Arreche, O. et al. [29]	2024	Random Forest	SHAP	CICIDS2017	0.9640	0.0002	0.0006
CICIoT2023 Dataset							
HIFI-XAI	**2025**	**XGBoost**	**CVAE**	**CICIoT23**	**0.9561**	**0.5856**	**1.952**
HIFI-XAI	**2025**	**Random Forest**	**CVAE**	**CICIoT23**	**0.9560**	**0.6293**	**2.097**
Wali, S. et al. [30]	2025	Random Forest	SHAP	CICIoT23	0.9750	0.0099	0.033
Bin hulayyil, S. et al. [15]	2025	XGBoost	SHAP	CICIoT23	1.0000	0.0156	0.052

TABLE II: Computational Overhead on Raspberry Pi 4

Metric	CICIDS2017	CICIOT2023
Total Disk Size of Artifacts	924.86 MB	927.79 MB
Time to Load Artifacts	9.36 seconds	3.27 seconds
Total Explanation Time	**124.94 s**	**70.98 s**
Peak Memory Usage (Exp.)	**16.81 MB**	**14.31 MB**

Security Incident Analysis (CICIDS2017)

Technique: Data exfiltration or data leak using a low-and-slow approach to covertly steal data over a prolonged period while bypassing detection mechanisms.

Security Incident Analysis (CICIOT2023)

Technique: Denial-of-Service (DoS) or reconnaissance scan, leveraging SYN or UDP floods to overwhelm IoT device resources and disrupt services.

While these examples demonstrate the qualitative effectiveness of our approach, we also conducted a quantitative analysis to evaluate the quality and coherence of the Semantic Engine. We created a test set of 100 semantic JSON blueprints randomly sampled from our generated corpus and manually authored a corresponding gold standard human-readable explanation for each. This reference set served as our ground truth. We then used the fine-tuned Gemma-2B model to generate an explanation for each of the 100 blueprints. The model-generated text was then programmatically compared against our human-authored ground truth using two standard NLP metrics:

- *ROUGE-L:* [32] This metric measures content overlap by identifying the longest common subsequence between the generated and reference texts. It is effective at assessing whether the key technical terms are present.
- *BERTScore:* [33]This advanced metric computes the semantic similarity between the texts using contextual embeddings and human-written ones. It goes beyond simple word matching to determine if the generated text accurately captures the *meaning* and *intent* of the reference explanation.

Our fine-tuned model achieved a *ROUGE-L score of 0.9147* and an average *BERTScore F1 of 0.9895* in comparison to a correctness score of 0.80 achieved in recent work [27]. The strong results, combining qualitative and quantitative evidence, successfully bridge the semantic gap for security analysts.

VI. Conclusion

A novel framework, HiFi-XAI, designed to solve the critical challenges of fidelity and semantic interpretability in XAI for Intrusion Detection Systems, is introduced and demonstrated in this paper. Traditional methods like SHAP and LIME provide untrustworthy explanations for network data by failing to account for feature dependencies in non-IID data scenarios. HiFi XAI develops higher trust by introducing CVAE, a Shapley value-based method that respects the underlying data distribution to generate high-fidelity feature attributions.

Our contributions are threefold. First, we empirically demonstrated through feature ablation studies that CVAE significantly outperforms existing methods in identifying the true features of a model's prediction for multiple instances at the global level. Second, we integrated CVAE with a fine-tuned, lightweight LLM gemma 2B to translate these faithful attributions into semantically rich, human-readable narratives. Finally, we validated by deploying the entire framework to check the feasibility for real-world application on a resource-constrained Raspberry Pi 4, proving that high-fidelity XAI is practical for edge security environments. In the Future, we will explore more advanced LLM methods and focus on developing architecture-aware XAI methods based on CVAE as a baseline.

Acknowledgement

The project (TTDF/6G/517) is sponsored by the Telecom Technology Development Fund (TTDF), Department of Telecommunication (DoT), Government of India. We gratefully acknowledge their support and encouragement.

REFERENCES

[1] C. Rajathi and P. Rukmani, "Hybrid learning model for intrusion detection system: A combination of parametric and non-parametric classifiers," *Alexandria Engineering Journal*, vol. 112, pp. 384–396, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016824012651

[2] S. Neupane, J. Ables, W. Anderson, S. Mittal, S. Rahimi, I. Banicescu, and M. Seale, "Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities," *IEEE Access*, vol. 10, pp. 112 392–112 415, 2022.

[3] N. Moustafa, N. Koroniotis, M. Keshk, A. Y. Zomaya, and Z. Tari, "Explainable intrusion detection for cyber defences in the internet of things: Opportunities and solutions," *IEEE Communications Surveys Tutorials*, vol. 25, no. 3, pp. 1775–1807, 2023.

[4] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017. [Online]. Available: https://arxiv.org/abs/1705.07874

[5] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016. [Online]. Available: https://arxiv.org/abs/1602.04938

[6] D. Gaspar, P. Silva, and C. Silva, "Explainable ai for intrusion detection systems: Lime and shap applicability on multi-layer perceptron," *IEEE Access*, vol. 12, pp. 30 164–30 175, 2024.

[7] D. M. J. G., D. Solans, M. Heikkila, A. Vitaletti, N. Kourtellis, A. Anagnostopoulos, and I. Chatzigiannakis, "Non-iid data in federated learning: A survey with taxonomy, metrics, methods, frameworks and future directions," 2024. [Online]. Available: https://arxiv.org/abs/2411.12377

[8] P. Goldschmidt and D. Chudá, "Network intrusion datasets: A survey, limitations, and recommendations," *Computers Security*, vol. 156, p. 104510, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404825001993

[9] Z. Deng, A. Torim, S. Ben Yahia, and H. Bahsi, "Generative ai in intrusion detection systems for internet of things: A systematic literature review," *IEEE Open Journal of the Communications Society*, vol. 6, pp. 4689–4717, 2025.

