Energy Cost of Enhancing Reliability of Machine Learning Models for Edge IoT Security

Hussein Fawaz[1,2,*], Omran Ayoub[1], Davide Andreoletti[1], Silvia Giordano[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*—Machine learning (ML) models are widely used for detecting malicious traffic in networked systems, including Internet of Things (IoT) environments. Recent work has shown that accuracy alone is insufficient, as models must also provide reliable and well-calibrated confidence estimates to avoid over-confident errors and unstable behavior in practical deployments. Improving model calibration typically relies on post-hoc techniques or additional training data, but both can increase model complexity and energy demand, which is in turn an important limitation for resource-constrained edge deployments. While the trade-off between predictive accuracy and energy consumption has been explored in prior work, the joint trade-off involving prediction reliability has received far less attention despite its critical importance for operational Intrusion Detection Systems (IDS). In this paper, we address this gap by investigating the relationship between predictive performance, reliability, and energy consumption in ML-based IDS for IoT networks. To this end, we develop an evaluation framework that applies post-hoc calibration techniques, namely, Platt scaling and isotonic regression, to ML classifiers and systematically quantifies their impact. Using real IoT traffic data, we evaluate models under varying training data sizes and measure predictive accuracy, calibration quality, and computational energy requirements. Experimental results show that post-hoc calibration improves confidence reliability considerably, with isotonic regression reducing calibration error (ECE) by over 95% and Platt scaling by 83%, while increasing inference energy by only 2–3%, demonstrating that reliability does not come at meaningful computational cost. Results further highlight that calibrated models achieve the same levels of predictive performance and reliability using up to 60–90% less training data, reducing both data needs and energy demand.**

***Index Terms*—Model Calibration, Energy Consumption, Reliability, Intrusion Detection, IoT Security**

I. Introduction

The rapid proliferation of Internet of Things (IoT) devices has significantly expanded the attack surface of IoT networks [1], [2]. Consequently, developing intrusion detection systems (IDS) capable of detecting attacks with high accuracy is a pressing requirement for ensuring robust IoT security.

Machine Learning (ML) has emerged as a key enabler for robust IDS, offering strong predictive performance through models that learn to distinguish between benign and malicious activities from network data [3], [4]. In IoT environments, these ML models are typically deployed at the network edge, to detect and act on packets as early as possible, for example to prevent malicious traffic from propagating through the network. However, edge devices operate under strict limitations in memory, computation, and energy consumption [5]. Consequently, developing IDS that can function effectively under tight resource constraints and low energy availability represents another pressing requirement. In response, research efforts have increasingly focused on reducing ML model complexity and designing lightweight ML approaches while maintaining the desired predictive performance [6]–[8].

While these efforts have quantified the trade-off between model size, energy demand, and predictive accuracy [9], [10], and have optimized models for feasible edge deployment, comparatively little attention has been given to other essential dimensions such as prediction reliability. Recent research in ML-based IDS has indeed demonstrated that high accuracy alone is insufficient to guarantee dependable decision-making [11]–[13]. In other words, a model may appear effective while masking unreliable behavior in the form of poorly calibrated predictions, where calibration refers to the alignment between the probabilities predicted by the model and the true outcome likelihoods (explained later in more detail). In this context, reliability requires that a model's confidence estimates be aligned with its actual performance, which is particularly critical in security-sensitive applications [14]–[16].

A variety of techniques, including Platt scaling, isotonic regression, and the incorporation of supplemental training data, can be employed to improve model calibration. Despite their effectiveness, these approaches may increase model complexity and introduce additional computational or energy overheads [17]. This creates a critical trade-off between achieving reliable confidence estimates and preserving efficiency in terms of, e.g., energy demand and model size, for continuous edge deployments.

In this context, ML models for IDS must satisfy multiple requirements, including high predictive performance, strong reliability, and minimal energy consumption [18], [19]. As previously mentioned, prior studies targeting energy reduction typically focus on compression or architectural simplification [20], [21] without examining the reliability of the resulting models, while research on calibration rarely evaluates its computational or energy costs [22], [23]. As a result, the combined effect of calibration on both reliability and energy consumption in resource-constrained IDS remains largely unexplored.

In this work, we address this gap by examining the energy cost of enhancing model reliability, measured through model calibration, in ML-based IDS for IoT networks. We develop an evaluation framework that applies post-hoc calibration

ISBN: 978-3-903176-79-9

techniques, namely, Platt scaling and isotonic regression, to a classifier trained on a given dataset and systematically measures predictive performance, prediction reliability, and energy consumption. Our framework enables a detailed analysis of the energy–performance–reliability trade-offs inherent to edge-deployed IDS. Accordingly, we investigate the following research questions (RQs):

- *RQ1: How does improving model reliability in ML-based IoT intrusion detection systems affect inference energy consumption, and can post-hoc calibration provide more reliable confidence estimates without degrading predictive accuracy?*
- *RQ2: To what extent does the application of calibration methods reduce the volume of training data required to achieve specified levels of predictive performance and reliability?*
- *RQ3: Which calibration technique provides the most favorable trade-off among energy consumption, reliability, and predictive performance in constrained edge settings?*

To address these RQs, we implement the proposed framework on two representative IoT intrusion detection datasets and apply both Platt scaling and isotonic regression to a baseline classifier. We then systematically assess the resulting models along three dimensions, predictive performance, calibration quality, and energy consumption, to quantify the trade-offs introduced by calibration in realistic edge-deployment scenarios. To the best of our knowledge, our work is the first to quantify the energy cost of enhancing ML model reliability and to systematically assess how different calibration approaches influence the energy demand of ML-based IDS.

