Towards Better-Calibrated ML Models for Reliable Network Intrusion Detection via Calibration-Aware SHAP-based Feature Selection

Hussein Fawaz[1,2,*], Fatima Ezzeddine[1,2], Silvia Giordano[1], Omran Ayoub[1]

[1]University of Applied Sciences and Arts of Southern Switzerland, Switzerland, [2]Università della Svizzera italiana, Switzerland
* *Corresponding author: hussein.fawaz@usi.ch*

***Abstract*—Model calibration and feature selection are two critical aspects in developing reliable and accurate machine learning models. Calibration ensures that the model's confidence scores accurately reflect the true likelihood of correctness, which is essential in security-critical applications, like Network Intrusion Detection Systems (NIDS). Feature selection (FS), meanwhile, enhances model efficiency, interpretability, and generalization by identifying the most relevant inputs. However, it may also degrade model's calibration if not explicitly considered. Recently, Explainable Artificial Intelligence (XAI) methods, particularly Shapley Additive Explanations (SHAP), have proven effective in guiding the FS process. Yet, existing SHAP-based FS techniques typically focus on improving accuracy, often giving little attention to the impact of FS on model calibration, an aspect that is critical for reliable decision-making in high-stakes applications such as NIDS. In this work, we propose a novel approach that incorporates a calibration-aware loss function for XGBoost with SHAP-based Recursive Feature Elimination for FS to jointly improve both predictive accuracy and model calibration. Experiments on two benchmark NIDS datasets show that our approach can reduce Brier score and Expected Calibration Error by up to 3.5% and 84.8%, respectively, over uncalibrated baselines, and by up to 2.6% and 58.3% over standard calibration methods, while enhancing or maintaining predictive accuracy and number of features.**

***Index Terms*—Intrusion Detection Systems, Model Calibration, Feature Selection, Explainable AI.**

I. Introduction

Machine Learning (ML) methods are widely used in Network Intrusion Detection Systems (NIDS) to detect and classify malicious activities in network traffic [1]. A key factor in the success of these models is the selection of input features, which directly impacts detection accuracy, computational efficiency, and interpretability [2], [3]. Feature selection (FS) is thus a critical step, aiming to identify the most informative features while reducing dimensionality. In the context of NIDS, recent work has increasingly explored FS for ML-based NIDS through *explainable artificial intelligence* (XAI), particularly using SHapley Additive exPlanations (SHAP) [4].

SHAP employs a cooperative game theory-based approach to assign importance scores to features based on their contributions to the model's predictions, revealing how features influence individual decisions. This is critical in NIDS as diverse intrusion behaviors may depend on different feature subsets and their correlations [5], and hence SHAP can highlight not only globally relevant features but also those specific to certain attack types or traffic patterns [6]. This allows practitioners to perform FS more effectively than without SHAP [7], [8].

While SHAP-based FS has shown considerable promise in enhancing predictive performance, most existing approaches tend to overlook an equally critical dimension, model calibration. Calibration refers to the extent to which a model's predicted probabilities align with actual outcomes, i.e., the model's confidence accurately reflects the likelihood of correctness. For instance, predictions made with 80% confidence should be correct approximately 80% of the time [9]. In this sense, calibration ensures that a model expresses high confidence only when warranted, and remains cautious when uncertainty is high. Conversely, poorly calibrated models may be overconfident, issuing high-probability predictions even when they are incorrect, or underconfident, hesitating when predictions are likely correct. This is particularly important in high-stakes settings where decision-making systems must be accurate and trustworthy.

In NIDS, particularly human-in-the-loop systems, calibration enables *confidence-based abstention*, allowing the model to refrain from making uncertain predictions [10]. By applying a confidence threshold, the system can determine whether to proceed with automated classification or defer to a human analyst [11]. This mechanism improves decision quality by ensuring only predictions with sufficient confidence are acted upon autonomously, while ambiguous cases are escalated for review.

Much of the existing work remains focused on optimizing classification accuracy on benchmark datasets, often neglecting the reliability of the model's confidence estimates. As our results will later demonstrate, FS can substantially reduce feature dimensionality and enhance predictive performance. Yet, when calibration is not explicitly addressed, this process can inadvertently compromise the model's output reliability. This trade-off is especially problematic in decision-critical applications like NIDS, and is more pronounced in human-in-the-loop scenarios, where workflows rely on calibrated uncertainty estimates to support abstention and deferral strategies [12].

