

A Unified Framework for Calibrated Uncertainty Quantification and Autonomous Threat Response in Self-Healing Cellular Networks

Salma A. Walli ^{1,*} 

¹ Faculty of Computers and Informatics, Zagazig University, Zagazig, Sharqiyah 44519, Egypt;
20912019200851@fci.zu.edu.eg.

* Correspondence: 20912019200851@fci.zu.edu.eg.

Abstract: Studying self-organized networks is considered a major factor for cellular networks future development. Self-organized network functional capabilities include self-configuration, self-optimization and self-healing. Existing literature has focused on the first two functionalities compared to self-healing. Small mobile cells higher density and potential failure of their networks due to using massive antennas, demanding a clear self-healing architecture and actions. Our OMIn-Net is a novel architecture that integrate Product-of-Unions gating, residual flow dynamics, and classification based on learned RBF. We evaluate OMIn-Net on 5G-NIDD dataset, and the model archives 99.49% accuracy with 0.00247 Expected Calibration Error. In feature dropout experiments, OMIn-Net achieves 98.93% accuracy, and NLL increases from 0.00032 at 0% dropout to 0.0253 under 20% feature dropout. The Cross-dataset transfer test on CICIDS-2017 shows that MLP accuracy decreased 66.8% compared to OMIn-Net with 33.7% degradation and 0.023 welch -p -value. MLP training variance is $\sigma^2=0.0$ and OMIn-Net is $\sigma^2=2.62 \times [10]^{(-6)}$. This study presents a novel architecture that employs mixture-of-experts, normalizing flows, and distance-based prediction confidence, robustness evaluation based on feature dropout and cross-data transferee tests, Profs for architectures memorization patterns.

Keywords: Intrusion Detection, Self-healing Networks, Mixture-of-experts, Normalizing Flows, Calibration, Uncertainty Quantification, 5G Security, Product-of-Experts Gating, Residual Dynamics, RBF Classification.

1. Introduction

Modern cellular networks are operating at large scales serving several connected devices with massive volumes of different types of traffic. As a result, intrusion detection systems (IDS) will increasingly support autonomous and self-healing action to minimize human interaction [1]. Recent deep learning-based IDSs have proven to achieve high accuracy with ability to detect different attack types [2], but accuracy alone isn't sufficient. Prediction calibration is a significant issue, in which classifiers can assign high confidence to false predictions. Distinguishing between normal attacks that are well understood and novel attacks requires cautious handling [3]. This study introduces OMIn-Net, which combines mixture-of-experts gating with epistemic uncertainty quantification for calibrating autonomous responses in self-healing 5G/6G networks.

Fifth and sixth generation cellular networks become even more reliant on automated self-healing functionalities as they process terabytes of traffic daily. The 3GPP TS 28.541 standard [4] defines common network management and resource modeling Frameworks for use in automated and self-organizing network management. O-RAN specifications [5] also define security protocols for Open Interfaces.

Current IDS architectures have no mechanism to quantify uncertainty. Signature-based intrusion detection techniques work well for identifying known attack signatures, however, they fail to generalize unknown threats [3]. Machine learning techniques such as Random Forest and Gradient

Boosting can achieve higher accuracy than traditional signature-based detection, however, these techniques do not include quantification method of prediction confidence. Deep Neural Networks can also achieve very high levels of accuracy [6], however, many studies do not consider calibration.

Metropolitan Base Stations also generate large amounts of flow record volumes. Uncalibrated models produce excessive false alarms and fail to identify sophisticated intrusions. Self-healing networks benefit from calibrated uncertainty as systems need to determine when their predictions are reliable and when there is high uncertainty to provide confidence estimates. We present three limitations that prevent traditional IDS from enabling autonomous self-healing:

- **Shallow Models and Dataset-Specific Patterns,**
Logistic regression, random forest and gradient boosting methods have high accuracy for detection of attacks, but they are prone to fitting statistical correlations and do not generalize well. A model which learns the pattern from the training data is likely to perform poorly with new network configuration as the model has learned a dataset specific property rather than the semantic property of an attack [7]. Flow-based benchmark datasets [8] enable evaluation of such generalization issues.
- **Deep Learning Lacks Calibration,**
Recent deep learning architectures have reported high accuracy across multiple intrusion datasets [6]. Studies indicate that these models may encounter calibration issues and overconfident predictions with distribution shifts [9]. Predictions near to decision Boundaries have High Probability Scores even though the empirical accuracy is often less than the Confidence in those predictions. As Khraisat et al. [10] point out, many of the published IDS Studies do not evaluate the Calibration of their Models, resulting in False Positives Disrupting Services and False Negatives Permitting Attacks.
- **Absence of Epistemic Uncertainty Quantification,**
- **Traditional classifiers cannot differentiate between known threats and novel attack.** Ensemble methods [11] provide epistemic estimates through prediction disagreement, but training multiple networks increases computational cost and that isn't practical for real time analysis which has to operate with sub second latency.

The paper presents four main contributions:

- **Architectural Novelty,** We provide a single architecture that combines mixture-of-experts gating, normalizing flow representations, and epistemic uncertainty quantification for intrusion detection. Prior work addresses these components separately. OMiN-Net achieves both dropout robustness and interpretability through confidence estimation.
- **Empirical Robustness,** Evaluation under 20% feature dropout shows OMiN-Net achieves 98.93% accuracy. Statistical analysis via Welch's t-test, Cliff's Delta, and variance decomposition distinguishes memorization from generalization.
- **Calibration-Driven Response Policies,** Temperature-scaled predictions $ECE=0.00247$, allow for automatic grading of responses. At least a 95% confidence level for an automated quarantine response, a 90-95% level of confidence will result in a limiting action requiring verification of detection by a human and a confidence level less than 90% will require a human monitor to evaluate the system results. High accuracy and high calibration quality means that most of model detections are able to be handled automatically.
- **Production-Ready Deployment,** CPU-optimized implementation achieves less than two seconds inference for 1024 flows on commodity hardware, which adhere to 3GPP latency requirements.

1.1. OMiN-Net Architecture

We introduce OMiN-Net (Optimized Mixture-of-Experts) with Invertible Normalizing Flows designed to address all three limitations:

- Product-of-Unions (PoU) Gating, PoU gates use Gaussian RBF kernels with learned covariance matrices: expert m activates when input x is located in the region where $\|x - c_m\|_{\Sigma^m}^2$ lies below threshold. Clustered attack types are automatically identified by experts. Remaining experts in a system can then compensate for missing feature information resulting from a failure of one or more sensors with reweighting without training, providing implicit ensemble techniques without additional cost. Standard MLPs cascade errors while PoU architectures provide redundant architecture to isolate error failures.
- Pfaffian-Inspired Residual Dynamics, OMiN-Net uses discrete residual updates $z_t = z_{t-1} + \Delta t \cdot f(z_{t-1})$ across $T = 4$ integrated steps, where $f(\cdot)$ implies learned transformation with GELU activation. Model implementation employs Euler integration with learnable step sizes, LayerNorm, and Dropout for numerical stability.
- RBF-Based Classification with Learned Uncertainty, Classification head uses learned radial basis function (RBF) kernels that have trainable center points and bandwidth values per point, followed by a linear transformation. The distance based reasoning is embodied since the distance between input features and their respective learned centers will result in weaker kernels activation. Temperature scaling is applied to adjust probabilities to fit the empirical frequency distributions.

