Unseen Attacks Identification in Intrusion Systems Using BERT and Logit Normalization

Frédéric Adjewa*†, Moez Esseghir*, Leïla Merghem-Boulahia*, Cheikh Kacfah†
* University of Technology of Troyes, LIST3N-Recy, France
† Computer Science and Information Systems departement, Institut Universitaire Saint Jean du Cameroun, Yaounde, Cameroon
Email: {frederic.adjewa, moez.esseghir, leila.merghem_boulahia}@utt.fr
cheikh.kacfah@institutsaintjean.org

***Abstract*—The exponential growth of internet-based services is enhancing human-machine interactions as we transition to next-generation networks. However, the threat landscape is also changing rapidly in tandem with this transformation, with increasingly complex cyberattacks that can pass through current security mechanisms because of their unidentified features. Identifying such novel threats is essential for an effective response, but existing AI-based intrusion detection systems methods often lack reliability when facing these unseen threats. Although transformer-based models have shown promise in threat detection and massive data volume management, they are prone to overconfident predictions on unfamiliar inputs, limiting their robustness. To address this, we propose a BERT-based intrusion detection approach trained with logit normalization, which enhances the model reliability by mitigating overconfidence in the model prediction, leading to better seen-unseen separation boundary. We evaluate our method on four benchmark datasets namely CICIDS2017, CICIoT2023, Edge-IIoTset and NSL-KDD, under intra- and cross-dataset conditions. Results show a substantial improvement in detecting unseen attacks, with accuracy rising from 3.9% to almost 100% while maintaining at least 95% detection for seen threats. This work results in a robust and unified model capable of detecting seen attacks and identifying unseen ones.**

***Index Terms*—Next-Generation Networks, Intrusion Detection Systems Robustness, Logit Normalization, Unseen Attack Detection.**

I. Introduction

The evolution of networks has significantly transformed the digital environment, notably through the exponential growth of data generated by the introduction of new services and applications. While these advancements create substantial value, they also expand the attack surface, leading to increasingly sophisticated and large-scale cybersecurity threats aiming at compromising these innovations. In recent years, intrusion detection systems (IDS) have benefited greatly from advances in machine learning and deep learning, enabling more effective analysis of network traffic to detect deviations from normal behavior and flag anomalies that may indicate malicious activity [1]. However, many of these AI-based IDS are developed under the *closed-world assumption* (CWA), where the model expects all inputs to belong to the classes seen during training.

However, in practice, models are deployed in an *open-world setting*, where they may encounter semantically novel, distributionally shifted or even adversarial inputs [2]. While AI-based IDS often achieve high performance on in-distribution (ID) test data (i.e., data drawn from the same distribution as the training set) their reliability degrades significantly when exposed to out-of-distribution (OOD) data that differ fundamentally from the training distribution.

The OOD problem is particularly critical in real-world cybersecurity, where network traffic is dynamic and novel attack types frequently emerge. In intrusion detection, we define OOD data as traffic that significantly deviates from the training distribution-here referred to as unseen attacks. Early detection of such anomalies is vital for effective threat response and ensuring system robustness. A reliable classifier must therefore be sensitive to semantic, covariate and broader distributional shifts [2].

Scalability is another key challenge for modern IDS, which must process ever-growing volumes of network traffic driven by the expansion of connected services. To address this, recent work has leveraged large language models, originally developed for NLP, to analyze complex network logs [3]. BERT, in particular, has shown strong potential due to its ability to capture contextual and sequential dependencies. However, these models are typically trained using the standard cross-entropy loss with a softmax activation function, which often results in overconfident predictions. This is especially problematic in high-risk environments such as cybersecurity. For instance, a model trained to detect brute-force and SQL injection attacks should not confidently misclassify a DDoS attack as one of these seen categories; instead, it should recognize the input as unfamiliar. This underscores the importance of integrating robust unseen threats detection mechanisms into the core detection process.

Several studies have explored Variational Autoencoders and LSTM networks for anomaly detection [4]–[7], though they struggle with sudden non-temporal anomalies and highly imbalanced datasets. Here, we propose a BERT-based model with logit normalization [8], which improves detection of unseen data while preserving strong seen threat identification performance.

To the best of our knowledge, this is the first application of logit normalization within a transformer-based architecture, such as BERT, for intrusion detection systems. Our approach not only enables accurate classification of seen attacks, but also

979-8-3503-9281-4/25/$31.00 ©2025 IEEE

identifies and isolates previously unseen traffic, rather than misclassifying it.