In this work, we address these limitations by proposing a FS framework that jointly optimizes predictive performance, measured by accuracy and F1-score, and model calibration,

979-8-3503-9281-4/25/$31.00

measured using the Brier Score (BS) and Expected Calibration Error (ECE) [13]. To this end, we design a novel calibration-aware loss function that balances accuracy and calibration objectives. We then integrate this loss into a SHAP-guided FS pipeline, specifically employing Recursive Feature Elimination (RFE), which iteratively removes the least important features based on SHAP values. Although our approach is compatible with other FS techniques, we choose SHAP-based RFE for its effectiveness in identifying optimal feature subsets while preserving model performance.We evaluate our method against several baselines, including uncalibrated models and standard post-hoc calibration techniques such as isotonic regression, both with and without FS, on two NIDS benchmark datasets. Cross-validated experimental results show that our approach significantly improves model calibration, reducing Brier Score by up to 3.5% and ECE by up to 84.8% compared to uncalibrated baselines, and by up to 2.6% and 58.3%, respectively, over calibration-aware baselines. Importantly, these calibration gains are achieved while maintaining or improving predictive performance. In terms of feature efficiency, our method selects between 9.8 and 38.0 features, representing a reduction of 20.2% and 3.3% compared to SHAP-based FS without calibration, demonstrating competitive compactness alongside improved reliability.

The remainder of this paper is structured as follows. Sec. II discusses related work. Sec. III and IV details our proposed approach and methodology, respectively. Sec. V presents the evaluation settings and Sec. VI discusses the experimental results. Sec. VII concludes the paper.

II. Related Work

In this section, we review related work on SHAP-based FS and model calibration in intrusion detection.

SHAP-based Feature Selection: XAI techniques have been widely investigated in the context of ML-based IDS [14]. Among these, SHAP has gained significant traction due to its ability to estimate the contribution of each feature to individual model predictions. SHAP has been used for various purposes, including interpreting model behavior by identifying the most influential features, with the final aim of extracting insights about the underlying problem [2], [3].

Beyond its use for model interpretability, SHAP has also been employed as a feature ranking mechanism for FS, often in combination with techniques such as RFE and Recursive Feature Addition (RFA) [15]. In [16], the authors propose a high-performance IDS using a SHAP-based FS approach, preserving predictive performance while reducing number of features by up to 80%. Similarly, [17] introduces a SHAP-based FS method for encrypted traffic intrusion detection, comparing its effectiveness to traditional techniques like Information Gain and RFE. Their approach reduces features by 87%, cuts training time by half, and shrinks the model size by 30% while maintaining strong detection performance. In [18], multiple SHAP-based FS strategies identify critical characteristics unique to various AI models and network intrusion types, revealing the most influential features for IDS models using feature subsets. In [19] authors propose an XAI-based methods using SHAP and Local Interpretable Model-agnostic Explanations to optimize efficiency, by leveraging XAI for FS, which maintains high accuracy.

Model Calibration in Intrusion Detection Systems: Recent work has increasingly focused on improving model's calibration and uncertainty estimation in security applications. [20] proposes a Deep Belief Network (DBN) with uncertainty-aware dynamic early stopping for ransomware detection, showing improved accuracy (94% to 98%) and reduced false positives (0.18 to 0.10). Similarly, [21] addresses overconfidence in neural network-based intrusion detection through a Bayesian CNN ensemble, while [22] enhances calibration in anomaly detection using Bayesian Autoencoders that model both aleatoric and epistemic uncertainty. In [23] authors explore similar benefits through uncertainty-enhanced autoencoder architectures with Monte Carlo dropout and Bayesian approaches. The work demonstrates how uncertainty quantification improves detection reliability.

Unlike prior works, our approach jointly addresses both predictive performance and calibration objectives within a unified framework that integrates SHAP-based FS with a novel calibration-aware loss function. While existing studies leveraging SHAP primarily focus on interpretability or efficiency through FS, and recent calibration methods focus on uncertainty estimation in IDS, to the best of our knowledge, no work explicitly combines FS with calibration-driven objectives to optimize both simultaneously. Our experimental results demonstrate that this combined approach achieves substantial calibration improvements without sacrificing performance.