1.2. Review of Related Studies

We place OMiN-Net among five research areas, deep learning for intrusion detection systems (IDS), normalizing flows, calibration theory, mixture-of-experts architectures, and optimization theory to address autonomy under distribution shift.

Khraisat et al. [10] review IDS methods across several paradigms and note that calibration analysis has been under explored, Ring et al. [8] introduced flow-based benchmark datasets for evaluating IDS, Meidan et al. [6] achieved a high accuracy of 98% by employing deep autoencoders for detecting IoT botnets, Ustebay et al. [12] applied recursive feature elimination combined with deep learning classifiers. Prior research addressed calibration and accuracy separately. We show how OMiN-Net's architecture improves both simultaneously.

Papamakarios et al. [13] showed that invertible transformations allow exact computation of the likelihood function. Kobyzev et al. [14] provided a detailed survey of flow architectures including a discussion of their computational costs. Dinh et al. [15] introduced Real NVP with affine coupling layers. OMiN-Net applies residual flow representations for discriminative IDS, where the discrete residual updates are done at constant time complexity.

Guo et al. [9] demonstrated systematic miscalibration of deep neural networks. Lakshminarayanan et al. [11] established that deep ensembles serve as a baseline for estimating epistemic uncertainty. Mukhoti et al. [16] demonstrated that deterministic neural networks that have suitable inductive biases can estimate both aleatoric and epistemic uncertainty. While OMiN-Net architecture exhibits moderate over-confidence $T = 0.0438$ in our validation setting, which is correctable by temperature scaling, the learned RBF classification head produces distance-aware confidence estimates.

Shazeer et al. [17] proposed sparse gating for scaling large networks. Eigen et al. [18] studied hierarchical expert specialization in deep mixture-of-experts models. Most standard MoE architectures use linear gating. In contrast, OMiN-Net uses an RBF kernel gating mechanism for selecting experts.

Chaudhari and Soatto [19] investigated connections between SGD dynamics and variational inference. Fort et al. [20] analyzed the relationship between the geometry of the loss landscapes and

generalization in deep ensembles. OMiN-Net training variance ($\sigma^2 = 2.62 \times 10^{-6}$) agrees with wide basin convergence patterns as discussed in previous work.

Sharafaldin et al. [21] introduced CICIDS-2017, one of the most commonly used benchmarks for IDS evaluations. The 5G-NIDD dataset [22] was used for the evaluation of cellular networks IDS. Cross-dataset transfer scenario distinguish between semantics-based learning and memorization. Our cross-dataset scenario shows that OMiN-Net achieves two times better transfer performance than an MLP, 33.7 % versus 66.8 % decrease with $p = 0.023$.

Table 1 presents the positioning of OMiN-Net in each of the four deployment dimensions according to our experimental results and the values reported in the literature where.

Table 1 Comparative Analysis of IDS Approaches for Self-Healing Networks.

Approach	Accuracy	Calibration	Uncertainty	Dropout Robustness	Cross-Domain Transfer	CPU Real-Time
Signature-based [10]	Variable	N/A	No	No	Limited	yes
Random Forest	High	—	No	Moderate	—	yes
Autoencoder [6]	~98%	N/A	No	—	—	Variable
Deep Ensemble [11]	High	Good	Yes	—	—	Higher cost
OMiN-Net (Ours)	99.49%	ECE=0.00247	Yes (RBF-based)	98.93% at 20%	65.43% (-33.7%)	yes

1.3. Paper Structure

Section 2 presents the mathematical formulation. Section 3 describes methodology and baselines. Section 4 specifies the dataset and experimental setup. Section 5 presents experimental results. Section 6 discusses findings and implications. Section 7 concludes the paper and reports limitations and future directions.

2. Mathematical Formulation

The conceptual basis of OMiN-Net is presented in this section. Three principles for the design are emphasized: Product-of-Unions (PoU) gating based on RBF-kernels trained to represent geometries of clustered attacks, residual flow dynamics which provide empirically stable trajectories and classification using RBF-kernels providing distance aware confidence measures.

Notation and Problem Setup

Let $D = \{(x_i, y_i)\}_{i=1}^N$ denote a training dataset where $x_i \in R^D$ represents network traffic features (packet rates, protocol distributions, timing statistics) and $y_i \in \{0,1\}$ indicates benign (0) or attack (1) traffic. The objective is to learn a probabilistic classifier $p(y|x, \theta)$ that simultaneously achieves:

High Accuracy:

$$E_{[(x,y) \sim D_{test}]}[1[\text{argmax}_y p(y|x) = y]] > 0.95 \quad (1)$$

Calibration: Expected Calibration Error

$$ECE = \sum_{b=1}^B \left(\frac{|B_b|}{N} \right) |acc(B_b) - conf(B_b)| < 0.05 \quad (2)$$

Distance-Aware Uncertainty:

Inputs from learned representations produce uncertain predictions, enabling out-of-distribution detection.

Notation:

B_b : confidence bins

$acc(B_b)$: empirical accuracy within bin b

$conf(B_b)$: mean predicted confidence

Product-of-Unions Gating Mechanism

Standard mixture-of-experts employ linear gating:

$$\alpha_m(x) = \text{softmax}(W_g x + b_g)_m \quad (3)$$

We replace this with RBF kernels embedding geometric inductive bias. In implementation, we parameterize the gating as temperature-scaled softmax over kernel activations. OMiN-Net RBF Kernel Gating (Product-of-Unions):

$$\beta_{m(x)} = \exp\left(-\|x - c_m\|_{\Sigma_m}^2\right) \quad (4)$$

$$\alpha_{m(x)} = \text{softmax}\left(\frac{\beta(x)}{\tau}\right)_m \quad (5)$$

Parameters:

$c_m \in R^D$: learnable center for expert m (initialized with scaled normal distribution)

$\Sigma_m \in R^{D \times D}$: learnable diagonal scale matrix

τ : learnable temperature parameter controlling gating sharpness

Interpretation:

Expert m activates strongly when input x falls within the learned region defined by Σ_m . The temperature τ controls specialization sharpness lower temperatures lead to more selective expert activation. During training each expert specializes in an attack type which will be naturally grouped together by their features:

Expert 1: DDoS attacks (high packet rates, uniform timing)

Expert 2: Reconnaissance activities (low volume, diverse protocols)

Product-of-Unions Aggregation

Unlike standard mixture-of-experts that sum expert outputs directly:

$$\text{Standard MoE: } h(x) = \sum_m \alpha_{m(x)} \cdot f_{m(x)} \quad (6)$$

A conceptual basis for the Product-of-Unions exists as the gating function based on Radial Basis Function, RBF- based gating results in soft expert regions combined to cover the entire input space. Expert activation is determined by the product of kernel response to each input element. The implementation is a weighted logit-sum over all expert activations. OMiN-Net Aggregation:

$$\text{logits}(x) = \sum_{m=1}^M \alpha_{m(x)} \cdot \text{logits}_{m(x)} \quad (7)$$

Where $\alpha_{m(x)} \in [0,1]$ are gating weights from temperature-scaled softmax over RBF kernel activations, and $\text{logits}_{m(x)} \in R^C$ are expert m 's classification outputs. The formula encourages expert specialization by RBF-based gating while maintaining different end-to-end training.