The remainder of this paper is structured as follows. Section II discusses related work. Section III presents our proposed framework. Section IV presents the experimental settings and discusses experimental results. Finally, Section V concludes the paper.

II. Related Work

Numerous studies in network security have focused on reducing model complexity or model size, both of which are key factors in influencing energy consumption, to ensure feasible deployment in resource-constrained environments [7], [8], [24]–[27]. Particular attention has been given to employing knowledge distillation to produce lightweight models for IDS while maintaining the desired predictive performance, as in [24], where the authors apply a student–teacher architecture and achieve substantial reduction in the number of model parameters and enhanced inference speed and detection accuracy. Similarly, [25] applies knowledge distillation for IoT traffic classification and shows how compact student networks can closely match the accuracy of larger teacher models. Model compression and quantization techniques have also been applied to IDS pipelines. For instance, [26] proposed a lightweight hybrid deep neural network architecture that integrates advanced feature engineering and quantization techniques to enhance detection performance of models with substantially smaller sizes, and hence suitable, for deployment on resource-constrained IoT devices. Moreover, [27] shows that model compression techniques can reduce energy use during inference and reduce model size while maintaining target accuracy levels. Additionally, in [28], authors introduce a constrained variational autoencoder designed to produce compact latent representations so that intrusion detection models receive lower-dimensional features, reducing resource usage on IoT devices. Their results show that their approach improves detection accuracy while keeping runtime extremely low and model size very small, enabling practical deployment on IoT IDS devices.

More recent works have focused primarily on the energy consumption of the ML models, always with the aim to balance predictive performance and energy consumption [20], [21], [29]–[32]. For instance, in [29], authors quantify the energy costs of on-device ML-based intrusion detection and demonstrate direct trade-offs between algorithm choice and energy consumption. The authors show that k-NN inference energy could rise significantly as dataset size increases, highlighting the importance of lightweight models for IoT devices. Moreover, the work in [30] investigates energy-aware anomaly detection in IoT-integrated systems, showing that sequential models can be optimized for both accuracy and efficiency, achieving high detection performance with very low energy usage and minimal detection delay. [31] propose a deep reinforcement learning-based IDS integrated with a self-adaptive control loop that dynamically adjusts detection and response strategies to balance security and resource usage, achieving high detection performance while keeping CPU and energy consumption within practical limits. Moreover, the authors in [32] propose an approach that combines traditional ML models with knowledge distillation and adaptive optimization to dynamically adjust model complexity under real-time energy constraints. Their results show reduced energy consumption while maintaining high detection performance.

Taken together, these works demonstrate that research on IoT and IDS has largely centered on optimizing model complexity, size, and energy consumption primarily in relation to predictive performance. However, as previously highlighted, performance alone is insufficient to ensure accurate and dependable IDS deployment, but instead the reliability of model outputs is equally critical, particularly in settings where decisions depend on well-calibrated confidence estimates. Our work complements these efforts by shifting attention toward model reliability and by examining how calibration techniques influence both the dependability of predictions and the associated energy cost of achieving such reliability.

Enhancing model reliability for IDS is primarily tackled through model calibration techniques and uncertainty quantification (UQ) [12]–[15], [23], [34], [35]. For instance, [15] propose an approach that integrates a calibration-oriented loss function into XGBoost together with SHAP-based recursive feature elimination to improve both predictive performance and the quality of confidence estimates. Moreover, [23] address the challenge of reliability in distributed settings by proposing a federated learning framework that enhances model

calibration. On the UQ side, [13] advocate integrating UQ into IDS to mitigate overconfidence and better handle open-set and zero-day attack scenarios. Similarly, [33] employ Bayesian autoencoders to provide uncertainty-aware anomaly detection, offering a measure of confidence for each security alert. In [34], the authors introduce an entropy-based reliability filter for IDS in IoT, leveraging prediction uncertainty to improve robustness against unseen attack types.

Although these works have significantly advanced IDS in terms of predictive performance, energy consumption, and model reliability, they show no particular focus on the additional model complexity or energy demand introduced by such techniques. Instead, our work focuses on characterizing the trade-offs among predictive performance, reliability, and energy consumption, and by explicitly quantifying the energy cost of improved model reliability.

III. Problem Statement and Methodology

We aim to quantify the trade-offs among predictive performance, calibration quality, and energy consumption in ML-based intrusion detection models designed for IoT edge environments. To this end, we formulate the problem[1] as follows. Given a baseline classifier f and a calibration strategy C, our objective is to apply the calibration to the model and evaluate the resulting calibrated model $C(f)$ along three dimensions: (i) predictive performance, (ii) reliability of the associated confidence estimates, and (iii) computational energy cost incurred during training, calibration, and inference.

We consider metrics measuring predictive performance, predictions reliability and energy demand. The predictive performance is assessed using the macro F1-score, which captures the classifier's ability to correctly identify both benign and attack classes. Model reliability is quantified through the Expected Calibration Error (ECE) and the Brier score. The ECE measures the discrepancy between predicted confidence, and empirical accuracy across M bins, and is defined as:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right| \tag{1}$$

where B_m is the set of samples falling into bin m, n is the total number of samples, $\mathrm{acc}(B_m)$ is the accuracy of samples in that bin, and $\mathrm{conf}(B_m)$ is their average predicted confidence. ECE indicated whether its confidence estimates can be trusted. The Brier score, on the other hand, captures the mean squared error of probabilistic predictions:

$$\mathrm{Brier} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2 \tag{2}$$

where p_i is the predicted probability for sample i and $y_i \in \{0, 1\}$ is the true label. The Brier score penalizes both overconfidence and underconfidence, indicating whether a model's probability estimates reliably reflect reality.