III. Reference Scenario and Objectives

The accurate quantification of uncertainty are critical for deploying reliable ML models in high risk NIDS. Different methods exist to address the issue of uncertainty, such as *Confidence-based uncertainty* and *Calibration-based uncertainty*. *Confidence-based uncertainty* interprets uncertainty directly from the predicted probabilities. A low maximum probability $\max_c p_c$ indicates the model is unsure. *Calibration-based uncertainty* model is said to be *well-calibrated* if its predicted probabilities reflect true likelihoods. For instance, among all predictions where the model assigns 70% confidence to the top class, that class should be correct approximately 70% of the time. A model can be confident (e.g., $p_c = 0.95$) but still be poorly calibrated if its true accuracy is much lower than 95%. To support safe and reliable predictions, uncertainty support abstention mechanism which is a post-processing step. Specifically, the model is allowed to abstain from making a prediction on input when it is uncertain or not sufficiently confident, specially when the model is calibrated.

Model uncertainty can arise from inherent miscalibration in powerful ML models, particularly ensemble methods such as XGBoost, or/and from irrelevant or redundant features. For instance, XGBoost aggregates decision trees that tend to produce overconfident probabilities at their leaf nodes, making

them naturally overconfident. The boosting process prioritizes classification accuracy (minimizing a loss function like log-loss) rather than explicitly optimizing for well-calibrated probabilities. Additionally, irrelevant or redundant features introduce noise into the learning process and distort confidence estimates, which can degrade both predictive performance and the reliability of uncertainty estimates. To address this, there is a need to optimize simultaneously for model performance, model uncertainty, and the effectiveness of FS methods.

Let $\mathcal{X} \subseteq \mathbb{R}^d$ denote the input space and $\mathcal{Y} = \{1, 2, \ldots, C\}$ the set of class labels for NIDS, where it can be to detect a specific class of intrusions or attacks. We consider a multi-class probabilistic classifier $h : \mathcal{X} \to \{0\} \cup \mathcal{Y}$ trained on dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, and where the output $h(x) = 0$ denotes abstention. A prediction of x is $h(x) = [p_1, p_2, \ldots, p_C]$ such that $\sum_{c=1}^{C} p_c = 1$).

A model is said to be *certain* about input x if the predicted probability is highly concentrated on a single class. That is,

$$\max_{c \in \{1,\ldots,C\}} p_c \approx 1.$$

In such cases, the model assigns nearly all its output to the corresponding correct class, reflecting high confidence in the prediction. Conversely, A model is *uncertain* when the predicted probability distribution is diffuse or close to uniform, indicating ambiguity across classes probabilities. Formally,

$$\max_{c \in \{1,\ldots,C\}} p_c \approx \frac{1}{C}.$$

This implies the model lacks a strong preference for any single class, and its output reflects high epistemic or aleatoric uncertainty. Additionally, FS plays a crucial role in improving generalization, reducing overfitting, and enhancing interpretability by removing the least informative and redundant features based on model weights or importance scores. Using SHAP values to assess feature informativeness is particularly beneficial, as SHAP's insights into feature contributions are more reliable than traditional feature importance and attribution methods, leading to a more accurate understanding of which features truly drive predictions. More importantly, when combined with uncertainty-aware calibration strategies, it can help produce models that are not only accurate and confident, but also robust to redundant correlations in the data, with maintaining model explainability with SHAP.

Our objective is to train a calibration-aware classifier and perform FS guided by SHAP[1], with the aim of maintaining high predictive performance and enhancing model's calibration for achieving a calibrated and trusted confidence estimates. Specifically, we aim to:

- Perform calibration-aware FS, potentially improving generalization and FS under uncertainty when integrated with the calibration-based uncertainty optimization methods.
- Learn an in-processing well-calibrated model, where predicted probabilities correspond to true likelihoods, by optimizing calibration metrics such as the *Brier score*.

[1]https://github.com/hussein-fawaz/Reliable-Network-Intrusion-Detection-via-Calibration-Aware-SHAP-based-Feature-Selection

IV. Methodology

Our methodology consists 1) designing a calibration-aware loss function for training a better calibrated model, specifically, XGBoost, and 2) integrating that with SHAP-guided FS.

A. Calibration-aware Loss Function

Let a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ used to train an XGBoost probabilistic classifier $h : \mathcal{X} \to \mathcal{Y}$ to predict a class $\mathcal{Y} = \{1, 2, \ldots, C\}$ the set of class labels. A prediction of x is probabilistic in the form of $h(x) = [p_1, p_2, \ldots, p_C]$. We aim to learn a well-calibrated model to improve the calibration of predicted probabilities compared to the baseline model. To this end, we propose an in-processing method, that incorporates a novel objective function that combines the standard multiclass softmax cross-entropy (L_{CE}) loss with the BS (L_{Brier}). To train the XGBoost model, the calibration-aware loss function have two main components:

$$L_{\text{Custom}} = \alpha\, L_{\text{CE}} + (1 - \alpha)\, L_{\text{Brier}} \tag{1}$$

where $\alpha \in [0, 1]$ is a hyperparameter that controls the trade-off between classification accuracy and probabilistic calibration. Setting $\alpha = 1$ recovers the standard cross-entropy loss, while $\alpha = 0$ corresponds to optimizing the BS alone.

