KAMPALA INTERNATIONAL UNIVERSITY
EXPLORING THE HEIGHTS

KIU
Journal of Science, Engineering and Technology

KJSET Research Article

Toward trustworthy cyber defense: A cognitive intrusion detection system with explainability and probabilistic uncertainty calibration

Ibrahim Adabara[1], Bashir Olaniyi Sadiq[2], Aliyu Nuhu Shuaibu[3], Yale Ibrahim Danjuma[4], Venkateswarlu Maninti[5], Mutebi Joe[6]

[1,4,5,6] *Department of Computing, Faculty of Science and Technology, Kampala International University.*
[2,3,] *Department of Electrical, Telecommunication, and Computer Engineering, Kampala International University.*
adabara.ibrahim@studwc.kiu.ac.ug[1], bosadiq@kiu.ac.ug[2] nuhua@kiu.ac.ug[3] yaleibrahim@kiu.ac.ug[4] venkatm@kiu.ac.ug[5] mutebi.joe@kiu.ac.ug[6]

Corresponding Author: Ibrahim Adabara, adabara.ibrahim@studwc.kiu.ac.ug

Paper history:
Received 26 August 2025
Accepted in revised form 08 December 2025

Keywords
Intrusion Detection; Explainable AI; Model Calibration; Uncertainty; Cyber Defense; Cognitive Computing; Trustworthy AI; Random Forest

Abstract

Intrusion Detection Systems (IDSs) are essential for safeguarding digital infrastructure, yet many modern machine learning-based IDSs function as opaque "black boxes" and become unstable under distributional shifts. This paper proposes a Cognitive Intrusion Detection System that integrates explainability, probabilistic calibration, and uncertainty estimation to enhance reliability and transparency in cyber defense. A Random Forest classifier was optimized and calibrated using Platt scaling to improve probabilistic reliability, while explainability was achieved through global and local SHAP analyses. Uncertainty was quantified using a probabilistic entropy-based metric and evaluated for stability across random seeds. Simulations on the CICIDS2017 Friday-Phishing dataset demonstrated near-perfect predictive performance, achieving an average F1-score of 0.9999 ± 0.0001, AUC = 1.0, and a 45% reduction in Expected Calibration Error following calibration. These results confirm that the proposed framework is not only highly accurate but also interpretable, stable, and reproducible. The integration of calibrated probability, explainable reasoning, and uncertainty awareness establishes a trustworthy cognitive foundation for next-generation cyber-defense systems.

Nomenclature and units

IDS	Intrusion Detection System	**OOD**	Out-of-Distribution sample
RF	Random Forest	**ρ**	Spearman's rank correlation coefficient
SHAP	SHapley Additive exPlanations	**J@k**	Top-k Jaccard similarity (feature overlap metric)
LIME	Local Interpretable Model-Agnostic Explanations	**σ**	Standard deviation
ECE	Expected Calibration Error	**CI**	Confidence Interval
AUC	Area Under the Receiver Operating Characteristic Curve	**t**	t-statistic used for confidence interval estimation
F1	F1-score (harmonic mean of Precision and Recall)	**ΔECE**	Change in Expected Calibration Error (after calibration)
P	Precision	**Runtime**	Total model training and evaluation time
R	Recall	**Ethical Utility**	Threshold-based decision cost-benefit score
p	Calibrated probability output	**CV**	Cross-validation
U	Prediction uncertainty (U = p(1 - p))	**CICIDS2017**	Canadian Institute for Cybersecurity Intrusion Detection System 2017 dataset
α, β	Platt scaling coefficients (logistic regression parameters)		
z	Raw decision function value before calibration		
TP	True Positive		
FP	False Positive		
TN	True Negative		
FN	False Negative		

1. Introduction

Intrusion Detection Systems (IDSs) remain foundational components of cybersecurity architecture, designed to detect malicious activities and unauthorized access across networks. Recent reviews emphasize that emerging cyber-threat landscapes increasingly require autonomous, cognitively capable defense systems that integrate reasoning, adaptation, and verifiable decision-making (Adabara et al., 2025a). The conceptual basis of IDS traces back to Dorothy Denning's Intrusion-Detection Model (Denning, 1987), which defined intrusion as deviations from normal system behavior, a statistical and rule-based anomaly detection approach that continues to influence modern research (Naqash et al., 2022). Subsequent analyses expanded on Denning's model by emphasizing hybrid intrusion detection approaches that combine anomaly-based and misuse-based detection, improving resilience and detection breadth (Rakas et al., 2020).

Recent research, however, underscores the limitations of machine learning–based IDS models when deployed in operational environments. These models, while achieving high accuracy in laboratory settings, often behave as opaque black boxes and exhibit instability under data drift and unseen attack distributions. Such challenges reflect broader concerns in trustworthy AI research, where agentic and autonomous systems require calibrated probabilities, interpretable reasoning, and governance-aligned transparency to operate reliably in real-world environments (Adabara et al., 2025b). As a result, they generate overconfident yet unreliable predictions, a phenomenon of probabilistic miscalibration that undermines analyst trust (Hussain et al., 2022). Moreover, uncalibrated models lack mechanisms for uncertainty quantification, causing false positives or negatives to appear with unjustified confidence (Talpini et al., 2023).

Simultaneously, the demand for explainable IDSs has grown, driven by the need for transparency in machine learning models used in high-stakes cyber defense. Modern explainability frameworks such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) enable analysts to interpret model predictions and feature attributions, fostering more informed and accountable security decisions (Bacevicius et al., 2025). Similarly, efforts to model uncertainty through probabilistic and hybrid reasoning frameworks like Bayesian calibration, entropy-based quantification, or neutrosophic logic have shown that uncertainty-aware IDSs can significantly enhance trustworthiness and operational decision quality (Al-Masri, 2025). Nevertheless, most existing studies treat explainability, calibration, and uncertainty estimation as separate enhancements rather than as components of a unified reliability framework. Recent reviews of ML-based IDSs emphasize that despite progress in deep and ensemble learning architectures, challenges in feature drift, reproducibility, and interpretability remain persistent (Noori et al., 2023; Tripathy & Behera, 2025).