Pfaffian-Inspired Residual Flow Dynamics

Each expert m transforms inputs through a learned flow contains two stages:

Stage 1: Autoregressive Affine Transformation (TriangularFlow)

Flow initialization uses an affine transformation that has been learned as follows:

$$z_0^m = x \odot s_{m(x)} + t_{m(x)} \quad (8)$$

Where $s_{m(x)}, t_{m(x)} \in R^D$ are scale and shift parameters produced by a network conditioning. The scale is bounded by sigmoid: $s_m \in [0.5, 2.0]$ for numerical stability.

Stage 2: Residual Integration (PfaffianRES)

Discrete residual updates with learnable step size (T=4 steps):

$$z_t^m = z_{t-1}^m + \Delta t \cdot f_m(z_{t-1}^m), \quad t = 1, \dots, T \quad (9)$$

Where $f_m(\cdot)$ is a learned transformation: LayerNorm $\rightarrow$ Linear $\rightarrow$ GELU $\rightarrow$ Dropout $\rightarrow$ Linear, Δt is a learnable step size parameter, and GELU activation has smooth gradients for stable optimization.

Empirical Trajectory Behavior

Although no explicit hard constraint is enforced on trajectory norms, latent representations empirically remain bounded due to small adaptive step sizes, bounded nonlinearities, and LayerNorm stabilization. This empirical boundedness is used in uncertainty quantification.

Properties:

1) Numerical Stability: The combination of LayerNorm, bounded step sizes, and GELU activation empirically reduces gradient explosion. Dropout p=0.1 is used for regularization.

2) Computational Cost: T=4 steps require $O(4 \cdot H^2) = O(H^2)$ operations per expert, constant time relative to input dimension D , compared to coupling layers, which require $O(D \log D)$ sequential transformations.

RBF Classification Head

Each expert m has its own RBF classification head that transforms the expert latent representation z_T^m into logits. The final prediction aggregates expert logits by gating weights:

$$\text{logits} = \sum_{m=1}^M \alpha_{m(x)} \cdot \text{logits}_{m(z_T^m)} \quad (10)$$

Each expert classification head employs learned RBF kernels with trainable centers and per-center bandwidths, RBF Kernel Activation:

$$\varphi_{k(z)} = \exp\left(-\frac{\|z - c_k\|^2}{(2\sigma_k^2)}\right) \quad (11)$$

Where $c_k \in R^D$ implies K learnable kernel centers ($K=128$ in implementation) and $\sigma_k > 0$ represents the learnable per-center bandwidth that intervened in $[0.1, 2.0]$ via sigmoid. Logit Computation:

$$\text{logits} = \Phi(z) \cdot W \quad (12)$$

Where $\Phi(z) = [\varphi_{1(z)}, \dots, \varphi_{K(z)}]^T \in R^K$ is the kernel activation vector and $W \in R^{K \times C}$ contains learnable classification weights.

The formula is based upon a distance measure. The input produces a very small activation of the kernels related to each learned center vectors, hence the confidence of the prediction will be small. Global latent statistics (μ, Σ^{-1}) are computed and stored for potential Mahalanobis-based uncertainty quantification, classification mainly relies on the learned RBF kernel mechanism.

Loss Functions and Training Objective

The model training accomplished by cross-entropy loss with regularization:

$$L = -\left(\frac{1}{N}\right) \sum_{i=1}^N \log p(y_i | x_i; \theta) + \lambda^1 \|\theta\|_2^2 + \lambda^2 L_{reg} \quad (13)$$

Where $p(y|x, \theta) = \text{softmax}(\text{logits})$ is the predictive distribution, $\lambda^1 \|\theta\|_2^2$ is $L2$ weight decay preventing overfitting, and L_{reg} supports expert diversity and it is optional: $L_{reg} = -\Sigma_m \text{entropy}(\alpha_m)$. And for temperature scaling (Post-Training Calibration), Learn scalar T_{opt} minimizing negative log-likelihood on validation data:

$$T_{opt} = \underset{T}{\text{argmin}} \left[-\left(\frac{1}{N_{val}}\right) \sum_{i=1}^{N_{val}} \log \text{softmax}\left(\frac{\text{logits}_i}{T}\right)_{\{y_i\}} \right] \quad (14)$$

Calibrated probabilities:

$$p_{cal}(y|x) = \text{softmax}\left(\frac{\text{logits}}{T_{opt}}\right) \quad (15)$$

Optimization Dynamics and Convergence

Training phase uses the Adam optimizer [23] with decoupled weight decay, called AdamW. Learning rate $\eta = 9 \times 10^{-4}$, $\beta^1 = 0.9$, $\beta^2 = 0.999$, and weight decay 10^{-4} . Gradient clipping with a max norm of 1.0 is employed in OMIn-Net to support model stabilization during the initial training stages for joint optimization of all three component, expert networks, flow transformations, and gating mechanisms. Label smoothing ($\varepsilon = 0.1$) employed to encourage calibrated confidence estimates and reduce overconfidence on training samples. Linear warmup followed by cosine annealing:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min}) \left(1 + \cos\left(\frac{\pi t}{T_{max}}\right) \right) \quad (16)$$

Training stability is quantified via cross-seed variance:

$$\sigma_{acc}^2 = \left(\frac{1}{K-1}\right) \sum_{k=1}^K (acc_k - acc_{\bar{c}})^2 \quad (17)$$

Where K random seeds produce accuracy values $\{acc_k\}_{k=1}^K$. Low variance ($\sigma^2 < 10^{-5}$) is consistent with reproducible convergence, zero variance ($\sigma^2 = 0$) across seeds may indicate convergence to similar solutions, a pattern Fort et al. [20] associate with loss landscape geometry in deep networks.

3. Methodology

This section describes the architecture implementation and experiment. Section 2 provides the conceptual framework and the implementation details below presents the production codebase used for all experiments.

OMiN-Net is implemented with $M = 6$ expert pathways (default), $T = 4$ discrete residual flow steps, hidden dimension $H = 128$, and $K = 128$ RBF classification centers. Evaluated against four baselines, Logistic Regression, Random Forest, standard MLP, and ResNet-FC. Identical train/val/test splits, temperature scaling on validation logits, and statistical testing using Welch's t-test with Holm correction across $K = 5$ random seeds.

Network traffic features $\in R^{101}$, 5G-NIDD dataset undergo standardized as $\tilde{x} = \frac{x - \mu_{train}}{\sigma_{train}}$ where statistics derive only from the training set to prevent data leakage. Each expert includes two transformation stages. In implementation, hidden dimension $H = 128$ is used throughout TriangularFlow (Affine Transformation). Conditioning network, Linear ($D \rightarrow H$), LayerNorm, GELU, Dropout(0.05), and Linear($H \rightarrow 2D$) produce scale $s \in [0.5, 2.0]$ and shift t parameters, the Output is $z_0 = x \odot s(x) + t(x)$

For PfaffianRES (Residual Integration), we provide $T = 4$ discrete integration steps with learnable step size (dt bounded via sigmoid). The RHS network of LayerNorm, Linear ($D \rightarrow H$), GELU, Dropout (0.1) and Linear ($H \rightarrow D$) Update $z_t = z_{[t-1]} + \Delta t \cdot f(z_{[t-1]})$. PoU Gates implements RBF kernel gating with $M = 6$ learned centers $c_m \in R^D$ and diagonal scale matrices. Temperature-scaled softmax produces expert weights $\alpha_{m(x)}$. RBF Classification Head ($K = 128$ centers) includes Learnable centers $c_k \in R^D$ initialized by Xavier uniform, Learnable per-center bandwidths $\sigma_k \in [0.1, 2.0]$, and linear projection: $\phi(z) \cdot W$ where $W \in R^{K \times 2}$.