[1]Code can be found at https://github.com/hussein-fawaz/Energy-Cost-of-Enhancing-Reliability-of-Machine-Learning-Models-for-Edge-IoT-Security

Regarding energy consumption, we monitor the energy footprint of the machine used to train the ML models and perform the inference. Specifically, we quantify the CPU and RAM energy used during training, the energy required for the calibration procedure itself, and the inference energy. We rely on the pyRAPL [38] toolkit for this aim.

To conduct our experiments, we consider a labeled dataset and partition it into a training set $\mathcal{D}_{\mathrm{train}}$, a calibration set $\mathcal{D}_{\mathrm{cal}}$, and a test set $\mathcal{D}_{\mathrm{test}}$. The baseline classifier f is first trained on $\mathcal{D}_{\mathrm{train}}$ to learn a mapping from input observations to intrusion detection labels. A post-hoc calibration strategy C is then fitted on the model's output scores using $\mathcal{D}_{\mathrm{cal}}$, yielding a calibrated model $C(f)$. We adopt two post-hoc calibration strategies:

Platt Scaling: A parametric logistic regression model fitted to the model's output probabilities. It provides a smooth and computationally efficient mapping from the model's raw confidence scores to calibrated probabilities, and is suitable when the relationship between logits and true probabilities approximates a sigmoid curve, meaning the model's confidence errors follow a smooth S-shaped pattern, where probabilities gradually increase with evidence. However, it may be limited when probabilities deviate significantly from that form.

Isotonic Regression: A non-parametric, monotonic mapping between uncalibrated and true probabilities that learns a flexible stepwise function, offering more capability to model intricate calibration curves when sufficient data is available. In practice, this means it does not assume any specific shape and can adapt to whatever pattern exists between predicted confidence and true likelihood, making it easier to capture complex or irregular calibration structures. It is particularly appropriate when model miscalibration is irregular or non-sigmoidal (i.e., when the relationship between predicted probabilities and observed outcome frequencies cannot be well approximated by a smooth S-shaped curve and instead exhibits plateaus, abrupt changes, or locally varying slopes across confidence regions), which is often the case in highly heterogeneous IoT traffic distributions. For this reason, isotonic regression is expected to better capture calibration structure under varying training sizes.

The uncalibrated model f itself serves as the baseline against which all calibrated configurations are evaluated. We analyze how f and $C(f)$ behave across different data sampling levels, allowing us to quantify the extent to which calibration improves reliability and its impact on both predictive performance and energy consumption, thus addressing RQ1. Then, by evaluating performance and reliability as the amount of training data varies, we quantify the extent to which calibration can reduce data requirements, hence addressing RQ2. Finally, by contrasting Platt and isotonic calibration, we identify which method offers the best energy–reliability–performance trade-off, thus addressing RQ3.

IV. Experimental Results

This section discusses the experimental results. We evaluate our framework using two datasets, with our analysis primarily focused on the *CICIoT-2023* dataset [37], while

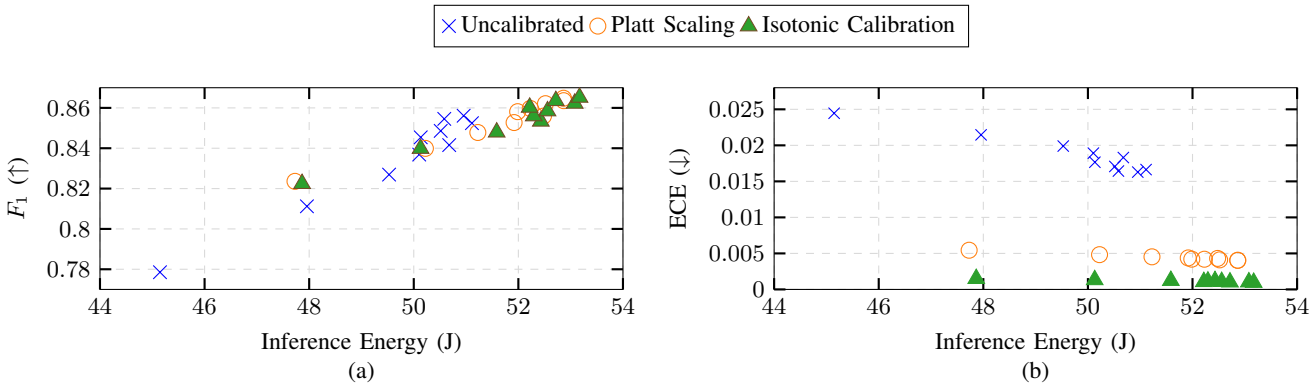


Fig. 1: (a) Predictive performance (F1) and energy demand for inference (J); (b) calibration quality (ECE) and energy demand for inference (J).