The BS is a strictly proper scoring rule that measures the mean squared error between the predicted probability distribution and the true outcome. For a multiclass problem with K classes and N examples, let $\hat{p}_{ik}$ be the predicted probability for class k on instance i, and define $y_{ik} = 1$ if the true label $y_i = k$, and 0 otherwise. Then the multiclass BS is defined as:

$$L_{\text{Brier}} = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \big(\hat{p}_{ik} - y_{ik}\big)^2 \tag{2}$$

Intuitively, the BS is the mean squared difference between the model's full probability vector and the one-hot encoding of the true class. Because it is a proper scoring rule, a lower BS indicates both higher accuracy and better calibration of the predicted probabilities. In particular, $\text{BS} = 0$ corresponds to perfectly confident and correct predictions.

During training, XGBoost requires the first and second derivatives (*gradient* and *Hessian*) of the loss function with respect to the raw predicted logits, z_{ik}. These derivatives guide the tree-building process. For each sample i and class k, the gradient of the cross-entropy term with respect to z_{ik} is:

$$\nabla_{z_{ik}}(L_{\text{CE}}) = \hat{p}_{ik} - y_{ik} \tag{3}$$

The approximate Hessian (second derivative) of the cross-entropy term with respect to z_{ik} is:

$$H_{z_{ik}}(L_{\text{CE}}) = \hat{p}_{ik}(1 - \hat{p}_{ik}) \tag{4}$$

For the BS term, we employ practical and widely used approximations for its gradient and Hessian with respect to the logits, which are derived in a form analogous to the cross-entropy terms when propagating through a softmax layer. The gradient of the BS term with respect to z_{ik} is approximated:

$$\nabla_{z_{ik}}(L_{\text{Brier}}) \approx 2\big(\hat{p}_{ik} - y_{ik}\big) \tag{5}$$

The approximate Hessian of the BS term with respect to z_{ik}:

$$H_{z_{ik}}(L_{\text{Brier}}) \approx 2\hat{p}_{ik}(1-\hat{p}_{ik}) \tag{6}$$

The overall gradient $\nabla_{z_{ik}}(L_{\text{Custom}})$ and Hessian $H_{z_{ik}}(L_{\text{Custom}})$ for the custom objective function are computed as the convex combination of the respective terms:

$$\nabla_{z_{ik}}(L_{\text{Custom}}) = \alpha\nabla_{z_{ik}}(L_{\text{CE}}) + (1-\alpha)\nabla_{z_{ik}}(L_{\text{Brier}}) \tag{7}$$

$$H_{z_{ik}}(L_{\text{Custom}}) = \alpha H_{z_{ik}}(L_{\text{CE}}) + (1-\alpha)H_{z_{ik}}(L_{\text{Brier}}) \tag{8}$$

These combined gradients and Hessians are then provided to the XGBoost algorithm to iteratively build the ensemble of trees, optimizing the model towards both accurate classification and well-calibrated probabilistic predictions. It is worth noting that the gradient and Hessian of the Brier loss (Eq. 5-6) are scaled versions of the cross-entropy loss (Eq. 3-4). This scaling does not affect the class predictions themselves, hence the accuracy values are expected to remain stable across settings when the same random initialization is used. Instead, the difference manifests in how leaf weights are adjusted during boosting, which directly influences the confidence scores assigned to predictions. As a result, our calibration-aware loss primarily impacts the calibration of probability estimates rather than accuracy, which is consistent with our reported improvements in terms of BS and ECE.

B. SHAP-Enhanced Recursive Feature Elimination

To apply FS, we follow the standard RFE procedure, iteratively removing the least important features to improve model performance and eliminate redundancy. Our integration combines SHAP-based feature importance with both standard and calibration objectives (Algorithm 1), creating a more robust selection process that considers both predictive power and model confidence. At each iteration, we train an XGBoost model either with the standard objective or with our calibration-aware loss function on the current subset of features. Unlike traditional RFE methods that rely on model coefficients or tree-based impurity scores, we compute feature importance using SHAP values, which better capture non-linear relationships and feature interactions. Features are ranked according to their average absolute SHAP values, and the least important feature is removed at each step until no further improvement in validation accuracy is observed.