Our contributions are as follows.

- We critically analyse the drawbacks of traditional LLM-based IDS models, especially their tendency to make too confident predictions on unseen inputs.
- We provide a BERT-based method that may precisely detect and isolate attacks that have not been seen before, reducing the possibility of misclassifying OOD traffic as seen threats.
- In order to guarantee that the model only generates reliable predictions when the input falls within the learnt feature space, our approach incorporates a way to flag unseen inputs before classification.

II. Related Work

A. Transformer Models for Intrusion Detection

The use of transformer-based models in cybersecurity, particularly for intrusion detection systems, has gained significant traction in recent years. Originally designed for natural language processing (NLP), LLMs like BERT are now being adapted for threat detection tasks.

In [1], BERT was used to extract semantic features from network traffic, which were then processed by a CNN for intrusion detection, a hybrid approach that, while effective, may complicate the deployment in resource-limited settings. [9] applied LLMs directly for vulnerability detection and addressed their computational cost using compression techniques such as knowledge distillation and Low-Rank Adaptation (LoRA), demonstrating their viability in security contexts. CyBERT [10] was an early adaptation of BERT for cybersecurity, supporting tasks like classification and named entity recognition, though it was not tailored for intrusion detection. In contrast, SecurityBERT [11] offered a lightweight, privacy-aware BERT variant for IoT threat detection by reducing the number of transformer layers. Expanding on this, [12] introduced a federated version of SecurityBERT to address scalability and data privacy, validating its performance in both i.i.d. and non-i.i.d. environments. BERT has also been used for log-based anomaly detection, as shown in [3], where system logs were treated as structured sequences similar to language. Lastly, FlowTransformer [13] provided a general framework for evaluating transformer-based NIDS. While thorough in its architectural exploration, it did not address unseen detection.

Despite these advances, a critical limitation persists across these cited works: the absence of rigorous evaluation of model robustness, particularly on flagging as unseen what is unfamiliar to their knowledge. As these models are ultimately used for classification tasks in dynamic environments, the lack of unseen evaluation raises questions about their reliability in unseen or evolving threat landscapes. This gap underscores the need for the integration of robustness evaluation into the core design and validation of transformer-based security models.

B. Out-of-Distribution Detection Methods

Most models assume all classes are seen a priori, a premise misaligned with real-world dynamics. OOD detection addresses this by identifying data drawn from distributions different from the training set and overlaps with related fields such as anomaly detection, open set recognition, novelty detection, and outlier detection [2]. While these differ subtly-e.g., anomaly detection targets deviations from normality, open set recognition identifies unseen classes among seen ones, novelty detection flags unseen test-time classes, and outlier detection detects statistical deviations-they all contribute to understanding OOD behavior. Our focus is specifically on OOD detection due to its relevance for enhancing the robustness of AI-driven intrusion detection systems (IDS).

OOD detection methods are broadly categorized into *post-hoc* and *training-integrated* approaches. Post-hoc methods operate after training and typically rely on model-derived confidence scores. Notable examples include ODIN [14], which employs temperature scaling and small input perturbations to enhance ID/OOD separation; energy-based methods [4], which exploit the lower energy levels typically observed in ID samples and Mahalanobis distance-based approaches [15], which classify samples based on distances in the feature space using intermediate layer representations. In contrast, training-integrated approaches incorporate OOD-awareness directly into the optimization objective. These include using energy regularization as an auxiliary objective alongside cross-entropy loss [4]; G-ODIN [6], which adapts ODIN as a training objective; and Logit Normalization (LogitNorm) [8], which enforces a constant logit norm to improve generalization and robustness.

Although detailed coverage of all methods exceeds this scope, OOD detection remains a vibrant field central to the safe deployment of AI. Despite advances in transformer-based IDS, many models neglect systematic evaluation of distributional robustness; a critical oversight this study seeks to address.