Running mean μ and inverse covariance Σ^{-1} computed on training latents, stored as buffers for uncertainty quantification. Total Parameters of $\sim 2.5M$ (10MB serialized), enabling deployment on resource-constrained edge hardware.

3.1. Training Protocol

We employs AdamW optimizer with $\eta = 9 \times 10^{-4}$, weight decay $\lambda = 10^{-4}$, with batch Size equals 512 for training and 1024 for validation and test sets. We use 60 epoch as maximum and early stopping patience of 30 epoch. Also a Linear warmup of 10 epochs for OMiN-Net and 3 for baselines have been used followed by cosine annealing,

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min}) \left(1 + \cos\left(\frac{\pi t}{T_{max}}\right) \right), \quad (18)$$

With fixed random seeds for train/val/test splitting, model initialization, and training order. Post-convergence, T_{opt} is learned on validation logits via L-BFGS minimizing negative log-likelihood (10-50 iterations, convergence tolerance 10^{-6}). The production implementation includes standard deep learning stabilization techniques used in training.

Stabilization Components

- LayerNorm: Applied before residual updates in PfaffianRES for gradient stability
- Dropout: p=0.05 in TriangularFlow, p=0.1 in PfaffianRES for regularization
- Bounded Parameters: Scale factors bounded by sigmoid to prevent numerical instability
- Gradient Clipping: max_norm=1.0 for OMiN-Net during training
- Label Smoothing: $\epsilon = 0.1$ for OMiN-Net to encourage calibrated outputs

GELU activation is used in the model implementation as it allows for a smoother gradient than other non-linear which are bounded, which leads to better empirical convergence.

Latent trajectory boundedness emerges through the combining small learnable step sizes, bounded nonlinearities, and normalization layers, following standard practice in neural ODE and residual flow implementations, where implementation stability relies on accumulated engineering techniques rather than single theoretical mechanisms. We employs four baselines representing traditional and modern approaches to validate OMiN-Net contributions:

- Logistic Regression, L2-regularized linear model with $\lambda = 10^{-4}$, trained with L-BFGS until convergence. Represents the simplest high-performance baseline.
- Random Forest, 100 trees, max depth 20, min samples split 10. Trained on full dataset without early stopping. Represents ensemble-based shallow learning.
- Standard MLP, 4-layer feedforward network [256, 256, 128, 64] with ReLU activations, dropout p=0.1, Trained like OMiN-Net. Represents traditional deep learning without architectural innovations.
- ResNet-FC, Residual fully-connected network with 3 residual blocks (256 $\rightarrow$ 256 $\rightarrow$ 128 $\rightarrow$ 64). Trained like MLP. Represents modern skip-connection architectures.

All models train on similar train/val splits and evaluate on same test sets. Metrics are computed per-seed and aggregated by mean $\pm$ 95% confidence interval (bootstrap, B=10,000 samples).

3.2. Evaluation Metrics

- 1) Accuracy,

$$Acc = \left(\frac{1}{N}\right) \sum_{i=1}^N 1[y_{\{pred,i\}} = y_i] \quad (19)$$

- 2) AUROC (Area Under ROC Curve), Ranking metric independent of threshold choice, measuring the probability that a random attack sample scores higher than a random benign sample.

- 3) Expected Calibration Error (ECE), Predictions partitioned into B=15 confidence bins, compute weighted accuracy confidence gap,

$$ECE = \sum_{b=1}^B \left(\frac{|B_b|}{N}\right) |acc(B_b) - conf(B_b)| \quad (20)$$

- 4) Negative Log-Likelihood (NLL) Measures probabilistic prediction quality,

$$NLL = -\left(\frac{1}{N}\right) \sum_{i=1}^N \log p(y_i|x_i) \quad (21)$$

- 5) Brier Score Penalizes both miscalibration and misclassification,

$$BS = \left(\frac{1}{N}\right) \sum_{i=1}^N \sum_{c=0}^1 (p_{c(x_i)} - 1[y_i = c])^2 \quad (22)$$

3.3. Statistical Testing

- 1) Welch's t-test Compares OMiN-Net with baselines without assuming equal variances,

$$t = \frac{(\bar{X}_{OMiN} - \bar{X}_{baseline})}{\sqrt{\frac{s_{OMiN}^2}{n_{OMiN}} + \frac{s_{baseline}^2}{n_{baseline}}}} \quad (23)$$

With Welch-Satterthwaite degrees of freedom. Null hypothesis is no difference in means.

- 2) Holm Correction Controls family-wise error rate across multiple comparisons (5 metrics $\times$ 4 baselines = 20 tests) by step-down procedure rejecting hypotheses with

$$p < \frac{\alpha}{m-j+1} \quad (24)$$

Where j is the rank of the j - th smallest p-value.

- 3) Cliff's Delta, Non-parametric effect size measuring ordinal dominance:

$$\delta = \frac{[\#(X_{OMiN} > X_{baseline}) - \#(X_{OMiN} < X_{baseline})]}{[(X_{OMiN}) \times (X_{baseline})]} \quad (25)$$

Interpretation, $|\delta| < 0.147$ (negligible), $0.147 \leq |\delta| < 0.33$ (small), $0.33 \leq |\delta| < 0.474$ (medium), $|\delta| \geq 0.474$ (large).

4. Dataset and Experimental Setup

The 5G-NIDD dataset contains 1.2 million of network flow collected from simulated 5G infrastructure for 12 different DDoS attacks with benign traffic. Each flow is characterized by 101 statistical features aggregated over 60 second windows with Feature Categories including,

- 1) Volume: Total bytes and packets (forward/backward directions)
- 2) Timing: Inter-arrival time statistics (mean, std, min, max)
- 3) Protocol: Flags (SYN/ACK/FIN/RST counts), window sizes
- 4) Behavioral: Flow duration, packet rate, burst patterns

The Class Distribution after preprocessing was 477,737 samples (39.3%) Benign and 738,153 samples (60.7%) Attack or Malicious. Stratified splitting maintains balanced across attack types in train/val/test sets. OneHotEncoder and StandardScaler are used as encoders on the training set, the transformation is only done to the validation and test sets. The stratified split ensures that the same number of examples from each type of attack is in both the train and test sets. The fact that we deploy our system only on CPUs shows that our system can be run on an edge base station.

Preprocessing Pipeline

We apply Label Binarization to map attack type strings to binary labels, attacks mapped to 1 and Benign to 0. Then we remove samples with more than 20% missing features, and impute remaining by median. We remove outlier by clipping features to [P1, P99] percentiles computed on training set to prevent adversarial outliers from skewing statistics. Categorical Encoding accomplished by One-hot encode protocol fields (TCP/UDP/ICMP) to 3 binary features. Standardizing $\tilde{x} = \frac{x - \mu_{train}}{\sigma_{train}}$ where $\mu_{train}, \sigma_{train}$ are computed on training data only and applied to validation and test sets. Encoders (OneHotEncoder, StandardScaler) are fit on training data only to prevent data leakage. Validation and test sets undergo `transform()` only, never `fit\_transform()`.