Algorithm 1 RFE with SHAP Values

Require: Training data $(X_{\text{train}}, y_{\text{train}})$, validation data $(X_{\text{val}}, y_{\text{val}})$, XGBoost training function h (standard or custom), initial feature set $\mathcal{F}$, minimum number of features d_{min}
Ensure: Selected feature set $\mathcal{F}_{\text{selected}}$

Initialize best accuracy: best_acc $\leftarrow 0$
Initialize best feature set: $\mathcal{F}_{\text{selected}} \leftarrow \mathcal{F}$
improved $\leftarrow$ True
while $|\mathcal{F}| > d_{\text{min}}$ **and** improved = True **do**
  Train model h on $(X_{\text{train}}[:, \mathcal{F}], y_{\text{train}})$
  Compute SHAP values on training data
  For each feature $f \in \mathcal{F}$, compute mean absolute SHAP value: $\phi_f \leftarrow \mathbb{E}[|\text{SHAP}_f|]$
  Identify least important feature: $f_{\min} \leftarrow \arg\min_{f\in\mathcal{F}} \phi_f$
  Update feature set: $\mathcal{F}_{\text{temp}} \leftarrow \mathcal{F} \setminus \{f_{\min}\}$
  Evaluate accuracy on validation set $(X_{\text{val}}[:, \mathcal{F}_{\text{temp}}], y_{\text{val}})$
  Let val_acc$_{\text{temp}}$ be the validation accuracy
  if val_acc$_{\text{temp}}$ > best_acc **then**
    Update: best_acc $\leftarrow$ val_acc$_{\text{temp}}$
    Update: $\mathcal{F}_{\text{selected}} \leftarrow \mathcal{F}_{\text{temp}}$
    Update: $\mathcal{F} \leftarrow \mathcal{F}_{\text{temp}}$
  else
    improved $\leftarrow$ False
  end if
end while
return $\mathcal{F}_{\text{selected}}$

V. Evaluation Settings

Approaches. We compare our proposed approach against uncalibrated baseline and standard calibration methods. In all approaches, we utilize XGBoost as an underlying classification model. The approaches differ along two key dimensions:

- Calibration Awareness: whether the model incorporates calibration techniques either using the calibration-aware loss function or post-hoc isotonic regression.
- FS: whether FS is applied using SHAP-based.

The six evaluated approaches are describe in Table. I:

TABLE I
Summary of evaluated approaches.

Approach	Calibration	SHAP-RFE
Uncalibrated	None	×
Uncalibrated-SHAP	None	✓
Isotonic	Isotonic	×
Isotonic-SHAP	Isotonic	✓
Calib.-Aware Loss	Calib.-Aware Loss	×
Calib.-Aware Loss-SHAP	Calib.-Aware Loss	✓

1) *Uncalibrated (No Calibration, No FS):* A standard model trained without any calibration technique or FS. This serves as the uncalibrated baseline.
2) *Uncalibrated-SHAP:* A model trained without any calibration technique, but with FS performed using SHAP-based RFE. This scenario is used to isolate the effect of SHAP-guided FS in the absence of calibration.
3) *Isotonic:* A model trained with isotonic regression, which is a standard non-parametric calibration method that learns a step-wise function to map uncalibrated output probabilities to well-calibrated ones that better reflect true likelihood of outcomes. This approach quantifies the effectiveness of isotonic calibration over an uncalibrated model.
4) *Isotonic-SHAP:* A model trained with SHAP-RFE for FS, followed by calibration using isotonic regression. This setup assesses the combined effect of FS and isotonic regression.
5) *Calibration-aware:* A model trained using the calibration-aware loss function, without FS. This scenario quantifies the contribution of our loss to improving calibration compared to the *Uncalibrated* and *Isotonic* approaches.
6) *Calibration-aware-SHAP:* Our calibration-aware loss function combined with SHAP-RFE. This scenario allows us to evaluate the joint optimization of predictive performance and calibration through integrated training and informed FS.