III. Background

Understanding distributional shifts is essential for effective OOD detection. Let $\mathcal{X}$ denote the input space and $\mathcal{Y} = \{1, \ldots, C_K\}$ the label space, with the joint data distribution represented by $P(X, Y)$; $P(X)$ the marginal input distribution and $P(Y)$ the label distribution. In practice, two primary types of distribution shifts are of interest:

- *Covariate shift*, where the marginal input distribution changes: $P'(X) \neq P(X)$, potentially influencing the conditional distribution $P(Y|X)$.
- *Semantic shift*, characterized by a change in the label distribution: $P'(Y) \neq P(Y)$. In our context, this corresponds to the appearance of previously unseen attack classes. Since $P(X)$ is often conditioned on Y, a semantic shift may also implicitly alter $P(X)$. Therefore, this type is of particular interest in the study.

Assuming intrusion detection as a supervised classification problem, let the training dataset be denoted as $D^{in} =$

$\{(x_i, y_i)\}_{i=1}^{K}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathcal{Y}$. The goal is to learn a function $f_\theta : \mathbb{R}^d \to \mathbb{R}^K$ that maps each input to a vector of class scores (logits), typically optimized by minimizing the risk expected in equation 1

$$\mathcal{R}_{\mathcal{L}}(f) = \mathbb{E}_{(x,y)_{\mathcal{P}_{\mathcal{X}\mathcal{Y}}}}[\mathcal{L}(f(x;\theta), y)] \tag{1}$$

In the OOD detection setting, the classifier is expected to distinguish whether a given input x is in-distribution (labeled as 0) or out-of-distribution (labeled as 1). This introduces an auxiliary binary classification problem on top of the original multi-class setup.

The logit output $\mathbf{f}$ can be decomposed into its magnitude and direction:

$$\mathbf{f} = \|\mathbf{f}\| \cdot \hat{\mathbf{f}} \tag{2}$$

As demonstrated in [8], the magnitude $\|\mathbf{f}\|$ plays a crucial role in determining the model's confidence. During training, even after correct classification, the magnitude may continue to grow due to the objective of minimizing loss, leading to overconfident predictions, particularly problematic for OOD detection.

To address this, *logit normalization* constrains the logit's norm to remain constant throughout training. The modified loss function is expressed as:

$$\mathcal{L}_{logit\_norm} = -\log\left(\frac{e^{f_y/(\tau\|\mathbf{f}\|)}}{\sum_{i=1}^{K} e^{f_i/(\tau\|\mathbf{f}\|)}}\right) \tag{3}$$

Here, τ is a temperature parameter controlling the scale of the normalized logits. By fixing the norm, the optimization focuses solely on adjusting the direction $\hat{\mathbf{f}}$, thus mitigating overconfidence and enhancing the model's robustness to OOD samples.

IV. Methodology

In our study, logit normalization is integrated into the training process as a loss function rather than a direct detection strategy. Consequently, to evaluate unseen attacks detection capabilities, we rely on post-hoc scoring method applied after model training. This section presents our overall methodology, including the scoring method, model architecture and training procedure, dataset details and the evaluation metrics used to assess performance.

A. Scoring Method

As OOD detection is typically framed as a binary classification task, we adopt a post-hoc approach to assess whether a given input x belongs to the in-distribution set $\mathcal{P}^{in}$. This is achieved using a scoring function $S(x)$ and a threshold γ taken when $x\%$ of the ID sample are well classified, x being often 95%. The scoring function is formalized in Equation 4:

$$g(x) = \begin{cases} 0 & \text{if } S(x) \geq \gamma \quad \text{(in-distribution)} \\ 1 & \text{if } S(x) < \gamma \quad \text{(out-of-distribution)} \end{cases} \tag{4}$$

In our experiments, we used the Maximum Softmax Probability (MSP) method [16], which is one of the simplest scoring approaches. We deliberately chose this baseline to demonstrate that if our method performs well even with a basic score like MSP, then adopting more advanced alternatives, which have shown better performance in prior works, would likely yield even stronger results.

B. Model Implementation

Following prior work [11], [12], we employ a fine-tuned version of the BERT model whose architecture is presented in table I; adapted for processing network flow data as structured text. This choice leverages the model's ability to extract semantic features and capture long-range dependencies, which is particularly valuable in the context of intrusion detection.

TABLE I: **Architecture details of the lightweight BERT variant.**

Component	Specification
Transformer Encoder Layers	4
Attention Heads per Layer	4
Hidden Layer Size	256
Feedforward Network (FFN) Size	1024
Vocabulary Size	30,522
Fully connected layer size	128
Dropout	0.3
# parameters	**11170560**

This architecture has demonstrated competitive results in prior intrusion detection tasks, especially in resource-constrained environments such as IoT networks. Contrary to prior works, our model is trained using the logit normalization (described in Section 3), which encourages robust and well-calibrated confidence estimates.