complementary results on the second dataset *(Edge-IIoT* [39]) are reported. Both datasets provide a realistic mix of benign traffic and diverse IoT attacks. Specifically, the CICIoT-2023 dataset contains over 46 million records of labeled network flows collected from heterogeneous IoT devices, encompassing both normal traffic and attack categories such as DDoS, DoS, Mirai, Spoofing, and Bruteforce variants. Each record reports standard NetFlow-style statistical features, such as packet counts, byte counts, inter-arrival times, and flow durations, along with protocol and flag information. A complete description of all feature fields can be found in [37]. Prior to training, we transform all categorical features to label-encoded features while continuous features are standardized. Experiments are conducted on an Intel-based workstation (Intel i9-14900HX, 32 GB RAM, Ubuntu 24.04) to ensure consistent energy profiling across runs. We first partition the dataset into a training set $\mathcal{D}_{\text{train}}$, a calibration set $\mathcal{D}_{\text{cal}}$, and an independent test set $\mathcal{D}_{\text{test}}$ to ensure strict separation between training, calibration, and evaluation. The base classifier used throughout all experiments is a Random Forest (RF), chosen for its simplicity, robustness, and strong performance on tabular intrusion detection data, making it an appropriate baseline for studying energy–reliability interactions. However, the overall pipeline remains agnostic to the deployed model and can be applied to any classifier. To understand how performance scales with data volume, we then vary the training set size from 10% to 100% of the available training data. For each fraction, we perform five random resamplings and then embed the entire process in a five-fold outer cross-validation. Calibration is performed exclusively on $\mathcal{D}_{\text{cal}}$, and the uncalibrated and calibrated models are subsequently evaluated on $\mathcal{D}_{\text{test}}$ in terms of predictive performance, calibration quality, and energy consumption.

We structure the discussion around five key aspects corresponding to our RQs: (1) we examine the energy–performance trade-offs by jointly assessing predictive performance, calibration reliability, and the energy required to achieve them; (2) we study the impact of training data fractions by varying the size of the training set and observing how this affects performance, reliability, and energy demand; (3) we analyze the energy breakdown by observing the energy consumption during inference to understand how much energy is required to reach a given performance level; (4) we investigate calibration reliability patterns by comparing how the different calibration methods consistently improve alignment between predicted and true probabilities; and (5) we evaluate the generality of our findings by validating the proposed analysis on an additional dataset.

A. *Energy–Performance-Reliability Trade-Off*

Figure 1(a) and (b) illustrate, respectively, the trade-off between the predictive performance measured in terms of F1 and the energy demand for inference (in Joules), and between model calibration, measured in terms of ECE, and the energy demand for inference. These comparisons are shown across the three settings: *uncalibrated*, *Platt Scaling Calibration*, and *Isotonic Calibration*. Each method is represented by ten points, corresponding to experiments conducted using different fractions of the training data (from 10% to 100%)[2].

Starting with Figure 1(a), results show that most configurations cluster toward the high F1 region (above 0.84), with inference energy ranging between 50–53 J. A closer examination reveals that, in the majority of cases, the uncalibrated model attains an F1 comparable to that of the calibrated variants while requiring up to 6% less energy. For example, several uncalibrated configurations reach an F1 of 0.85 with an inference energy of approximately 51 J, whereas the calibrated models achieve similar F1s (between 0.85 and 0.86) but require around 52–53 J for inference. Taken in isolation, these results would suggest a preference for the uncalibrated model, as it offers essentially the same predictive performance at a consistently lower energy cost. This indicates that, in terms of the F1–energy trade-off alone, calibration provides only marginal advantage, if any. However, this perspective is incomplete: calibration affects not only accuracy but also the reliability of predicted probabilities.

Moving to Figure 1(b), we observe that the calibration variants achieve a near-optimal performance in terms of ECE, with a slight advantage for isotonic regression (0.001) over Platt scaling (0.005). Both calibrated models operate within an inference energy range of approximately 48–53 J. On the

[2]We investigate the impact of the fraction of training data in next subsection.

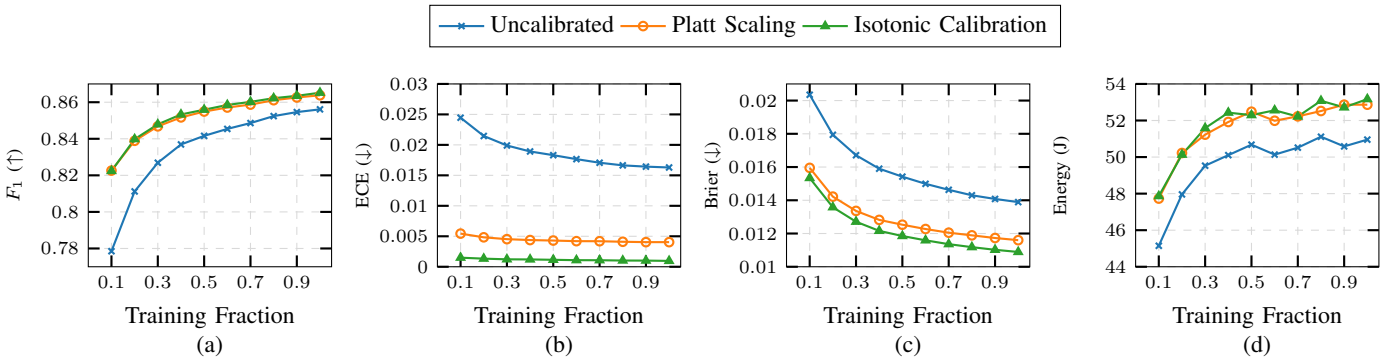


Fig. 2: Effect of training size on (a) predictive performance (F1-score), (b–c) model reliability measured by ECE and Brier score, respectively, and (d) energy consumption.

contrary, the uncalibrated model exhibits substantially higher ECE values, typically around 0.020 and reaching up to 0.025, while offering at most the previously noted 6% reduction in energy demand and, in some cases, no energy savings at all. These results show that calibration can significantly reduce ECE, yielding a more reliable model, for only a modest increase in energy demand. This finding highlights a crucial insight: models with similar predictive performance can differ significantly in terms of calibration. Consequently, relying solely on accuracy–energy trade-offs risks selecting models that are efficient but poorly calibrated. When reliability is accounted for, calibrated models offer more reliable probability estimates for marginal energy overhead. These results directly contribute to addressing RQ1, showing that improving model reliability can be achieved with only a modest and comparable energy increase without compromising predictive performance.