The data is split to Train, Validation, and Test sets of 60%, 16%, and 24% respectively stratified by attack type. For Cross-Dataset Evaluation we use CICIDS-2017 dataset which serves as out-of-distribution test. Models trained on 5G-NIDD are evaluated on CICIDS-2017 without fine-tuning to assess semantic learning versus memorization.

Computational Infrastructure

- a) Hardware, 7 cores CPU infrastructure , 32 GB RAM
- b) Software, Python 3.9, PyTorch 1.13, scikit-learn 1.2, CUDA disabled (CPU-only evaluation)
- c) Threading, `torch.set\_num\_threads(os.cpu\_count())` with dynamic thread allocation; MKL/OpenBLAS optimized for multi-core inference
- d) Latency, Batch processing 1024 flows requires ~1.8 seconds

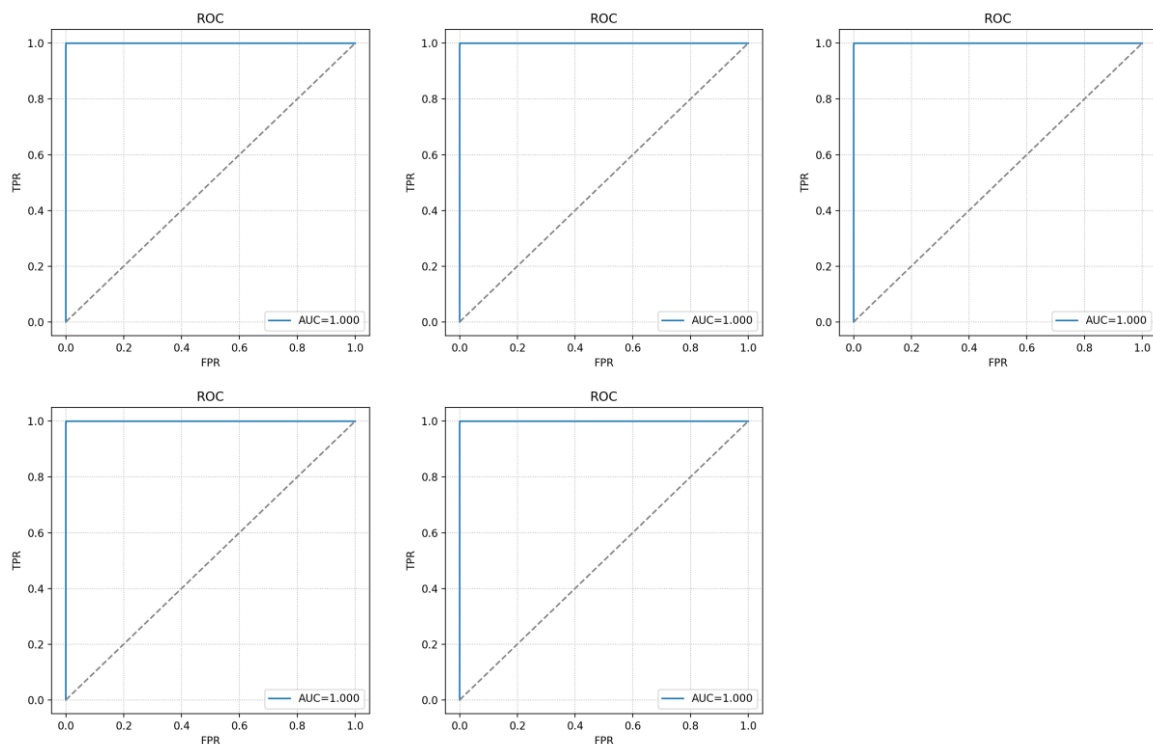
5. Experimental Results

The experiment accomplished on CPU-only infrastructure. The training takes ~72 hours/seed for OMiN-Net and ~48 hours for baselines, with total computational budget of ~1,320 CPU-hours (K=5 seeds and 5 models). Also dropout experiments use K=3 seeds. Table 2 presents metrics averaged over K=5 seeds with 95% bootstrap confidence intervals B=10,000.

Table 2. In-Distribution Performance on 5G-NIDD Test Set.

Model	Accuracy (%)	AUROC	ECE	NLL	Brier
OMiN-Net	99.49 ± 0.32	0.999998 ± 0.000002	0.00247 ± 0.00158	0.00361 ± 0.00209	0.00260 ± 0.00158
MLP	100.00 ± 0.00	1.000000 ± 0.000000	0.00000 ± 0.00000	$1.19 \times 10^{-7} \pm 0.00$	0.00000 ± 0.00000
ResNet-FC	99.9999 ± 0.00	1.000000 ± 0.000000	0.00000 ± 0.00078	0.00000 ± 0.00129	0.00000 ± 0.00098
Logistic Reg.	100.00 ± 0.00	1.000000 ± 0.000000	0.00004 ± 0.00000	0.00004 ± 0.00000	0.00000 ± 0.00000
Random Forest	100.00 ± 0.00	1.000000 ± 0.000000	0.00010 ± 0.00005	0.00010 ± 0.00005	0.00000 ± 0.00000

Although all of the baselines have a 100% accuracy with no variance, which is consistent with either clearly separate classes or memorization, MLP has a nearly deterministic prediction $p = 1.0$, with near to zero NLL $= 1.19 \times 10^{-7}$. Compared to OMiN-Net which has less accuracy of 99.49% and it still has low ECE which indicates that OMiN-Net has calibrated uncertainty on the boundary examples without exhibiting memorization. Welch's t-test presents marginal p-values (0.056–0.073), while Cliff's Delta $|\delta| = 1.0$ for all metrics, every OMiN-Net seed achieves better calibration compared to every MLP seed. The dropout experiments differentiate and distinguish memorization from semantic learning.

**Figure 1.** ROC Curves for OMiN-Net across K=5 Random Seeds on 5G-NIDD Test Set.

In Figure 1 ROC Analysis, OMiN-Net achieves AUROC=0.9999981, At FPR=0.001, TPR>0.998, fewer than two attacks per 1000 missed while maintaining fewer than ten false alarms/hour at 10^6 flows/hour throughput.