C. Integration into the Process

In a traditional AI-enhanced intrusion detection system at inference, illustrated in Figure 1a, once the input is pre-processed, it is passed directly to the model for prediction. Regardless of whether the input is familiar or not, the model assigns it to one of the seen classes.

Our approach introduces an additional step before final classification, as shown in Figure 1b. After generating prediction scores, the system compares the highest score to a predefined threshold. If the score falls below this threshold, the input is flagged as potentially unseen and stored for further investigation. If the score exceeds the threshold, the system proceeds with classification as usual.

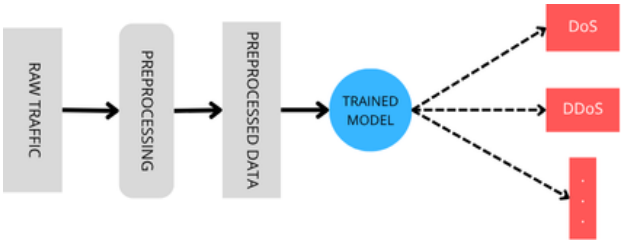
This modification helps the system isolate suspicious inputs instead of making blindly confident predictions, improving its ability to respond to novel or emerging threats.

V. Experimental Setup

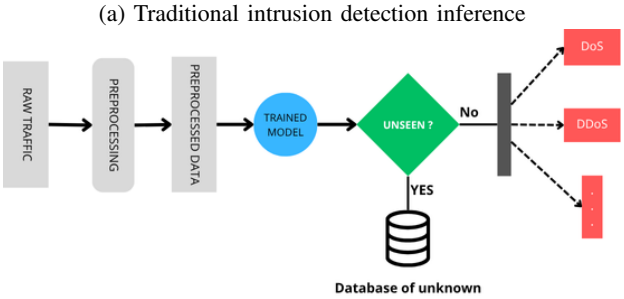
A. Datasets and Evaluation Scenarios

To evaluate the effectiveness of our approach, we selected four widely-used network traffic datasets, covering both traditional and IoT-based environments namely CICIDS2017 [1],

[1]https://www.unb.ca/cic/datasets/ids-2017.html



(a) Traditional intrusion detection inference



(b) Proposed approach with OOD detection

Fig. 1: **Intrusion detection during inference. (a) Traditional pipeline with direct classification, lacking OOD awareness. (b) Proposed pipeline with an additional step to detect and separate unseen inputs.**

NSL-KDD [2], CICIoT2023[3] and EdgeIIoTset[4]. Using these datasets, we constructed three distinct experimental scenarios to assess the robustness and generalization capabilities of our model under varying degrees of distributional shift.

a) Scenario 1: Intra-Dataset Evaluation: This scenario evaluates the model's ability to detect previously unseen attacks within the same dataset. The model is trained on a subset of seen attacks and tested on assumed unseen attack types from the same source. This mimics the real-world challenge of detecting new variants of seen attack families.

b) Scenario 2: Cross-Dataset Generalization: Here, we examine whether a model trained on one dataset can identify anomalous traffic from a completely different dataset as out-of-distribution. This setup evaluates the model's generalization to new environments and traffic patterns.

Each scenario includes a comparison between two training objectives: the standard *cross-entropy loss* and the *logit normalization loss*. This setup allows us to isolate the impact of logit normalization on OOD detection performance under different distribution shift conditions.

B. Evaluation Metrics

When training the model, the primary goal is to accurately classify the seen classes in the training set. To evaluate this, we use standard classification metrics such as accuracy, precision, and recall.

For OOD detection, however, the task is reframed as a binary classification problem. In this context, we adopt evaluation metrics commonly used in the literature: the Area Under the Receiver Operating Characteristic Curve (AUROC), and FPR@TPR95, which measures the false positive rate when the true positive rate is fixed at 95% [2].

[2]https://www.unb.ca/cic/datasets/nsl.html
[3]https://www.unb.ca/cic/datasets/iotdataset-2023.html
[4]https://www.kaggle.com/datasets/sibasispradhan/edge-iiotset-dataset

VI. Results and Discussion

A. Evaluation of Scenario 1: Intra-dataset

This first scenario, whose results are presented in Table II, evaluates the model's ability to detect unseen attacks within the same dataset. Each evaluation began with a baseline model trained using standard cross-entropy loss. As expected, the vanilla model performed poorly in identifying unseen inputs, with an average FPR of 73.25% across all datasets. It consistently misclassified unfamiliar attacks as belonging to one of the seen classes.