The evaluations are conducted using a 5-fold cross-

TABLE II
MODEL PERFORMANCE COMPARISON (5-FOLD AVERAGES).
(BOLD INDICATES BEST, UNDERLINE INDICATES SECOND-BEST)

Dataset	Approach	Acc.	F1	Brier	ECE	Feat.
NF-UNSW-NB15	Uncalibrated	81.8	73.2	0.115	0.019	48.0
	Uncalibrated-SHAP	82.2	73.7	0.115	0.023	11.4
	Isotonic	82.5	**74.8**	0.113	0.011	48.0
	Isotonic-SHAP	**82.6**	74.5	0.112	0.011	11.4
	Calib.-aware	82.3	73.3	0.113	0.007	48.0
	Calib.-aware-SHAP	**82.6**	73.6	**0.111**	**0.005**	9.8
CIC-MalMem-2022	Uncalibrated	75.0	74.9	0.157	0.033	52.0
	Uncalibrated-SHAP	75.4	75.3	0.157	0.034	36.8
	Isotonic	75.5	75.4	0.154	0.023	52.0
	Isotonic-SHAP	75.5	75.4	0.154	0.023	36.8
	Calib.-aware	**75.7**	**75.6**	**0.152**	**0.010**	52.0
	Calib.-aware-SHAP	75.6	75.5	**0.152**	0.013	38.0

validation, employing a fixed data split of 70%, 10%, and 20% for training, validation and testing, respectively. The test sets are kept identical across all evaluated methods to ensure a fair comparison. For each fold, the trade-off parameter α in Eq. (1) is selected using the validation set by performing a search over $\alpha \in [0, 1]$ and retaining the value that minimizes the ECE on the validation split. This ensures that calibration quality is prioritized without sacrificing accuracy, and the best α value is carried forward to final testing.

Datasets. We conduct our evaluations on two publicly available ID datasets, namely, NF-UNSW-NB15 [24] and CIC-MalMem-2022 [25]. *NF-UNSW-NB15* is a NetFlow-based adaptation of the UNSW-NB15 dataset [26] that extends the original by incorporating NetFlow-specific features, resulting in a total of 53 features. The dataset comprises approximately 5.4% attack flows (nine different attack categories) and 94.6% benign flows. *CIC-MalMem-2022* targets memory-based malware detection (categorized into various classes) constructed from memory dumps and consists of 55 features and is balanced between benign and malicious samples.

VI. NUMERICAL RESULTS

Table II reports the performance of our approach and the benchmark approaches across the two datasets.

Predictive performance: In terms of accuracy, results show that all approaches show comparable performance, between 81.8 % and 82.6% on NF-UNSW-NB15, with a slight advantage for *Isotonic-SHAP* and *Calib-aware-SHAP*, and 75.0% and 75.7% on CIC-MalMem-2022, with a slight advantage for *Calib.-aware*. Similarly, in terms of F1-score, all approaches show very similar performance. Moreover, results show that applying isotonic regression (*Isotonic* and *Isotonic-SHAP*) enhances performance, albeit little, with respect to their respective uncalibrated counterparts (*Uncalibrated* and *Uncalibrated-SHAP*) across both datasets. Additionally, results show that our calibration-aware loss function, despite deprioritizing accuracy for calibration, achieves an equal or even slightly better performance than other approaches. For instance, in NF-UNSW-NB15, *Calib-aware-SHAP* shows an improved accuracy (82.6%) over *Uncalibrated-SHAP* (81.8%) and same accuracy as *Isotonic-SHAP*. Moreover, SHAP-guided FS enhances or maintains same level accuracy, despite showing a slight decrease in F1-score in some cases.

Model Calibration: We compare the approaches in terms of calibration metrics. Results show that both isotonic regression and the calibration-aware loss consistently improve calibration over uncalibrated models, with lower BS and ECE across both datasets. In particular, the calibration-aware approaches achieve the best calibration performance overall. On NF-UNSW-NB15, *Calib.-aware-SHAP* reduces the BS from 0.115 (Uncalibrated) to 0.111, marking a relative improvement of approximately 3.5%, while ECE improves from 0.019 to 0.005, corresponding to a 74% reduction. Similarly, on CIC-MalMem-2022, BS improves from 0.157 to 0.152 (a 3.2% improvement), and ECE reduces from 0.033 to 0.013 (60% reduction). These improvements highlight the effectiveness of explicitly optimizing for calibration, particularly in reducing ECE, which is critical for producing reliable probability estimates. Moreover, SHAP-guided FS preserves calibration, while yielding better calibrated and more compact models.

Number of features: In terms of FS, results show that SHAP-guided FS consistently reduces the number of features while maintaining or even improving performance. For the NF-UNSW-NB15 dataset, the number of features drops from 48 in the uncalibrated and calibration-aware baselines to as low as 9.8 (average number of features across folds) with *Calib.-aware-SHAP*, achieving nearly 80% reduction. Similarly, in the CIC-MalMem-2022 dataset, SHAP-based approaches reduce the feature set from 52.0 to around 36.8–38.0 features, corresponding to approximately 27-30% reduction, demonstrating the effectiveness of SHAP in identifying the most relevant features. These results confirm that SHAP-guided FS enables compact models without sacrificing predictive performance or calibration, which is beneficial for model interpretability and computational efficiency.