Comparing the calibration approaches, results show that isotonic calibration improves reliability by a large magnitude, reducing ECE from roughly 0.02 to 0.001 on average while maintaining similar F1 values (0.84–0.86) and requiring less than 2 J of additional inference energy. Platt scaling attains an ECE around 0.004 with slightly lower energy overhead than isotonic regression.

B. Impact of Training Size

We now turn our attention to the impact of the fraction of the training dataset on predictive performance, calibration quality, and inference energy demand. Figure 2 summarizes these effects, showing how varying the proportion of training data influences F1-score (Fig. 2 (a)), ECE (Fig. 2 (b)), the Brier score (Fig. 2 (c)), and the energy required for inference (Fig. 2 (d)).

Examining Figure 2(a) and (b), the results show that, as expected, using larger fractions of the training dataset improves both the F1 and the ECE. Notably, the calibration techniques achieve a given predictive performance with substantially less data than the uncalibrated model. For example, both calibration methods reach an F1 of 0.85 using only 30% of the training data, whereas the uncalibrated model requires roughly 70% to obtain the same performance. Similarly, achieving an F1

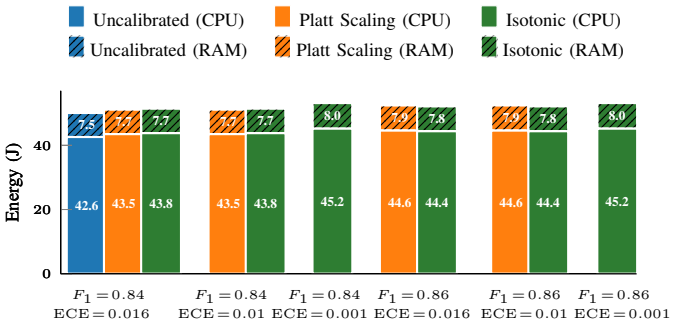
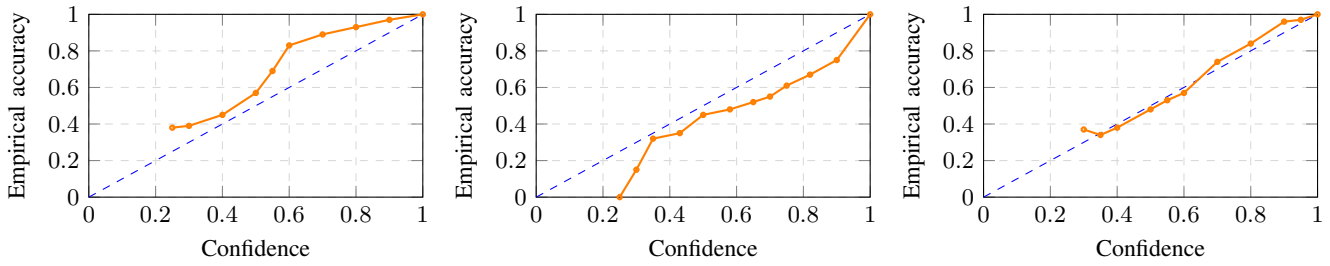


Fig. 3: Energy demand for meeting predictive and reliability (F1 and ECE) performance targets, showing also the breakdown across CPU and RAM. Each bar group corresponds to the lowest-energy configuration meeting the target performance.

of 0.84 requires approximately 50% of the data for the uncalibrated model, while the calibrated models (Platt scaling and isotonic calibration) attain this level using only about 20%.

The trend shows that both calibrated methods produce well-calibrated models even when trained on relatively small fractions of the dataset. Moreover, while the ECE of the uncalibrated model improves as more data is used for training, it never meets that of the calibrated models. For instance, an ECE below 0.02 is attainable with just 10% of the data using isotonic regression, whereas the uncalibrated model requires nearly the full training set to approach comparable reliability. A similar pattern is observed with the Brier score, which is shown in Figure 2(c). The uncalibrated model achieves its lowest Brier score only when trained on the full dataset, whereas the calibrated approaches reach comparable or better calibration using just 20% of the training data. This result highlights that post-hoc calibration not only improves probabilistic alignment but can also compensate for reduced training data, thereby lowering computational costs at training time and significantly lessening the model's dependence on extensive training data, which is a crucial property in IoT edge settings. This directly supports RQ2 by showing that calibration can reduce training-time energy and computation without compromising inference quality.

Figure 2(d) shows that inference energy rises moderately as more training data is used, that is, from about 45 J at 10% up to roughly 51–53 J at full data, regardless of



(a) Uncalibrated (b) Platt-Calibrated (c) Isotonic-Calibrated

Fig. 4: Reliability diagrams under different calibration methods: (a) uncalibrated, (b) Platt scaling, and (c) isotonic regression. The diagonal dashed line indicates perfect calibration.

whether calibration is applied. However, the difference in the energy demand required by the two calibration methods and the uncalibrated model remain marginal across all fractions, indicating that calibration does not introduce any meaningful inference-time energy overhead. Instead, its contribution lies in achieving improved reliability and performance with far less data. This observation connects to the next section, where we evaluate how much energy is required to meet concrete accuracy–reliability targets.

C. Energy Demand for Accuracy–Reliability Targets

Having analyzed the trade-offs between predictive performance, calibration quality, and energy consumption, we now focus on a practical scenario in which a target level of performance is required. In this section, we identify a representative operating point, defined by a desired F1 and ECE, and examine the energy consumption associated with achieving this target under different calibration settings. This analysis allows us to quantify the overhead introduced by calibration and show how costly it is, in energy terms, relative to the uncalibrated baseline.