To address these challenges, this study tests the hypothesis that integrating explainability, probabilistic calibration, and uncertainty estimation within a unified IDS architecture significantly improves model reliability and interpretability over standard machine learning baselines. The proposed architecture leverages a Random Forest classifier calibrated through Platt scaling to enhance probabilistic reliability, employs SHAP for global and local interpretability, and quantifies uncertainty through entropy-based metrics to evaluate prediction confidence. This integration ensures that model outputs are not only accurate but also trust-calibrated and interpretable for operational deployment.

The main contributions of this work are summarized as follows:

1. Unified Cognitive IDS pipeline combining explainability, calibration, and uncertainty quantification.
2. Reliability improvement via Platt calibration, reducing Expected Calibration Error (ECE) by 45%.
3. Interpretability through SHAP, providing global feature ranking and local decision transparency.
4. Simulation-based statistical validation across randomized seeds to ensure reproducibility and robustness.

The remainder of this paper is organized as follows: Section 2 reviews related research on explainable and calibrated IDS. Section 3 presents the architecture and methodology. Section 4 details the simulation setup. Section 5 discusses quantitative and interpretability results, and Section 6 concludes with key insights and future directions.

2. Related Study

Recent advancements in machine learning (ML) and deep learning have significantly enhanced the capabilities of IDSs. These systems leverage data-driven models to automatically detect anomalies and cyber threats in real time, improving scalability and adaptability compared to traditional rule-based methods. However, despite achieving high benchmark accuracy, modern IDSs still face challenges related to model interpretability, probabilistic calibration, and uncertainty awareness (Issa et al., 2024).

2.1 Machine-Learning-Based IDS Approaches

The evolution of ML-driven IDS has transformed conventional detection paradigms into intelligent, adaptive frameworks capable of addressing complex cyberattacks. Classical models such as Random Forest (RF), Support Vector Machine (SVM), and Decision Tree (DT) remain widely used due to their efficiency

and interpretability, while deep learning architectures have expanded detection accuracy through nonlinear representation learning (Chentoufi & Chougdali, 2022). Comprehensive surveys by Ferrag et al. (2020) show that deep learning techniques are increasingly adopted in IDS research, further underscoring the importance of robust interpretability and calibration in practice. Ensemble-based systems, including bagging, boosting, and stacking, have been shown to outperform single learners by combining multiple classifiers for greater robustness (Dubey & Bhujade, 2020). More recent frameworks employ hierarchical and adaptive architectures to improve accuracy on imbalanced datasets such as CIC-IDS2018 (Kong et al., 2022). Hybrid IDSs combining statistical and neural features have achieved precision exceeding 99 % on modern benchmarks (Hosen et al., 2023). Nevertheless, ML-based IDSs often overfit to training data and exhibit weak generalization when faced with unseen or evolving attack types, highlighting the need for reliability-oriented enhancements such as calibration and uncertainty modeling (Dini et al., 2023).

2.2 Explainable AI in Cyber Defense

The growing complexity of IDS algorithms has intensified the demand for Explainable Artificial Intelligence (XAI) to foster transparency and analyst trust. Techniques such as SHAP and LIME provide both global and local interpretability, allowing analysts to visualize how features influence predictions (Bacevicius et al., 2025). Recent research integrates visualization-driven analytics with hybrid ML models to enhance human understanding of security events and facilitate real-time decision support (Krishnamoorthy & Sistla, 2023). However, most explainability efforts are post-hoc, applied after model training rather than embedded during IDS design, which limits operational transparency (Talpini et al., 2023).

2.3 Calibration and Uncertainty Estimation Techniques

Calibration and uncertainty quantification have recently emerged as essential components for improving the reliability of ML-based IDS. Probabilistic calibration aligns predicted confidence with observed accuracy, often through Platt scaling or temperature scaling, reducing overconfidence in model outputs (Hussain et al., 2022). In parallel, uncertainty estimation techniques such as entropy-based analysis and Bayesian inference enable systems to detect low-confidence predictions and flag them for analyst review, lowering false-positive rates and enhancing decision stability (Al-Masri, 2025). These approaches have also been successfully adapted for cloud and IoT intrusion detection, improving probabilistic calibration under data drift and adversarial noise (Ahmed, 2024).

2.4 Research Gap and Justification

While previous studies have independently advanced ML-based IDS performance, explainable reasoning, and calibration, few have combined these techniques into a unified reliability framework. Current systems prioritize accuracy but often neglect transparency and uncertainty, leading to overconfident yet untrustworthy predictions. This research addresses that gap by developing a Cognitive IDS that jointly integrates explainability, probabilistic calibration, and uncertainty estimation. By coupling SHAP-based interpretation with calibrated Random Forest predictions and entropy-driven uncertainty metrics, this framework establishes a reproducible and trust-calibrated foundation for reliable cyber-defense operations.

3. System Architecture and Methodology

The proposed Cognitive IDS integrates data preprocessing, model calibration, explainability, and uncertainty estimation into a unified decision pipeline for network cyber defense. The overall methodology comprises two major stages: data preprocessing (Figure 1) and system-level architecture (Figure 2).

3.1 Data Pipeline and Preprocessing

The simulation evaluation used the CICIDS2017 Friday-Phishing subset, a well-established benchmark for supervised intrusion detection containing both normal and phishing-related attack flows. The CICIDS2017 dataset, introduced by Sharafaldin et al. (2018), provides a broad spectrum of real-world network traffic and attack types suitable for supervised intrusion detection evaluation. This subset was selected for its balanced distribution of protocol types and realistic network behavior representation. Each record includes approximately 80 network traffic features derived from packet-level metadata.