[10] R. Kumar, D. Javeed, A. Aljuhani, A. Jolfaei, P. Kumar, and A. K. M. N. Islam, "Blockchain-based authentication and explainable ai for securing consumer iot applications," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 1145–1154, 2024.

[11] A. Aljuhani, A. Alamri, P. Kumar, and A. Jolfaei, "An intelligent and explainable saas-based intrusion detection system for resource-constrained iomt," *IEEE Internet of Things Journal*, vol. 11, no. 15, pp. 25 454–25 463, 2024.

[12] D. Javeed, T. Gao, P. Kumar, S. Shoukat, I. Ahmad, and R. Kumar, "An intelligent and interpretable intrusion detection system for unmanned aerial vehicles," in *ICC 2024 - IEEE International Conference on Communications*, 2024, pp. 1951–1956.

[13] D. Attique, W. Hao, W. Ping, D. Javeed, and M. Adil, "Ex-dfl: An explainable deep federated-based intrusion detection system for industrial iot," in *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2024, pp. 358–364.

[14] M. Adil, M. A. Jan, S. Bin Hakim, H. H. Song, and Z. Jin, "xids-ensembleguard: An explainable ensemble learning-based intrusion detection system," in *2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, 2024, pp. 93–100.

[15] S. B. hulayyil, S. Li, and N. Saxena, "Explainable ai-based intrusion detection in iot systems," *Internet of Things*, vol. 31, p. 101589, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660525001027

[16] D. Javeed, T. Gao, P. Kumar, and A. Jolfaei, "An explainable and resilient intrusion detection system for industry 5.0," *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 1342–1350, 2024.

[17] N. N. Tai, N. N. Tai, N. D. Tan, N. D. Tan, T.-N. To, T.-N. To, P. T. Duy, P. T. Duy, V.-H. Pham, and V.-H. Pham, "A robust and trustworthy intrusion detection system using adversarial machine learning and xai," in *2024 International Conference on Advanced Technologies for Communications (ATC)*, 2024, pp. 407–412.

[18] M. Umair, M. S. Khan, W. A. Malwi, F. Asiri, I. Nafea, F. Saeed, and J. Ahmad, "Knowledge distillation for lightweight and explainable intrusion detection in resource-constrained consumer devices," *IEEE Transactions on Consumer Electronics*, pp. 1–1, 2025.

[19] C. S. Wickramasinghe, K. Amarasinghe, D. L. Marino, C. Rieger, and M. Manic, "Explainable unsupervised machine learning for cyber-physical systems," *IEEE Access*, vol. 9, pp. 131 824–131 843, 2021.

[20] M. Fan, J. Zuo, J. Zhu, and Y. Lu, "Explainable anomaly-based intrusion detection for specialized iot environments enabled by rule extraction from autoencoder," *IEEE Internet of Things Journal*, vol. 12, no. 12, pp. 19 504–19 521, 2025.

[21] E. Lanfer, S. Sylvester, N. Aschenbruck, and M. Atzmueller, "Leveraging explainable ai methods towards identifying classification issues on ids datasets," in *2023 IEEE 48th Conference on Local Computer Networks (LCN)*, 2023, pp. 1–4.

[22] S. Jeong, S. Lee, H. Lee, and H. K. Kim, "X-canids: Signal-aware explainable intrusion detection system for controller area network-based in-vehicle network," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 3, pp. 3230–3246, 2024.

[23] A. Oki, Y. Ogawa, K. Ota, and M. Dong, "Evaluation of applying federated learning to distributed intrusion detection systems through explainable ai," *IEEE Networking Letters*, vol. 6, no. 3, pp. 198–202, 2024.

[24] A. Reddy, M. Ordway-West, M. Lee, M. Dugan, J. Whitney, R. Kahana, B. Ford, J. Muedsam, A. Henslee, and M. Rao, "Using gaussian mixture models to detect outliers in seasonal univariate network traffic," in *2017 IEEE Security and Privacy Workshops (SPW)*, 2017, pp. 229–234.

[25] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *International Conference on Information Systems Security and Privacy*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4707749

[26] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "Ciciot2023: A real-time dataset and benchmark for large-scale attacks in iot environment," *Sensors*, vol. 23, no. 13, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/13/5941

[27] A. Alnahdi and S. Narain, "Towards transparent intrusion detection: A coherence-based framework in explainable ai integrating large language models," in *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, 2024, pp. 87–96.

[28] D. Attique, W. Hao, W. Ping, D. Javeed, and P. Kumar, "Explainable and data-efficient deep learning for enhanced attack detection in iiot ecosystem," *IEEE Internet of Things Journal*, vol. 11, no. 24, pp. 38 976–38 986, 2024.

[29] O. Arreche, T. Guntur, and M. Abdallah, "Xai-based feature selection for improved network intrusion detection systems," 2024. [Online]. Available: https://arxiv.org/abs/2410.10050

[30] S. Wali, Y. A. Farrukh, and I. Khan, "Explainable ai and random forest based reliable intrusion detection system," *Computers Security*, vol. 157, p. 104542, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404825002317

[31] G. Team, M. Riviere, S. Pathak, P. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, J. Ferret, P. Liu, P. Tafti, A. Friesen, M. Casbon, S. Ramos, R. Kumar, C. Lan, S. Jerome, and A. Andreev, "Gemma 2: Improving open language models at a practical size," 07 2024.

[32] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," 01 2004, p. 10.

[33] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020. [Online]. Available: https://arxiv.org/abs/1904.09675