Calib.-aware-SHAP vs. Benchmark Approaches: We now specifically focus on comparing the performance of *Calib.-aware-SHAP* to its relevant counterparts. First, when compared to *Uncalibrated-SHAP*, results highlight the benefit of the calibration-aware loss. In NF-UNSW-NB15, *Calib.-aware-SHAP* slightly improves accuracy from 82.2% to 82.6% and substantially enhances calibration, reducing Brier score from 0.115 to 0.111 and ECE from 0.023 to 0.005, while also slightly reducing the feature set size from 11.4 to 9.8. Similarly, in CIC-MalMem-2022, it achieves comparable accuracy (75.6% vs. 75.4%) and improves calibration (Brier: 0.152 vs. 0.157; ECE: 0.013 vs. 0.034). Second, compared to *Calib.-aware* (without SHAP), results confirm the added benefit of SHAP-guided FS. *Calib.-aware-SHAP* reduces the feature set size significantly (from 48.0 to 9.8 in NF-UNSW-NB15, and from 52.0 to 38.0 in CIC-MalMem-2022) while maintaining comparable accuracy and calibration performance. Finally, when compared to *Isotonic-SHAP*, results show that *Calib.-aware-SHAP* achieves similar or better accuracy (82.6% vs. 82.6% in NF-UNSW-NB15, and 75.6% vs. 75.5% in CIC-MalMem-2022) while providing superior calibration, in terms

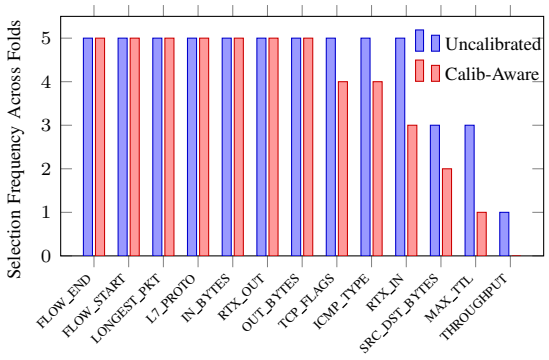

Fig. 1. Features selected across all folds for the NF-UNSW-NB15 dataset.

of ECE (0.005 vs. 0.011 in NF-UNSW-NB15 and 0.013 vs. 0.023 in CIC-MalMem-2022). These results confirm the effectiveness of combining calibration-aware training with SHAP-guided FS for building compact, accurate, and well-calibrated models. Across both datasets, the proposed *Calib.-aware-SHAP* approach achieves the best trade-off between predictive performance, calibration, and model compactness.

Selected Features: Figure 1 reports the selection frequency of features across the 5 folds for the NF-UNSW-NB15 dataset of *Uncalibrated-SHAP* and *Calib.-aware-SHAP*. The aim of this analysis is to investigate whether the calibration-aware loss influences the FS process. Results show that while both approaches consistently select the same core features, *Calib.-aware-SHAP* tends to select fewer secondary features across folds, indicating a stronger focus on the most relevant predictors. While this analysis is left for future work, we can consider that it suggests that incorporating calibration-awareness promotes more compact and stable feature subsets by de-prioritizing less impactful features.

VII. Conclusion

We proposed a calibration-aware loss function combined with SHAP-based Recursive Feature Elimination to jointly optimize predictive performance and model calibration while reducing feature set for the use case of Network Intrusion Detection. Experiments on two benchmark datasets show that our proposed approach improves calibration while maintaining predictive performance and identifying a subset of reliable core features. While effective, the proposed method is currently limited to the XGBoost model architecture. Future work will focus on generalizing the algorithm to other gradient-boosting models, such as LightGBM and CatBoost. We also plan to evaluate the approach on additional benchmark datasets and explore other SHAP-based FS methods beyond RFE. This will help assess computational feasibility and the method's ability to provide reliable confidence estimates and abstention capabilities in deployment for building more reliable and efficient NIDS models.

Acknowledgment

This work has been partially supported by the EUREKA CELTIC-NEXT project SUSTAINET-Advance, funded by the Swiss Innovation Agency Innosuisse. H. Fawaz and F. Ezzeddine are supported by the Swiss Government Excellence Scholarship (ESKAS) No. 2024.0474 and 2022.0547, respectively.