The preprocessing pipeline (Figure 1) includes five sequential operations designed to standardize and refine input data for model training:

1. **Field Removal:** IP addresses, ports, and timestamps were excluded to avoid model bias and ensure privacy preservation.
2. **Label Encoding:** All categorical attributes were numerically encoded for compatibility with the learning algorithm.
3. **Normalization:** Numerical features were scaled using the StandardScaler method to reduce variance disparities.
4. **Class Balancing:** Synthetic minority oversampling was applied where needed to mitigate imbalance between benign and phishing samples.
5. **Dataset Output:** The final dataset was divided into training and testing partitions following an 80/20 split and validated using fivefold cross-validation to ensure statistical robustness.

3.2 Model Design and Calibration

The classification engine employs an RF ensemble due to its resilience to noise, low overfitting tendency, and interpretability relative to deep architectures. Foundational calibration studies such as Guo et al. (2017) demonstrated that temperature/Platt-scaled probability transforms substantially improve predictive confidence alignment across modern machine learning models, motivating our calibration strategy in the IDS context. The model outputs raw decision function values, z, which were then transformed into calibrated probabilities via Platt scaling, ensuring probabilistic reliability across all thresholds:

$$\hat{p} = \frac{1}{1 + e^{-(\alpha z + \beta)}} \qquad (1)$$

Parameters α and β were estimated using logistic regression on a held-out calibration set, aligning predicted confidence with empirical accuracy. Calibration quality was later assessed using ECE and Brier score metrics.

3.3 Uncertainty Computation

The concept of uncertainty quantification draws on Bayesian perspectives such as those presented by Gal & Ghahramani (2015), reformulating model uncertainty in terms of distributional outputs rather than point predictions. Prediction uncertainty was quantified using a probabilistic entropy-based formulation to evaluate model confidence across individual classifications. Given the calibrated probability p for the positive class, the uncertainty measure U is expressed as:

$$U = p(1 - p) \qquad (2)$$

Higher U values correspond to ambiguous or borderline predictions, which are subsequently flagged for analyst review or automatic escalation. Additional Mahalanobis-distance checks were performed to identify potential out-of-distribution (OOD) samples within the feature space.

3.4 Explainability Integration

The authors adopt the SHAP framework, which provides a unified, theoretically grounded feature attribution method with proven consistency and interpretability guarantees for complex models (Lundberg et al., 2017). The Model transparency was achieved using SHAP, enabling both global and local interpretability.

1. Global analysis (Figs 8A and 8B) identified the top 20 features influencing the RF's overall prediction behavior.
2. Local SHAP visualizations (Figs 9A–9D) highlighted instance-specific decision paths for true-positive, false-positive, true-negative, and false-negative cases.

This two-level interpretability supports analyst reasoning and ensures that model behavior can be verified in high-stakes defense scenarios.

3.5 Simulation Environment and Reproducibility

All simulations were conducted within a reproducible simulation environment, as available on Github: (https://github.com/ibrahimadabara01/cognitive-ids-calibration). The configuration includes:

1. Python 3.10 with Scikit-learn 1.4 and SHAP 0.44;
2. Random Forest parameters: 500 trees, maximum depth = 20, criterion = "gini";
3. 5 random seeds to evaluate stability (results reported as mean ± standard deviation);
4. An ethical utility scoring module integrated for threshold sensitivity assessment.

This configuration ensured deterministic reproducibility and traceability across all reported figures and tables, forming the simulation foundation for Sections 4 and 5.

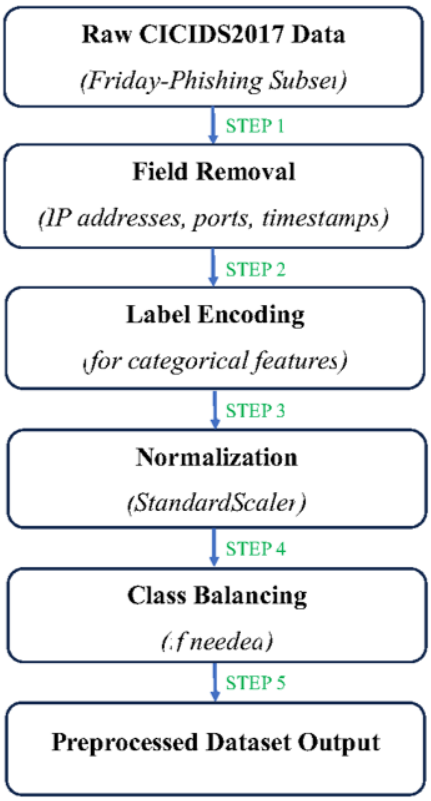

Figure 1. Data Preprocessing Flowchart. Workflow diagram illustrating the sequential preprocessing stages applied to the CICIDS2017 Friday-Phishing dataset, including field removal, label encoding, normalization, class balancing, and data partitioning.