We consider five operating points defined by combinations of target F1 levels (0.84 and 0.86) and calibration performance (ECE = 0.016, 0.010, and 0.001). Figure 3 reports the total inference energy required in these cases while also showing the breakdown across CPU and RAM components. For each case, we identify the lowest-energy configuration that meets or exceeds the required target performance, per approach. An absence of an approach for a given target performance indicates that the approach was unable to meet the requirement.

Results show that achieving a well-calibrated model often incurs only a marginal energy increase. For instance, moving from ECE = 0.016 to ECE = 0.01 at F1 = 0.84 requires an additional energy of around 1 J (from 42.6 J to 43.5 J). Furthermore, when targeting an ECE of 0.001 at the same F1 level, only isotonic calibration is able to meet this requirement, and it does so with the same additional energy cost relative to the uncalibrated configuration. Moving to a more stringent requirement (e.g., F1-score of 0.86 and ECE of 0.01 and F1-score of 0.86 and ECE of 0.001), we notice that these requirements can be met for negligible additional energy cost. The total inference energy demand increases linearly with training size ($\approx$ 45 J $\rightarrow$ 53 J) as shown in Figure 2(d), yet the improvement in reliability is disproportionately large (ECE $\downarrow$ from 0.024 to 0.001). This suggests a favorable energy–reliability frontier where small additional energy yields significant reliability gains. Importantly, while uncalibrated models may appear more energy efficient, their lower reliability metrics make them less suitable for real-world IoT security systems that depend on stable confidence estimates, thereby addressing RQ1. Finally, the breakdown further reveals that RAM contributes approximately 15% of the total energy, with CPU accounting for roughly 85% of total inference energy.

D. Reliability Evaluation and Energy Demand for Model Training

We now look at two complementary aspects of our analysis, namely, model reliability across confidence measures and energy needed for model training and calibration. Checking reliability helps ensure that the models give stable and trustworthy results in different situations while understanding energy use highlights the energy cost of keeping these models accurate and up to date.

Figure 4 shows the reliability diagrams, which visualize the relationship between predicted confidence and empirical accuracy across confidence bins. For each bin, the empirical accuracy corresponds to the fraction of correct predictions among samples with similar confidence values. The diagonal dashed line represents perfect calibration, where predicted probabilities match observed correctness (e.g., predictions with confidence 0.7 are accurate 70% of the time). Points above the diagonal indicate underconfidence, whereas points below indicate overconfidence. For clarity of presentation, we display a separate diagram for each method. Figure 4(a) shows that the uncalibrated model demonstrates clear underconfidence, with most points lying above the diagonal, indicating that predicted probabilities underestimate true accuracy. Platt scaling (Fig. 4(b)) improves calibration but tends to produce overconfident predictions, particularly in higher probability bins. In contrast, isotonic regression yields (Fig. 4(c)) the closest alignment to the ideal calibration line, confirming its superior ability to map raw probabilities to true likelihoods. These visual patterns mirror the quantitative reliability metrics: isotonic calibration yields the lowest ECE ($\approx$ 0.001) and Brier score ($\sim$ 0.011), aligning most closely with the ideal diagonal. Platt scaling shows moderate improvement

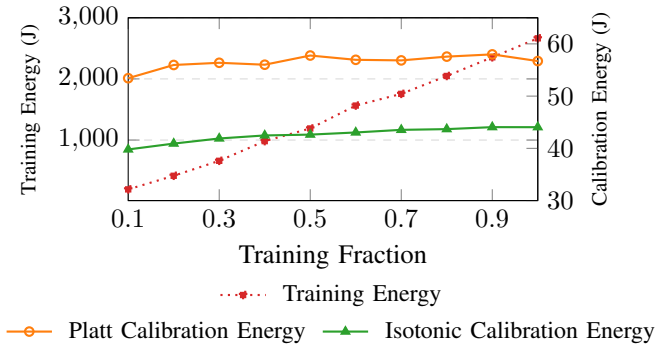


Fig. 5: Energy required for model training (left y-axis) and for calibration using Platt scaling and isotonic regression (right y-axis). The x-axis shows the fraction of the training dataset used.

(ECE $\approx$ 0.004), while the uncalibrated model remains systematically underconfident. Together with the energy analysis, this analysis further confirms that isotonic regression achieves the most balanced energy–reliability–performance trade-off, effectively delineating the desired energy–reliability frontier outlined in RQ3. More generally, these results show that calibration provides a practical and low-overhead pathway to more trustworthy IDS behaviour, as it substantially improves the alignment between predicted and true probabilities while keeping energy costs minimal.

Figure 5 reports the training and the calibration energy demand per training fraction. The results show, as expected, that training energy increases almost linearly with the fraction of training data used. The dotted red curve shows this progression clearly, rising steadily from around 200 J at 10% of the data to nearly 2,400 J at 90%. Moreover, the results show that the energy required for calibration, independent of the calibration method, is in the range of 40–60 J, two orders of magnitude less than that of training. Indeed, these values are similar in scale to the energy demand for inference, underscoring that calibration adds only a minimal overhead. Moreover, the results support the finding that using a small fraction of training data combined with calibration can yield better predictive performance than using substantially more training data without calibration. From an energy-efficiency standpoint, this means that high-quality, well-calibrated predictions can be obtained at a fraction of the training cost. For example, training with 20–30% of the data plus calibration consumes about 1,000 J for training and 50 J for calibration, while training with 80–90% of the data without calibration requires more than double the energy yet may still produce inferior reliability. This demonstrates that calibration offers a highly efficient path to improving model trustworthiness without incurring the significant energy cost of additional training.