In contrast, applying logit normalization significantly improved the model's ability to separate seen from unseen attacks. This improvement was sensitive to the value of the scaling parameter τ, which controls how sharply the model distinguishes unfamiliar inputs. Across all evaluation sets, $\tau = 0.05$ consistently yielded the most effective decision boundary, reducing the average FPR by approximately 39%. Based on these results, we adopted $\tau = 0.05$ for the remainder of our experiments.

TABLE II: **Intra-dataset evaluation on four benchmarks. LN = Logit Normalization, FPR95 = False Positive Rate at 95% TPR, AUC = Area Under the Curve. Dataset abbreviations: CIC17 = CICIDS2017, IoT23 = CICIoT2023, Edge = Edge-IIoTset, NSL = NSL-KDD. ↓ indicates lower is better; ↑ indicates higher is better.**

Data	ID / OOD attacks	LN	τ	FPR95 ↓	AUC ↑
CIC17	DDoS, DoS / PortScan, BruteForce, Web	✗	-	70.15	65.22
		✓	0.5	74.23	63.13
		✓	0.05	**27.55**	**96.34**
		✓	0.005	64.40	88.12
IoT23	ICMP, UDP / PSH-ACK, Mirai-greeth, RSTFIN	✗	-	69.46	78.15
		✓	0.5	75.34	77.67
		✓	0.05	**33.16**	**97.37**
		✓	0.005	62.47	89.58
Edge	ICMP, UDP / SQLi, Scanner, Password	✗	-	72.80	87.34
		✓	0.5	73.56	87.22
		✓	0.05	**35.90**	**96.78**
		✓	0.005	46.43	90.67
NSL	DoS, Probe / R2L, U2R	✗	-	80.58	60.65
		✓	0.5	81.55	63.45
		✓	0.05	**48.21**	**89.54**
		✓	0.005	50.35	88.66

Figure 2 illustrates the behavior of the logit magnitude for a randomly selected sample during training. For the model trained with cross-entropy loss (CE), the logit magnitude increases rapidly and continuously throughout the training process. In contrast, for the model trained with logit normalization

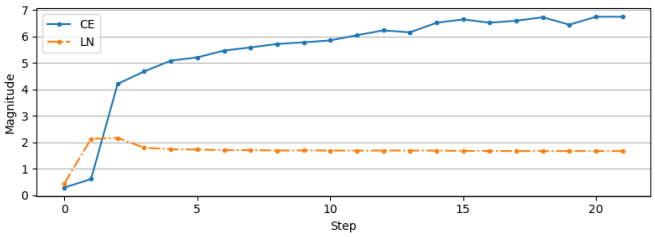


Fig. 2: **Random sample's logits magnitude chaging after 99 step**

(LN), the magnitude rises initially but stabilizes after the fourth training step. This plateau indicates that LN helps regulate the model's confidence, effectively preventing the unchecked growth of logits that often leads to overconfidence.

B. Evaluation of Scenario 2: Cross-Dataset Generalization

In this scenario, we evaluate the ability of a model trained with logit normalization to detect unseen threats from a different dataset. Specifically by positionning ourselves in context of the internet of things, caracterized by an exponentially growing traffic, we used CICIoT2023 as the ID dataset, containing DDoS-ICMP_Flood and DDoS-UDP_Flood attacks, and Edge-IIoTset as the OOD dataset, which includes SQL_Injection, Vulnerability_Scanner and Password attacks.

Figure 3a shows the confusion matrix of a model trained with standard cross-entropy loss. When the detection threshold is set to correctly classify 95% of the ID inputs, the model incorrectly labels nearly all unseen attacks as seen with only less than 4% of them being detected. This confirms a critical limitation: models often perform well on familiar data but fail to generalize when confronted with new, unseen threats.

In contrast, the confusion matrix in Figure 3b illustrates the behavior of a model trained with logit normalization. With the same 95% detection rate for seen threats, the detector is able to correctly flag 100% of the OOD instances as unseen, while maintaining a false positive rate of only 5%. This result highlights the improved robustness of the logit-normalized model in cross-dataset scenarios, making it a more reliable choice for real-world intrusion detection systems.