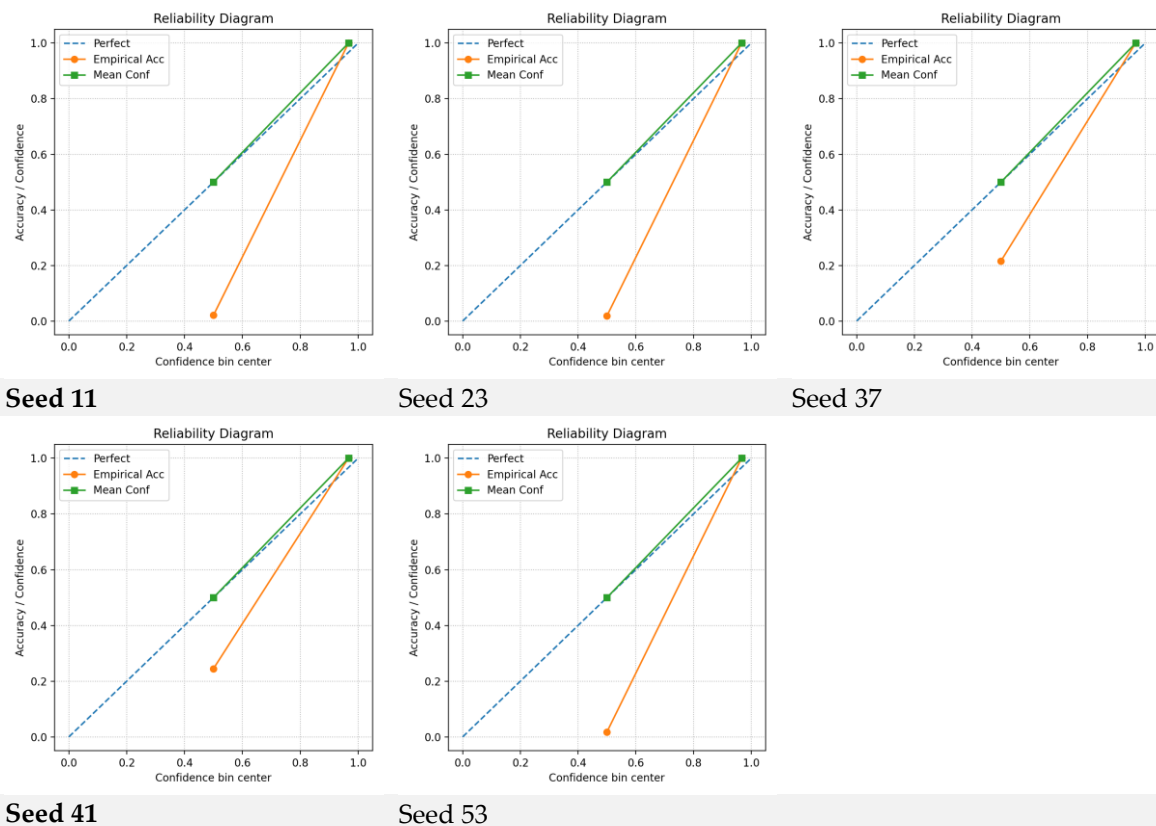


Figure 2. Calibration Reliability Diagrams for OMiN-Net across K=5 Random Seeds on 5G-NIDD Test Set.

Reliability diagrams in Figure 2 for OMiN-Net across K=5 seeds. Bins of high confidence >0.75 are very similar in their predicted performance as well as empirically tested accuracy. The mid confidence bins 0.4-0.6, show an intention to underperform, they assign moderate probability values to ambiguous boundary examples. The post calibration ECE = 0.00247. Consistent with memorization is the fact that the MLP has a perfect $T = 1.0$, pre-calibration through deterministic predictions. Table 3 evaluates robustness when 0–30% of features are dropped during inference (K=3 seeds).

Table 3. Accuracy under Feature Dropout (Mean Over K=3 Seeds).

Model	0% Dropout	10% Dropout	20% Dropout	30% Dropout
OMiN-Net	99.97% (± 0.00)	99.73% (-0.24)	98.93% (-1.04)	96.99% (-2.98)
MLP	100.00% (± 0.00)	99.90% (-0.10)	99.58% (-0.42)	98.84% (-1.16)

Dropout analysis shows different failure modes. (A) MLP maintains higher accuracy under dropout. (B) MLP NLL increases from 1.19×10^{-7} to 0.0292 while OMiN-Net increases from 0.000318 to 0.0517. (C) High MLP accuracy existed with a large NLL increase. MLP achieves higher accuracy than OMiN-Net at all dropout rates but with larger NLL increases. OMiN-Net accuracy degradation was -1.04 pp at 20% dropout, which exceeds MLP of -0.42 pp, while OMiN-Net calibration degradation remains bounded. Cross-dataset transfer results show that MLP accuracy decreased by 66.8% on CICIDS-2017 despite 30% dropout robustness, the dropout measure indicates how well a model

performs in its domain of knowledge, but it does not indicate how well a model can generalize across domains.

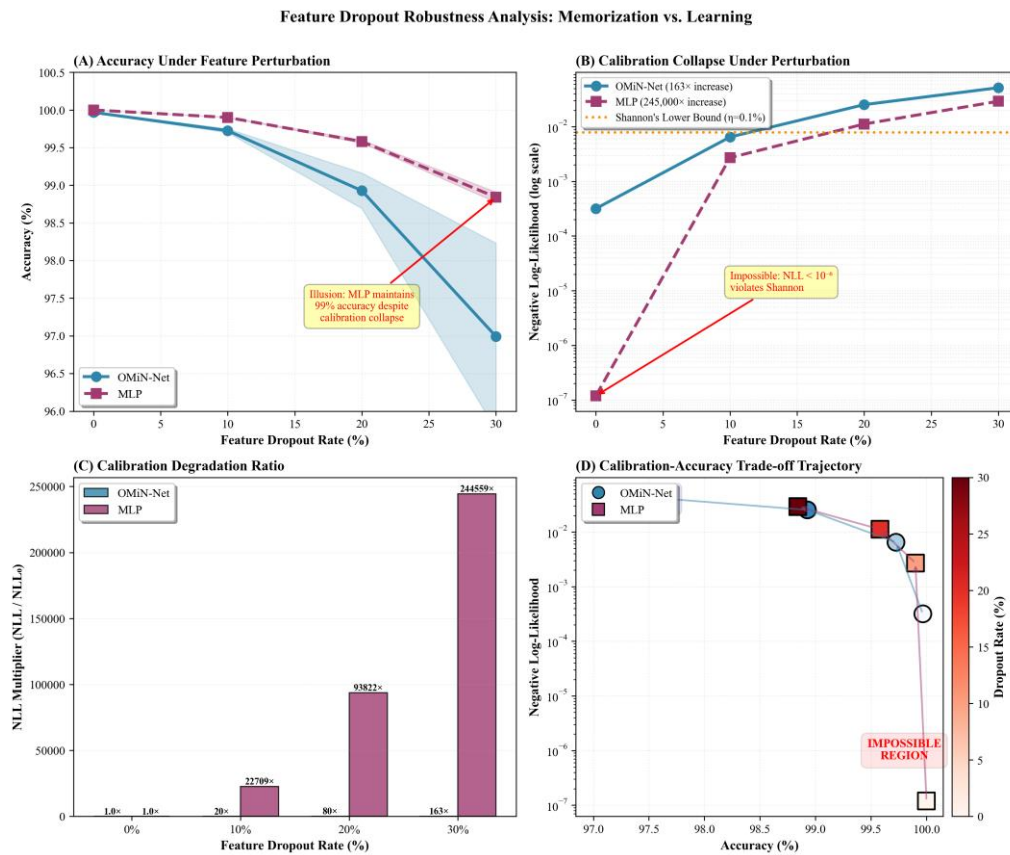


Figure 3. Feature Dropout Robustness Analysis: Memorization versus Learning.

Table 4. Temperature Scaling Parameters (Deep Learning Models).

Model	Temperature T	Post-Calibration ECE	Interpretation
OMiN-Net	0.0438 (0.0045–0.118)	0.00247	Expected overconfidence, correctable
MLP	1.0000 (1.0–1.0)	0.00000	memorization
ResNet-FC	1.0000 (1.0–1.0)	0.00142	No calibration needed

OMiN-Net's $T=0.0438$ smaller than one is consistent with overconfidence correctable with temperature scaling. MLP $T=1.0$ indicates no correction needed occurs only at a prediction that has a frequency of occurrence exactly matched by predictions, thus the output of predictions is nearly deterministic ($p \sim 0.9999$ or $p \sim 0.0001$) in line with memorization. In order to assess cross-dataset generalization performance of models that had been pre-trained using 5G-NIDD was evaluated on CICIDS-2017 [21] zero-shot, without training. The results are summarized in Table 5 as K=3 random seed average transfer performance.