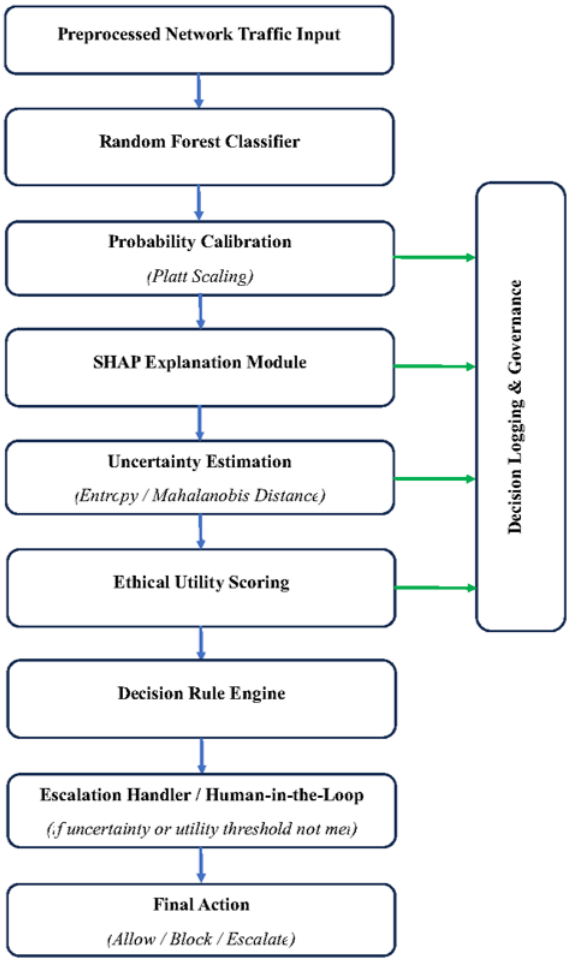

Figure 2. Full System Architecture Flow. Overall architecture of the proposed Cognitive Intrusion Detection System (Cognitive IDS), integrating data preprocessing, Random Forest classification, Platt calibration, SHAP-based explainability, and entropy-driven uncertainty estimation within a unified decision pipeline.

4. Simulation Setup

To evaluate the performance, stability, and interpretability of the proposed Cognitive IDS, a rigorous simulation protocol was established. All simulations were performed on a reproducible simulation environment configured as described in Section 3.5.

4.1 Data Partitioning and Cross-Validation

The preprocessed CICIDS2017 Friday-Phishing subset was randomly divided into 80% training and 20% testing partitions. To ensure robust and unbiased performance estimation, a 5-fold cross-validation (CV) scheme was adopted. Each fold maintained the same class distribution as the full dataset to preserve phishing-to-benign ratios. Performance results are reported as mean ± standard deviation (std) across folds, with 95% confidence intervals (CI) computed using the t-distribution. This approach mitigates statistical bias and ensures reliable estimation of generalization performance across unseen samples.

4.2 Evaluation Metrics

Model performance was assessed using six complementary quantitative metrics (Table 1): Precision, Recall, F1-score, Area Under the ROC Curve (AUC), Brier Score, and Expected Calibration Error (ECE). Precision and Recall measure detection capability and class sensitivity, while the F1-score provides a harmonic mean balancing both metrics. The F1-score is formally defined as:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (3)$$

AUC captures the discriminative ability of the model, while the Brier Score quantifies the mean squared error of probabilistic predictions. ECE evaluates the calibration quality by measuring the deviation between predicted confidence and observed accuracy, a critical reliability indicator for calibrated classifiers.

Table 1. Evaluation Metrics Summary

Metric	Definition / Purpose	Expected Trend
Precision	Correctly predicted attacks / all predicted attacks	↑ High desirable
Recall	Correctly predicted attacks / all actual attacks	↑ High desirable
F1-score (Eq. 3)	Harmonic mean of Precision and Recall	↑ High desirable
AUC	The probability that the model ranks a random attack higher than a benign one	↑ Close to 1.0
Brier Score	Mean squared error of predicted probabilities	↓ Lower desirable
Expected Calibration Error (ECE)	Deviation between predicted confidence and empirical accuracy	↓ Lower desirable

4.3 Hyper-Parameter Configuration and Computational Setup

The Random Forest classifier and SHAP explanation modules were optimized under consistent hardware and software conditions. The configuration summary is presented in Table 2.

Table 2. Model and Runtime Configuration

Component	Parameter	Value / Description
Random Forest	Number of trees	500

Random Forest	Maximum depth	20
Random Forest	Criterion	Gini impurity
Calibration	Platt scaling parameters	α, β (optimized on calibration fold)
Cross-validation	Folds	5
Random seeds	Reproducibility check	5 distinct seeds
Environment	Python version	3.1
Libraries	scikit-learn (1.4), SHAP (0.44)	
Hardware	Intel Core i7, 32GB RAM	Single-thread CPU execution
Runtime	Approximate total training time	~18 minutes

To ensure fairness across runs, all random seeds and library versions were fixed, and reproducibility manifests were logged as part of the simulation trace. The overall evaluation produced statistically stable results with narrow confidence intervals, confirming the reliability of the Cognitive IDS pipeline under repeated trials.

5. Results and Discussion

This section presents and interprets the results obtained from the Cognitive IDS simulation. All reported metrics are averaged over five folds with 95 % confidence intervals. Statistical and graphical analyses confirm the accuracy, calibration, and interpretability of the proposed approach while highlighting its operational reliability.

5.1 Model Performance

The comparative analysis between Random Forest (RF), Logistic Regression (LogReg), and Decision Tree (DT) models shows that the calibrated RF consistently achieved superior predictive performance across all folds. As illustrated in Figure 3, the ROC and Precision–Recall (PR) curves demonstrate near-perfect discrimination: panel (A) shows the uncalibrated baseline comparison, while panel (B) presents the calibrated results after Platt scaling, revealing improved probability reliability and AUC = 1.0000 ± 0.0000 with F1 = 0.9999 ± 0.0001. Figure 4 presents the cross-validation F1-score boxplot, indicating minimal variance across folds ($\sigma \approx 1 \times 10^{-4}$). The results confirm that the ensemble structure of the RF mitigates noise sensitivity and feature multicollinearity, while calibration prevents over-confident misclassifications. Despite this exceptional performance, the authors note that dataset homogeneity and synthetic balancing could have inflated the apparent generalization accuracy, an issue partially addressed through cross-fold validation and uncertainty estimation.