To further illustrate these results, Figure 4 shows the distribution of softmax scores for seen and assumed unseen samples. For the model trained with cross-entropy loss, most ID scores are close to one, but many unseen inputs's scores also cluster near one. This overlap makes it difficult to define a clear threshold γ to distinguish between seen and unseen inputs. In this case, the best achievable threshold is $\gamma = 0.997$, which in this case is very high and still not very effective.

In contrast, when using logit normalization, the distributions are clearly separated. Most ID scores remain high, while unseen inputs's scores drop significantly as shown in figure 4b, with a distinct separation point emerging around $\gamma = 0.75$. This sharp boundary allows for more reliable identification of unseen inputs during inference.

C. Comparative study

TABLE III: **Average probability outputs from a model trained to detect DDoS-ICMP_Flood and DDoS-UDP_Flood attacks from the CICIoT2023 dataset. In-distribution classes DDoS-UDP_Flood and DDoS-ICMP_Flood from CICIoT2023 and out of distribution classes, SQL_inection, Vulnerability_scanner, Password from EdgeIIoTset.**

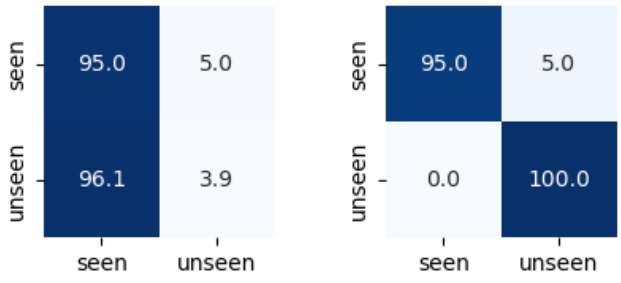
Model Variant	Knwon	Attack	Predicted Probs
SecurityBERT	✓	DDoS-ICMP_Flood	0.99
	✓	DDoS-UDP_Flood	0.99
	✗	SQL_injection	0.94
	✗	Vulnerability_scanner	0.946
	✗	Password	0.9137
Logit-Norm SecurityBERT	✓	DDoS-ICMP_Flood	0.8295
	✓	DDoS-UDP_Flood	0.8429
	✗	SQL_injection	0.6039
	✗	Vulnerability_scanner	0.608
	✗	Password	0.6028

To further assess the effectiveness of our approach, we conducted a comparative study. Table III presents the results obtained using our adaptation of SecurityBERT [11] on the task of detecting unseen attacks. As shown, the model trained with standard cross-entropy loss produces extremely high maximum softmax probabilities for both seen and unseen inputs. This indicates that the model is highly confident even when presented with unfamiliar data, showing no hesitation in misclassifying unseen inputs in seen classes. In contrast, when logit normalization is applied during training, the model maintains relatively high but more moderated confidence scores for in-distribution classes, while assigning significantly lower probabilities to unseen inputs. This behavior reflects an appropriate degree of uncertainty, which can be exploited for unseen detection.

TABLE IV: **Performance metrics by loss function and dataset**

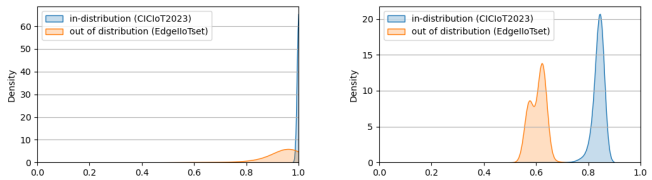
Loss Function	Dataset	Prec	Rec	F1-score	Acc
Cross Entropy	CICIDS2017	0.99	0.99	0.99	0.99
	CICIOT2023	0.999	0.999	0.999	0.999
	EdgeIIoT	1.00	1.00	1.00	1.00
	NSL-KDD	0.99	0.99	0.99	0.99
Logit Norm	CICIDS2017	0.999	0.999	0.999	0.999
	CICIOT2023	0.9988	0.9988	0.9988	0.9988
	EdgeIIoT	1.00	1.00	1.00	1.00
	NSL-KDD	0.99	0.99	0.99	0.99

Table IV compares the standard classification performance of the model when trained with cross-entropy loss versus logit normalization. Each model was trained on only two selected attack classes per dataset to facilitate comparison against unseen attack types. The results demonstrate that replacing cross-entropy loss with logit normalization does not degrade the model's classification capabilities. The observed consistency in standard metrics confirms that logit normalization can serve as



(a) CE-trained (b) LN-trained

Fig. 3: **Confusion matrices of the model performance on CICIoT2023 (id-distribution) and Edge-IIoTset (out-of-distribution) for CE and LN loss functions.**



(a) CE model (b) LN model

Fig. 4: **Softmax score distributions from CE and LN trained models.**

a reliable alternative, offering improved robustness to unseen inputs without compromising in-distribution performance.