E. Evaluation on a Second Dataset

To assess the robustness of our findings, we conduct the same evaluation on a second dataset, namely the Edge-IIoT dataset [39]. This experiment is intended to examine whether the observed energy–reliability trade-offs persist under different IoT traffic characteristics and attack distributions. The results obtained on Edge-IIoT, as shown in Figure 6, confirm that the main trends observed on CICIoT-2023 remain consistent, particularly the substantial improvement in reliability achieved through post-hoc calibration with only marginal inference-time energy overhead. We notice that the predictive performance, in terms of F1, follows a consistent and nearly linear trend with inference energy, ranging from approximately 0.85 to 0.88 across all configurations, with calibrated models achieving slightly higher F1 values than the uncalibrated baseline at comparable energy levels. In terms of reliability, calibration yields substantial improvements: isotonic regression reduces ECE from roughly 0.09 in the uncalibrated model down to about 0.035, corresponding to an improvement of over 50%, while Platt scaling achieves intermediate ECE values around 0.060. These results closely mirror the trends observed on CICIoT-2023, confirming that post-hoc calibration consistently enhances reliability with only marginal inference-time energy overhead across datasets.

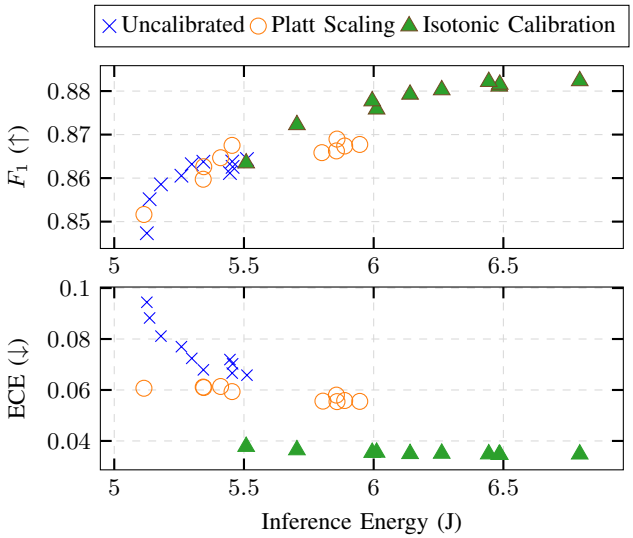


Fig. 6: (a) Predictive performance (F1) and energy demand for inference (J); (b) calibration quality (ECE) and energy demand for inference (J) for the Edge-IIoT dataset.

V. Conclusion

In this work, we conducted a systematic experimental evaluation of ML-based IoT intrusion detection systems to quantify the trade-offs between predictive performance, reliability, and energy consumption. Using post-hoc calibration methods, namely, Platt scaling and isotonic regression, we systematically analyzed their impact on energy demand and model reliability across varying training data volumes. Across all analyses, we observe that model calibration plays a crucial role in achieving reliability without incurring additional inference energy costs. Moreover, calibration effectively reduces the need for large training datasets while enhancing confidence alignment, allowing enhanced performance with smaller models and lower energy demand. These findings establish an empirical link between calibration quality and energy consumption, showing that reliability can often be improved

without a corresponding rise in energy demand, and highlight calibration as a practical and low-overhead mechanism for sustainable and trustworthy IoT intrusion detection at the edge.

While our results consistently demonstrate these advantages, they are obtained under a specific experimental setup involving a single model family and a fixed hardware platform. Different classifiers and hardware architectures may exhibit distinct calibration dynamics and energy profiles. Nevertheless, the observed trends suggest that the energy–reliability relationship is primarily driven by the lightweight nature of post-hoc calibration rather than model-specific architectural complexity. Future work will therefore extend this analysis to multiple model families and heterogeneous edge hardware platforms, as well as explore calibration-aware training strategies that further reduce energy consumption while preserving predictive performance and reliability.

ACKNOWLEDGMENT

This work has partially been supported by the Swiss Government Excellence Scholarship (ESKAS) No. 2024.0474 and by Innosuisse, the Swiss Innovation Agency, through the innovation project SUSTAINET (No. 119.588 INT-ICT), carried out within the EUREKA Cluster CELTIC-NEXT under the project SUSTAINET-Advance.

REFERENCES

[1] Alwahedi, Fatima, et al. "Machine learning techniques for IoT security: Current research and future vision with generative AI and large language models." Internet of Things and Cyber-Physical Systems 4 (2024): 167-185.

[2] Sarker, Iqbal H., et al. "Internet of things (iot) security intelligence: a comprehensive overview, machine learning solutions and research directions." Mobile Networks and Applications 28.1 (2023): 296-312.

[3] Altulaihan, Esra, et al. "Anomaly detection IDS for detecting DoS attacks in IoT networks based on machine learning algorithms." Sensors 24.2 (2024): 713.

[4] Cerasuolo, Francesco, et al. "Adaptable, incremental, and explainable network intrusion detection systems for internet of things." Engineering Applications of Artificial Intelligence 144 (2025): 110143.

[5] Hossain, Md Alamgir. "Deep learning-based intrusion detection for IoT networks: a scalable and efficient approach." EURASIP Journal on Information Security 2025.1 (2025): 28.

[6] Khanday, S Ahmad, et al. "Implementation of intrusion detection model for DDoS attacks in Lightweight IoT Networks." Expert Systems with Applications 215 (2023): 119330.

[7] Wang, Z, et al. "A novel lightweight IoT intrusion detection model based on self-knowledge distillation." IEEE Internet of Things Journal (2025).

[8] Anurupam, Kumar, et al. "Lightweight representation learning for network traffic towards malicious traffic detection in edge devices." Journal of Information Security and Applications (2025): 104186.