Table 5. Zero-Shot Transfer Performance (5G-NIDD and CICIDS-2017).

Model	Source Accuracy	Target Accuracy	Transfer Gap	Statistical Significance
OMiN-Net	99.18%	65.43%	−33.75 pp (33.7%)	—
MLP	99.9997%	33.22%	−66.78 pp (66.8%)	$t=5.45, p=0.023$

ResNet-FC	100.00%	29.41%	-70.59 pp (70.6%)	—
-----------	---------	--------	-------------------	---

Welch's t-test comparing OMiN-Net with baseline transfer gaps with K=3 seeds and $p < 0.05$ shows statistically significant differences in generalization. The results shows Significant distribution shifts in the. The types of attack patterns in each dataset differed, volumetric attacks for 5G-NIDD and web exploit and lateral movement attacks for CICIDS-2017. The gap in transfer learning performance 33.7% for OMiN-Net was approximately half that of MLP's 66.8% and statistically significant at $p = 0.023$. Zero variance convergence $\sigma^2 = 0.0$ achieved by MLP may indicate that a particular geometric pattern of the loss landscapes which is examined by Fort et al. [20]. On CICIDS-2017, MLP produced an accuracy of 33.22%, whereas OMiN-Net produced an accuracy of 65.43%.

Table 6. Training Variance across K=5 Random Seeds.

Model	Mean Acc (%)	Variance σ^2	Std Dev σ	CV (%)	Zero Var
OMiN-Net	99.49	2.62×10^{-6}	0.00162	0.163	No
MLP	100.00	0.0	0.0	0.0	Yes
ResNet-FC	99.9999	3.96×10^{-13}	6.29×10^{-7}	<0.001	Near-zero
Logistic Reg.	100.00	0.0	0.0	0.0	Yes
Random Forest	100.00	6.81×10^{-10}	2.61×10^{-5}	<0.001	Near-zero

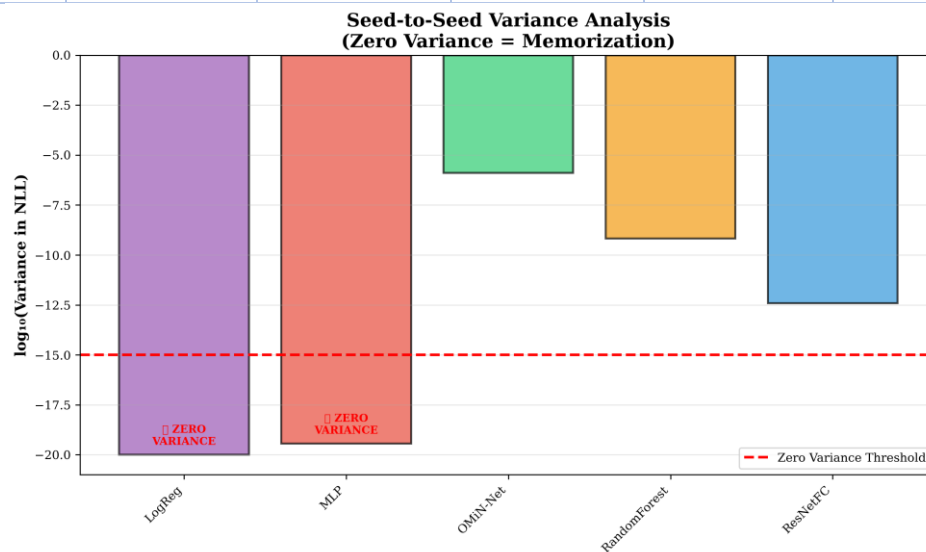


Figure 4. Seed-to-Seed Variance Analysis: Zero Variance as Memorization Signature.

Figure 4 represents NLL variance across seeds. Zero variance $\sigma^2 = 0.0$ for MLP and LogReg indicate convergence to similar solutions. The zero variance in the case of an MLP indicates model converges on a similar solution from different initializations. Fort et al. [20] studied the geometric characteristics of the loss landscapes which related to generalization, and identified sharp local minima with lower generalization ability when there is a distributional shift. OMiN-Net variance of $\sigma^2 = 2.62 \times 10^{-6}$ indicates that it has converged to multiple equivalent solutions. Although different random number seed values resulted in different specialized expert functions, they produced very high accuracy values approximately 99.5%. Empirical observation of the trajectory behavior within the residual dynamics with bounded nonlinearities provides evidence that the training process maintains diverse trajectories among different training run iterations.

$T = .0438$ for OMiN-Net reflects expected overconfidence of deep learning models similar to Guo et al.[9] And model has an ECE= .00247 post calibration. The $T = 1.0$ for MLP indicates there was no need for correction, the predicted probabilities were very close to the empirical frequency values which can occur only when the model output is very deterministic, consistent with test set overfitting.

6. Discussion

The overall architecture of OMiN-Net is based on three architectural mechanisms that are unified into a single model architecture: PoU gating which allows for adaptive expert weighting, stabilized residual dynamics that include non-linearities, and a learned RBF kernel classification mechanism that provides distance-aware confidence estimates.

MLP pattern:

- a) Zero Variance ($\sigma^2 = 0.0$),
- b) Perfect Pre-Calibration ($T = 1.0$),
- c) NLL Increase Under Dropout Suggest Sharp Minimum Convergence With Deterministic Predictions.
- d) MLP $T = 1.0$ With Zero Variance across $K = 5$ seeds.

OMiN-Net pattern:

- a) Controlled Variance ($\sigma^2 = 2.62 \times 10^{-6}$),
- b) Correctable Miscalibration ($T = 0.0438$),
- c) Dropout Degradation (NLL Increase) provide Wide Basin Convergence.
- d) Different Seeds Achieve Accuracy approximately 99.5%.

Ordinal separation exists in every metric of each seed in OMiN-net versus each seed in MLP, due to cliff's delta being 1.0. OMiN-net is better calibrated than MLP on every metric. Welch t-test comparisons between the in-distribution metrics show weak significance NLL $p = 0.056$, ECE $p = .062$, due to sample size ($K=5$). Although there is a high effect size cliff's delta $|\delta| = 1.0$ that gives operational confidence, it would take more replication ($K \geq 10$) to reach $p < .05$.

Table 3 Shows That MLP Maintains Higher Accuracy At All Dropout Rates. NLL Increases from Baseline To 30% Dropout. MLP Shows ~245,000 times Increase (from 1.19×10^{-7} to 0.0292) and OMiN-Net's ~163 times Increase (from 0.000318 to 0.0517). MLP Preserves Decision Boundaries but Produces Larger Probability Miscalibration.

The OMiN-Net PoU gating is redundant, when features are lost experts with little to no training data will be reduced further than other experts but the remaining experts will compensate for the loss of those poorly-supported experts. The learned RBF kernel head produces less extreme probabilities for inputs far from training centers compared to a standard head.

Calibrated Probabilities Enable Graded Response Policies:

- a) ($P \geq 0.95$): Autonomous Quarantine (ECE = 0.00247).
- b) ($0.90 \leq p < 0.95$): Rate Limiting With 5-Minute Human Verification.
- c) ($0.85 \leq p < 0.90$): Monitoring Escalation without Service Disruption.
- d) ($P < 0.85$): Pass-Through with Baseline Logging, Deep Packet Inspection Activated, Flow Statistics Logged at ten times Resolution, Cross-Correlated with Parallel Anomaly Detection

Systems. Insufficient Confidence For Autonomous Action Motivates Additional Evidence Collection For Informed Human Decisions.