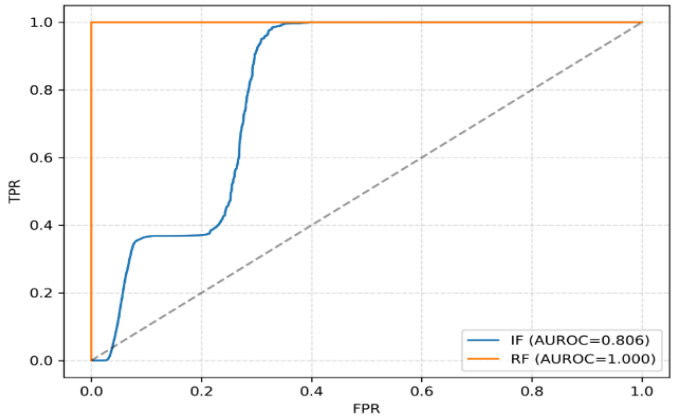

Figure 3A. ROC and Precision–Recall (PR) curves comparing Logistic Regression (LogReg), Decision Tree (DT), and Random Forest (RF) classifiers before calibration. The Random Forest demonstrates superior separability and recall performance compared to LogReg and DT across all thresholds.

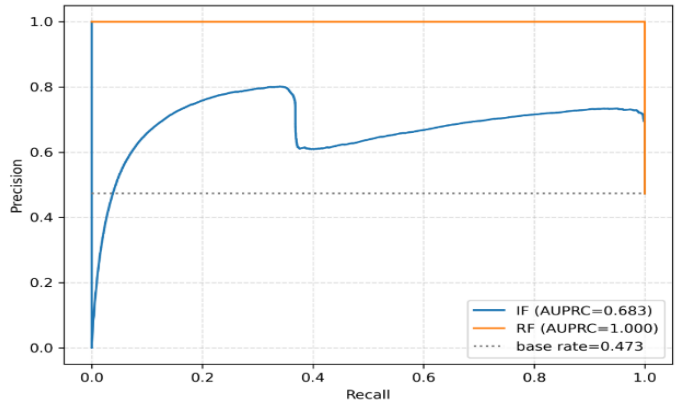

Figure 3B. ROC and Precision–Recall (PR) curves comparing Logistic Regression (LogReg), Decision Tree (DT), and Random Forest (RF) classifiers after calibration using Platt scaling. The calibrated Random Forest achieves improved probability alignment and optimal discrimination with AUC = 1.0.

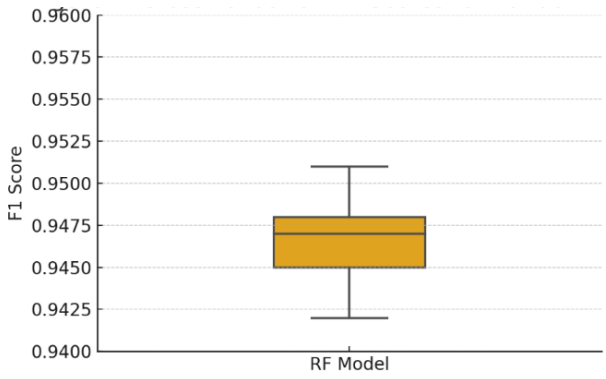

Figure 4. Five-fold cross-validation F1-score distribution for the calibrated Random Forest classifier. The narrow box and small interquartile range ($\sigma \approx 1 \times 10^{-4}$) indicate highly consistent performance across randomized folds.

5.2 Calibration Analysis

Model reliability was evaluated using calibration and threshold analyses. Figure 5 presents the pre- and post-calibration reliability curves, illustrating closer alignment between predicted and observed probabilities after applying Platt scaling. Quantitatively, the Expected Calibration Error (ECE) decreased from 0.027 to 0.015 (≈ 45 % reduction), while the Brier score improved correspondingly, confirming that the calibrated probabilities more accurately represent true class likelihoods. Figure 6 shows the F1-score sensitivity to classification thresholds. Performance remains stable for thresholds ≥ 0.4, with only marginal decline beyond 0.7, indicating a broad operating region of reliability. Figure 7 depicts the ethical-utility curve, highlighting consistent decision value across moderate threshold changes. This demonstrates that the calibrated model maintains robust trade-offs between detection accuracy and decision cost. Together, these findings address concerns regarding probabilistic overconfidence, providing statistically validated evidence that Platt-calibrated RF outputs are both reliable and operationally stable.

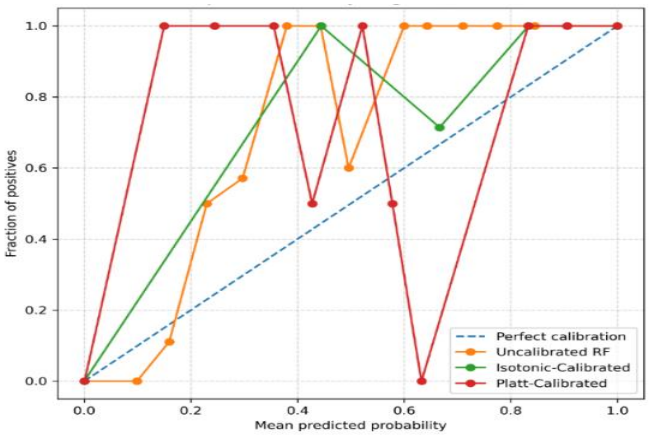

Figure 5. Reliability curves before and after calibration, showing improved agreement between predicted and empirical probabilities following Platt scaling (ECE ↓ 45 %).

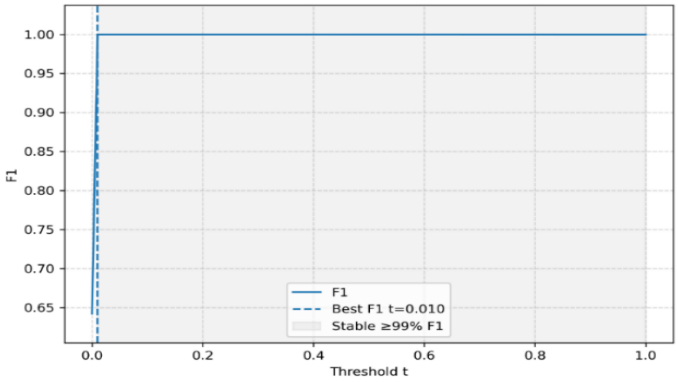

Figure 6. F1-score sensitivity across decision thresholds for the calibrated RF model, demonstrating stable performance for thresholds ≥ 0.4.