References

[1] Chou, D. et al. "survey on data-driven network intrusion detection." ACM Computing Surveys (CSUR) 54.9 (2021): 1-36.
[2] Tritscher, Julian, et al. "Evaluating feature relevance XAI in network intrusion detection." World Conference on Explainable Artificial Intelligence. Cham: Springer Nature Switzerland, 2023.
[3] Gaspar, Diogo, Paulo Silva, and Catarina Silva. "Explainable AI for intrusion detection systems: Lime and shap applicability on multi-layer perceptron." IEEE Access (2024).
[4] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
[5] Nascita, A., et al. "Improving performance, reliability, and feasibility in multimodal multitask traffic classification with XAI." IEEE Transactions on Network and Service Management 20.2 (2023): 1267-1289.
[6] Nascita, Alfredo, et al. "A survey on explainable artificial intelligence for internet traffic classification and prediction, and intrusion detection." IEEE Communications Surveys & Tutorials (2024).
[7] Wang, Huanjing, et al. "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods." Journal of Big Data 11.1 (2024): 44.
[8] Marcílio, WE., and Danilo ME. "From explanations to feature selection: assessing SHAP values as feature selection mechanism." 2020 33rd SIBGRAPI conference on Graphics, Patterns and Images. Ieee, 2020.
[9] Talpini, J. et al. "Enhancing trustworthiness in ML-based network intrusion detection with uncertainty quantification." Journal of Reliable Intelligent Environments 10.4 (2024): 501-520.
[10] Al-Masri, Eyhab. "Deciding When Not to Decide: Indeterminacy-Aware Intrusion Detection with NeutroSENSE."arXiv:2507.00003 (2025).
[11] Roponena, Evita, et al. "Towards a Human-in-the-Loop Intelligent Intrusion Detection System." Doctoral Cons/Forum@ DB&IS. 2022.
[12] Kim, Yeongwoo, György Dán, and Quanyan Zhu. "Human-in-the-loop cyber intrusion detection using active learning." IEEE Transactions on Information Forensics and Security (2024).
[13] Nixon, Jeremy, et al. "Measuring Calibration in Deep Learning." CVPR workshops. Vol. 2. No. 7. 2019.
[14] Neupane, Subash, et al. "Explainable intrusion detection systems (x-ids): A survey of current methods, challenges, and opportunities." IEEE Access 10 (2022): 112392-112415.
[15] Ayoub, Omran, et al. "Explainable artificial intelligence in communication networks: A use case for failure identification in microwave networks." Computer Networks 219 (2022): 109466.
[16] Asry, Chadia EL, et al. "Enhancing cybersecurity: A high-performance intrusion detection approach through boosting minority class recognition." PloS one 20.3 (2025): e0317346.
[17] Chen, Xuejiao, et al. "Explainable deep learning-based feature selection and intrusion detection method on the internet of things." Sensors (Basel, Switzerland) 24.16 (2024): 5223.
[18] Arreche, O. et al. "Let us Unveil Network Intrusion Features: Enhancing NIDS via XAI-based Feature Selection." (2024).
[19] Yacoubi, M. et al. "Explainable AI-Driven Feature Selection for Improved Intrusion Detection Systems in the Internet of Medical Things." IFIP International Conference on AI Applications and Innovations. Springer, Cham, 2025.
[20] Gazzan, M. et al. "Novel Ransomware Detection Exploiting Uncertainty and Calibration Quality Measures Using Deep Learning." Information 15.5 (2024): 262.
[21] Zhang, J. et al. "An ensemble-based network intrusion detection scheme with bayesian deep learning." ICC 2020-2020 IEEE International Conference on Communications. IEEE, 2020.
[22] Yang, T. et al. "Towards trustworthy cybersecurity operations using Bayesian Deep Learning to improve uncertainty quantification of anomaly detection." Available at SSRN 4609553 (2024).
[23] Birihanu, E. et al. "Enhancing Industrial Control Systems Security: Real-Time Anomaly Detection with Uncertainty Estimation." International Conf. on Discovery Science. Cham: Springer Nature Switzerland, 2024.
[24] Luay, Majed, et al. "Temporal Analysis of NetFlow Datasets for Network Intrusion Detection Systems." arXiv preprint arXiv:2503.04404 (2025).
[25] Carrier, Tristan. "Detecting obfuscated malware using memory feature engineering." (2021).
[26] Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems" IEEE Military Communications and Information Systems Conference, 2015.