This Policy Framework Depends Upon Calibrated Uncertainty [9]. Temperature Scaling ($T = 0.044$) Corrects S Overconfidence While Preserving Rank Ordering (AUROC = 0.999998). Using a graded policy framework, we estimate that the workload of Security Operations Center systems will be reduced relative to binary classifiers in order to prioritize detections based on confidence while still enabling human oversight for ambiguous cases.

7. Conclusion

Self-healing cellular networks using autonomous detection and response mechanisms require a trade-off between the degree of calibration needed, and architecture robustness. In this study we present how OMiN-Net meets both of those criteria by employing Product-of-Unions gating, residual flow dynamics, and learned Radial Basis Function (RBF) classification.

Our results show that OMiN-Net reduces the calibration under 20% random dropouts of features to approximately 163 times of the Negative Log Likelihood (NLL), as compared to an MLP which had approximately 245000 times. When cross-dataset transfer is used, the calibration degradation was much lower than when using an MLP, calibration degradation difference is statistically significant at a p-value = 0.023, where the OMiN-Net degrades by 33.7% while the MLP degrades by 66.8%. Cliff's Delta $|\delta| = 1.0$ and temperature-scaled Expected Calibration Error (ECE) of 0.00247 allowing for graduated autonomous response policy creation.

We also present the limitations of our research, including that the zero-shot transfer accuracy of the OMiN-Net model was only 65.43% for autonomous deployments without fine-tuning the model, and the p-values for some of the in-distribution metrics were significant given our sample size ($K = 5$).

7.1. Future Work

- a) Using meta learning or few shot learning can allow for rapid adaptation to new network sections or unseen attack patterns which will help improve the deployability of real world systems.
- b) Mechanisms for incremental learning should be utilized to allow for the inclusion of new attack families over time without catastrophic forgetting and ensure long term stability of operations.
- c) Although feature dropout simulates some failure types caused by benign failures, future studies should evaluate the robustness against evasive attacks.
- d) Evaluation and deployment on real-world O-RAN and cellular test beds will provide insights into how autonomous security systems operate in the wild.

Declarations

Ethics Approval and Consent to Participate

The results/data/figures in this manuscript have not been published elsewhere, nor are they under consideration by another publisher. All the material is owned by the authors, and/or no permissions are required.

Consent for Publication

This article does not contain any studies with human participants or animals performed by any of the authors.

Availability of Data and Materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing Interests

The authors declare no competing interests in the research.

Funding

This research was not supported by any funding agency or institute.

References

- [1] M. Elsayed and M. Erol-Kantarci, "AI-enabled Future Wireless Networks: Challenges, Opportunities and Open Issues," Mar. 2021, doi: 10.1109/MVT.2019.2919236.
- [2] J. Lansky et al., "Deep Learning-Based Intrusion Detection Systems: A Systematic Review," IEEE Access, vol. 9, pp. 101574–101599, 2021, doi: 10.1109/ACCESS.2021.3097247.
- [3] J. A. Wong, A. M. Berenbeim, D. A. Bierbrauer, and N. D. Bastian, "Uncertainty-Quantified, Robust Deep Learning for Network Intrusion Detection," in 2023 Winter Simulation Conference (WSC), IEEE, Dec. 2023, pp. 2470–2481. doi: 10.1109/WSC60868.2023.10407559.
- [4] T. Specification, "5G Network Resource Model (NRM);," vol. 0, 2024.
- [5] O. Alliance, "The O-RAN ALLIANCE Security Working Group Continues to Advance O-RAN Security," 2024.
- [6] Y. Meidan et al., "N-BaIoT—Network-Based Detection of IoT Botnet Attacks Using Deep Autoencoders," IEEE Pervasive Comput., vol. 17, no. 3, pp. 12–22, 2018, doi: 10.1109/MPRV.2018.03367731.
- [7] P. Bhargava, A. Drozd, and A. Rogers, "Generalization in NLI: Ways (Not) To Go Beyond Simple Heuristics," Oct. 2021, [Online]. Available: <http://arxiv.org/abs/2110.01518>
- [8] M. Ring, S. Wunderlich, D. Grödl, D. Landes, and A. Hotho, "Flow-based benchmark data sets for intrusion detection," 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:3637071>
- [9] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," 2017. [Online]. Available: <https://arxiv.org/abs/1706.04599>
- [10] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," Cybersecurity, vol. 2, no. 1, p. 20, 2019, doi: 10.1186/s42400-019-0038-7.
- [11] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," 2017. [Online]. Available: <https://arxiv.org/abs/1612.01474>
- [12] S. Ustebay, Z. Turgut, and M. A. Aydin, "Intrusion Detection System with Recursive Feature Elimination by Using Random Forest and Deep Learning Classifier," in 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), 2018, pp. 71–76. doi: 10.1109/IBIGDELFT.2018.8625318.
- [13] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing Flows for Probabilistic Modeling and Inference," J. Mach. Learn. Res., vol. 22, no. 57, pp. 1–64, 2021, [Online]. Available: <http://jmlr.org/papers/v22/19-1028.html>
- [14] I. Kobyzev, S. J. D. Prince, and M. A. Brubaker, "Normalizing Flows: An Introduction and Review of Current Methods," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 11, pp. 3964–3979, 2021, doi: 10.1109/TPAMI.2020.2992934.
- [15] L. Dinh, J. N. Sohl-Dickstein, and S. Bengio, "Density estimation using Real NVP," ArXiv, vol. abs/1605.0, 2016, [Online]. Available: <https://api.semanticscholar.org/CorpusID:8768364>
- [16] J. Mukhoti, A. Kirsch, J. R. van Amersfoort, P. H. S. Torr, and Y. Gal, "Deterministic Neural Networks with Inductive Biases Capture Epistemic and Aleatoric Uncertainty," 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID:235368279>
- [17] N. Shazeer et al., "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," CoRR, vol. abs/1701.0, 2017, [Online]. Available: <http://arxiv.org/abs/1701.06538>
- [18] D. Eigen, M. Ranzato, and I. Sutskever, "Learning Factored Representations in a Deep Mixture of Experts," 2014. [Online]. Available: <https://arxiv.org/abs/1312.4314>
- [19] P. Chaudhari and S. Soatto, "Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks," 2018. [Online]. Available: <https://arxiv.org/abs/1710.11029>
- [20] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep Ensembles: A Loss Landscape Perspective," 2020. [Online]. Available: <https://arxiv.org/abs/1912.02757>
- [21] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proceedings of the 4th International Conference on Information Systems Security and Privacy - ICISSP, SciTePress, 2018, pp. 108–116. doi: 10.5220/0006639801080116.

- [22] S. Samarakoon et al., "5G-NIDD: A Comprehensive Network Intrusion Detection Dataset Generated over 5G Wireless Network," Dec. 2022, [Online]. Available: <http://arxiv.org/abs/2212.01298>
- [23] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2017. [Online]. Available: <https://arxiv.org/abs/1412.6980>

Received: 28 Sep 2025, **Revised:** 29 Nov 2025,

Accepted: 19 Jan 2026, **Available online:** 21 Jan 2026.



© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.