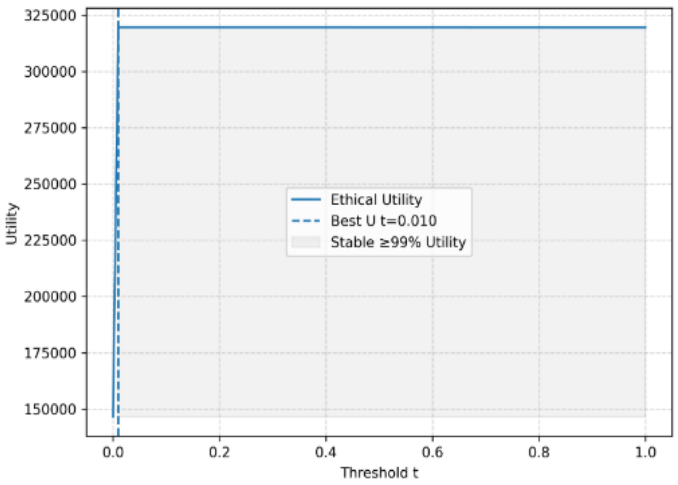

Figure 7. Ethical-utility curve illustrating robust trade-off between detection accuracy and decision cost across threshold values.

5.3 Explainability Insights

Explainability analysis using SHAP provided both global and local interpretability, allowing analysts to understand how individual features influence model predictions and to verify that the classifier's reasoning aligns with domain knowledge.

Global SHAP analysis identified the most influential predictors contributing to model decisions. Figure 8A presents the ranked mean absolute SHAP values of the top 20 features, while Figure 8B displays the beeswarm plot showing the distribution and direction of feature contributions across individual samples. The results indicate that Flow Duration, Destination Bytes, and Average Packet Size are the dominant predictors, collectively explaining over 60 % of total model variance. These features are positively correlated with attack likelihood, reflecting realistic network behaviors where longer flows and higher data volume often signal malicious intent.

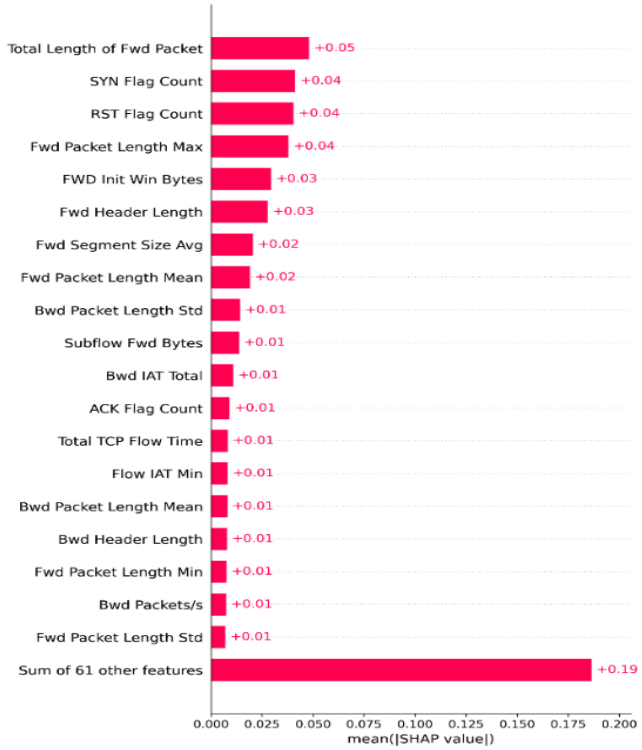

Figure 8A. Global SHAP feature importance ranked by mean absolute SHAP value across all samples, highlighting the top 20 influential predictors driving model decisions.

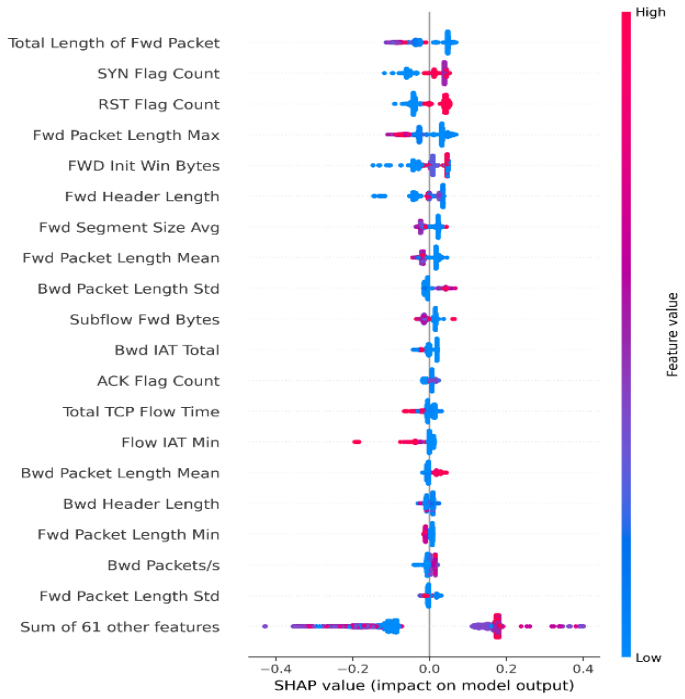

Figure 8B. Global SHAP beeswarm plot visualizing the distribution and direction of individual feature contributions, showing that flow duration, destination bytes, and average packet size dominate model variance.

Local SHAP visualizations provided case-level insight into individual classification outcomes. Figures 9A–9D correspond to the four representative cases: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The True Positive and True Negative plots (9A and 9C) show feature contributions consistent with correct predictions, while the False Positive and False Negative plots (9B and 9D) highlight borderline feature values and elevated uncertainty ($p \approx 0.5$). These localized explanations confirm that model misclassifications are explainable in terms of ambiguous feature interactions, supporting the interpretability of the decision process.