VII. Conclusion

In this work, we evaluated the robustness of intrusion detection systems powered by large language models, with a focus on BERT-based architectures for threat detection. We demonstrated that the standard training approach, typically based on cross-entropy loss, fails to account for unseen inputs, posing a critical limitation in real-world scenarios. To overcome this, we introduced a simple yet effective alternative: training with logit normalization. This modification improves the model's ability to recognize unfamiliar inputs by adding a confidence-based filtering step before final classification. As a result, the system can flag and isolate potential threats that deviate from seen patterns. Our experiments showed strong performance in identifying unseen attacks from the same distribution. More importantly, cross-dataset evaluations confirmed that the proposed method can reliably detect unfamiliar threats, while significantly reducing false positives and maintaining a high TPR.

Future work will explore the model's behavior in the presence of benign noise and natural variability in network traffic, factors not addressed in the current experimental setup. These elements are important for assessing the model's robustness in more realistic deployment conditions. Additionally, we plan to investigate the stored database of identified unseen traffic, enabling better tracking, analysis and potential reclassification of emerging threats over time.

Acknowledgement

The authors express their gratitude to Saint Jean Ingenieur, a Cameroonian Private Institution of Higher Education, for their generous funding and support, which made this research possible.

References

[1] F. Ullah, S. Ullah, G. Srivastava, and J. C.-W. Lin, "Ids-int: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic," *Digital Communications and Networks*, vol. 10, no. 1, pp. 190–204, 2024.

[2] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635–5662, 2024.

[3] S. Chen and H. Liao, "Bert-log: Anomaly detection for system logs based on pre-trained language model," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2145642, 2022.

[4] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in neural information processing systems*, vol. 33, pp. 21464–21475, 2020.

[5] A. Corsini and S. J. Yang, "Are existing out-of-distribution techniques suitable for network intrusion detection?," in *2023 IEEE Conference on Communications and Network Security (CNS)*, pp. 1–9, IEEE, 2023.

[6] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Y. Xu, Q. Zhang, H. Deng, Z. Liu, C. Yang, and Y. Fang, "Unknown web attack threat detection based on large language model," *Applied Soft Computing*, vol. 173, p. 112905, 2025.

[8] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with logit normalization," in *International conference on machine learning*, pp. 23631–23644, PMLR, 2022.

[9] J. Liu, G. Lin, H. Mei, F. Yang, and Y. Tai, "Enhancing vulnerability detection efficiency: An exploration of light-weight llms with hybrid code features," *Journal of Information Security and Applications*, vol. 88, p. 103925, 2025.

[10] P. Ranade, A. Piplai, A. Joshi, and T. Finin, "Cybert: Contextualized embeddings for the cybersecurity domain," in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 3334–3342, IEEE, 2021.

[11] M. A. Ferrag, M. Ndhlovu, N. Tihanyi, L. C. Cordeiro, M. Debbah, T. Lestable, and N. S. Thandi, "Revolutionizing cyber threat detection with large language models: A privacy-preserving bert-based lightweight model for iot/iiot devices," *IEEE Access*, vol. 12, pp. 23733–23750, 2024.

[12] F. Adjewa, M. Esseghir, and L. Merghem-Boulahia, "Efficient federated intrusion detection in 5g ecosystem using optimized bert-based model," in *2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pp. 62–67, IEEE, 2024.

[13] L. D. Manocchio, S. Layeghy, W. W. Lo, G. K. Kulatilleke, M. Sarhan, and M. Portmann, "Flowtransformer: A transformer framework for flow-based network intrusion detection systems," *Expert Systems with Applications*, vol. 241, p. 122564, 2024.

[14] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *arXiv preprint arXiv:1706.02690*, 2017.

[15] H. Ghorbani, "Mahalanobis distance and its application for detecting multivariate outliers," *Facta Universitatis, Series: Mathematics and Informatics*, pp. 583–595, 2019.

[16] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.