[9] Yuan, Xinwei, et al. "A simple framework to enhance the adversarial robustness of deep learning-based intrusion detection system." Computers & Security 137 (2024): 103644.

[10] Sajid, Muhammad, et al. "Enhancing intrusion detection: a hybrid machine and deep learning approach." Journal of Cloud Computing 13.1 (2024): 123.

[11] Wali, Syed, et al. "Explainable AI and random forest based reliable intrusion detection system." Computers & Security (2025): 104542.

[12] Nascita, Alfredo, et al. "A survey on explainable artificial intelligence for internet traffic classification and prediction, and intrusion detection." IEEE Communications Surveys & Tutorials (2024).

[13] Talpini, Jacopo, et al. "Enhancing trustworthiness in ML-based network intrusion detection with uncertainty quantification." Journal of Reliable Intelligent Environments 10.4 (2024): 501-520.

[14] Nascita, Alfredo, 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.

[15] Fawaz, Hussein et al. "Towards Better-Calibrated ML Models for Reliable Network Intrusion Detection via Calibration-Aware SHAP-based Feature Selection," in Proc. 1st IEEE Int. Workshop on Generative and eXplainable Artificial Intelligence for Networking (GenXNet), WiMob (2025).

[16] Sayadi, Hossein, et al. "Redefining trust: Assessing reliability of machine learning algorithms in intrusion detection systems." 2024 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2024.

[17] Balanya, Sergio A., et al. "Adaptive temperature scaling for robust calibration of deep neural networks." Neural Computing and Applications 36.14 (2024): 8073-8095.

[18] Jaddoa, Ali, et al. "Toward Scalable and Sustainable Detection Systems: A Behavioural Taxonomy and Utility-Based Framework for Security Detection in IoT and IIoT." IoT 6.4 (2025): 62.

[19] Ahakonye, Love Allen Chijioke, et al. "Eco-Secure SCADA: Towards Machine Learning Reliability for Green Cybersecurity." 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, 2025.

[20] Tekin, Nazli, et al. "A review of on-device machine learning for IoT: An energy perspective." Ad Hoc Networks 153 (2024): 103348.

[21] Jamshidi, Saeid, et al. "Evaluating machine learning-driven intrusion detection systems in IoT: Performance and energy consumption." Computers & Industrial Engineering 204 (2025): 111103.

[22] Yousef, Waleed A., et al. "Classifier calibration: with application to threat scores in cybersecurity." IEEE Transactions on Dependable and Secure Computing 20.3 (2022): 1994-2010.

[23] Talpini, Jacopo, et al. "A Federated Approach to Enhance Calibration of Distributed ML-Based Intrusion Detection Systems."

[24] Wisanwanichthan, Treepop, and Mason Thammawichai. "A lightweight intrusion detection system for IoT and UAV using deep neural networks with knowledge distillation." Computers 14.7 (2025): 291.

[25] Abbasi, Mahmoud, et al. "Unleashing the potential of knowledge distillation for IoT traffic classification." IEEE Transactions on Machine Learning in Communications and Networking 2 (2024): 221-239.

[26] Misrak, S F, and Henock M M. "Lightweight intrusion detection system for IoT with improved feature engineering and advanced dynamic quantization." Discover Internet of Things 5.1 (2025): 97.

[27] Umar, Hafiz Gulfam Ahmad, et al. "Energy-efficient deep learning-based intrusion detection system for edge computing: a novel DNN-KDQ model." Journal of Cloud Computing 14.1 (2025): 32.

[28] Dinh, Phai Vu, et al. "Constrained twin variational auto-encoder for intrusion detection in iot systems." IEEE Internet of Things Journal 11.8 (2023): 14789-14803.

[29] Tekin, Nazli, et al. "Energy consumption of on-device machine learning models for IoT intrusion detection." Internet of Things 21 (2023): 100670.

[30] Rawat, Romil, et al. "Energy-aware detection of threat information propagation speed in social network of things using XGBoost and sequential pattern mining." Discover Computing 28.1 (2025): 223.

[31] Jamshidi, Saeid, et al. "Self-adaptive cyber defense for sustainable IoT: A DRL-based IDS optimizing security and energy efficiency." Journal of Network and Computer Applications 239 (2025): 104176.

[32] Ranpara, Ripal, et al. "A simulation-driven computational framework for adaptive energy-efficient optimization in machine learning-based intrusion detection systems." Scientific Reports 15.1 (2025): 13376.

[33] Yang, Tengfei, et al. "Towards trustworthy cybersecurity operations using Bayesian Deep Learning to improve uncertainty quantification of anomaly detection." Available at SSRN 4609553 (2024).

[34] Alturki, Badraddin, and Abdulaziz A. Alsulami. "Semi-Supervised Learning with Entropy Filtering for Intrusion Detection in Asymmetrical IoT Systems." Symmetry 17.6 (2025): 973.

[35] Wong, Joshua A., et al. "Uncertainty-quantified, robust deep learning for network intrusion detection." 2023 Winter Simulation Conference (WSC). IEEE, 2023.

[36] Wang, Xiaojie, et al. "A survey on trustworthy edge intelligence: From security and reliability to transparency and sustainability." IEEE Communications Surveys & Tutorials (2024).

[37] Neto, E. C. P., et al. "CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment." Sensors 23.13 (2023): 5941.

[38] PyRapl. INRIA, University of Lille. 2023. Python Running Average Power Limit. https://pyrapl.readthedocs.io/en/latest/index.html.

[39] Ferrag, Mohamed Amine, et al. "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning." IEEe Access 10 (2022): 40281-40306.