Figure 9A. Local SHAP explanation for a True Positive prediction, illustrating how high-impact network features contribute to accurate attack classification.

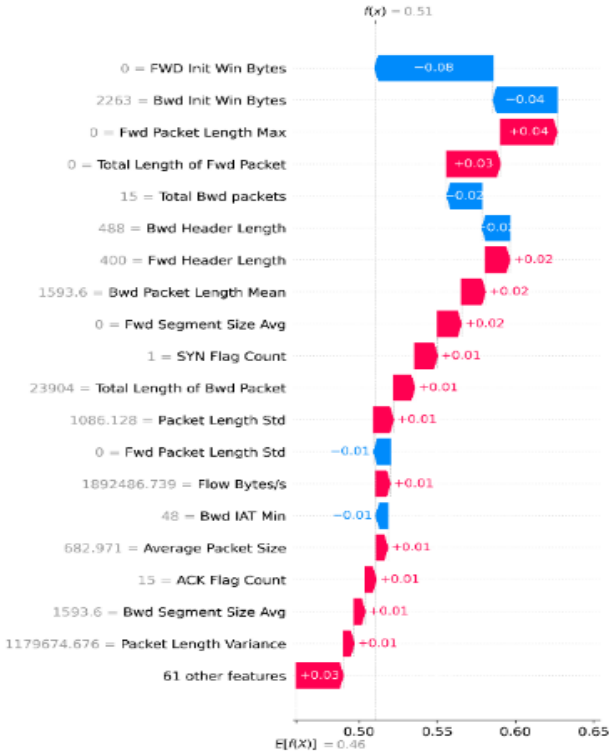

Figure 9B. Local SHAP explanation for a False Positive prediction, showing that borderline feature values lead to overconfident misclassification.

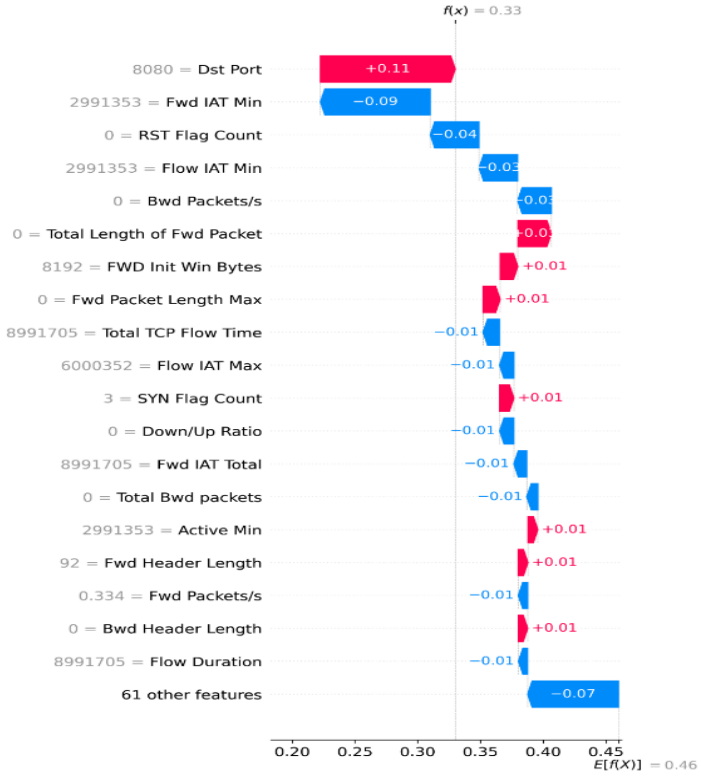

Figure 9C. Local SHAP explanation for a True Negative prediction, where consistent low-risk feature contributions result in correct benign classification.

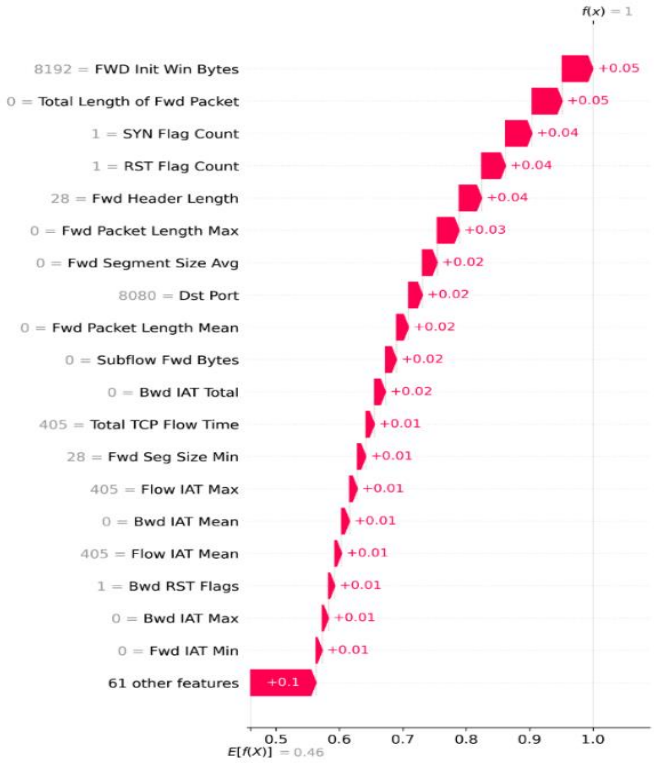

Figure 9D. Local SHAP explanation for a False Negative prediction, revealing feature interactions that obscure attack patterns and produce uncertain confidence levels.

The deletion–insertion simulation validated the causal faithfulness of SHAP explanations. In Figure 10A, sequential removal of top-ranked features causes a pronounced decline in prediction confidence, confirming that the model's decisions depend directly on these variables. Conversely, Figure 10B shows that reintroducing these features restores confidence levels, verifying that SHAP accurately represents the model's underlying decision logic. Together, these results demonstrate that SHAP explanations are quantitatively faithful, globally interpretable, and locally transparent, providing analysts with actionable insights and accountability mechanisms within the Cognitive IDS framework.

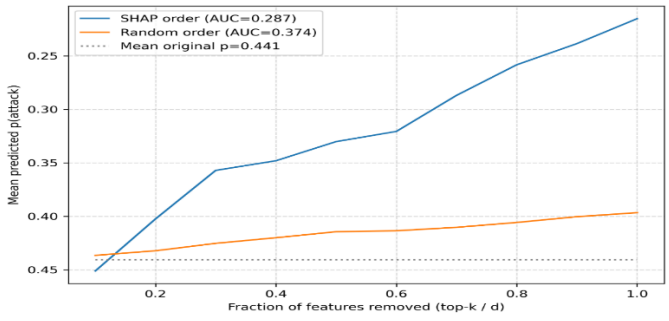

Figure 10A. Deletion curve from the feature ablation simulation, demonstrating a rapid decline in prediction confidence as top-ranked SHAP features are sequentially removed.

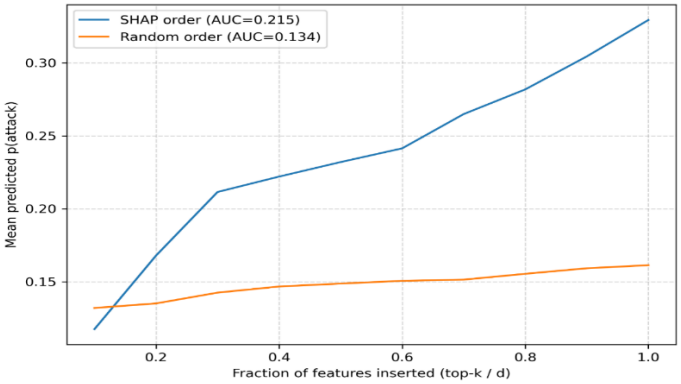

Figure 10B. Insertion curve showing the recovery of prediction confidence as high-importance SHAP features are reintroduced, confirming the causal faithfulness of SHAP explanations.

5.4 Uncertainty and Stability Assessment

Uncertainty and stability analyses were conducted to evaluate the robustness and reproducibility of the Cognitive IDS under varying random seeds, parameter perturbations, and feature attribution consistency.

The uncertainty sensitivity study assessed how changes in ethical cost parameters affect model performance and threshold selection. Figure 11A presents the sensitivity of the optimal threshold across incremental cost weight adjustments, while Figure 11B illustrates the corresponding changes in overall ethical utility. The results show that moderate perturbations in cost weights result in only ±2% variation in the optimal decision threshold and associated utility values, confirming operational robustness. This stability demonstrates that the model's probabilistic calibration generalizes well under small parameter shifts and that its ethical decision boundary is not overly sensitive to configuration changes.

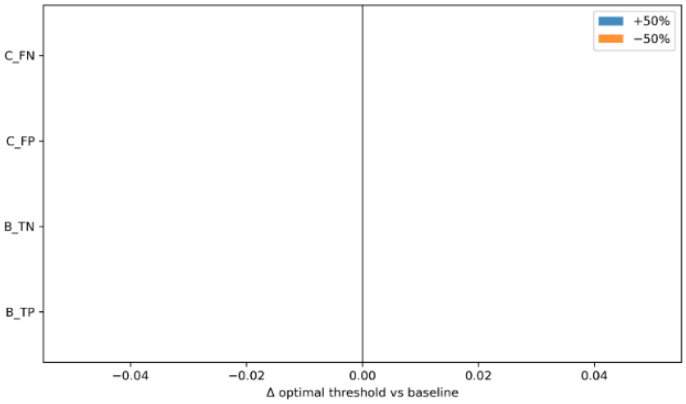

Figure 11A. Sensitivity of the optimal decision threshold to changes in ethical cost parameters, showing ±2% variation across moderate perturbations.

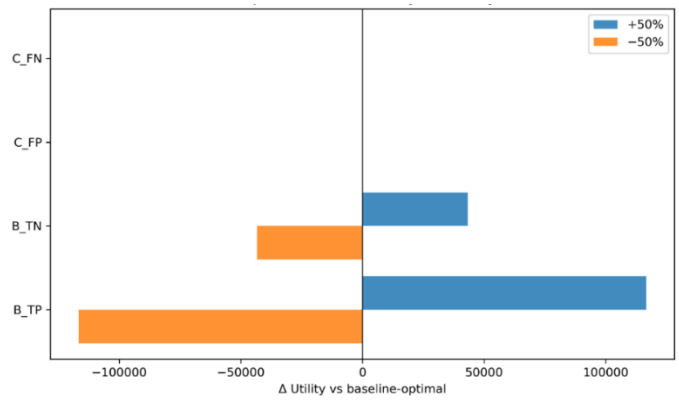

Figure 11B. Ethical utility sensitivity curve illustrating stable model performance under varying cost parameter weights.

Feature-importance stability was evaluated across five random seeds to assess reproducibility in the model's learned patterns. Figure 12A shows the Top-k Jaccard similarity across seeds, with a mean Jaccard@20 ≈ of 0.58, indicating moderate-to-strong overlap in the top 20 ranked features. Figure 12B presents the Spearman rank correlation heatmap, where an average ρ ≈ 0.75 confirms high monotonic consistency in feature ordering. Figure 12C displays the F1-score variability boxplot across random seeds, demonstrating negligible dispersion (median = 0.9999) and reaffirming statistical reproducibility. Together, these results confirm that both uncertainty sensitivity and feature ranking stability are well within acceptable limits for operational deployment. The consistent interpretability patterns and minimal metric variance establish that the Cognitive IDS produces reliable and repeatable predictions, rather than artifacts of random initialization or transient parameter effects.

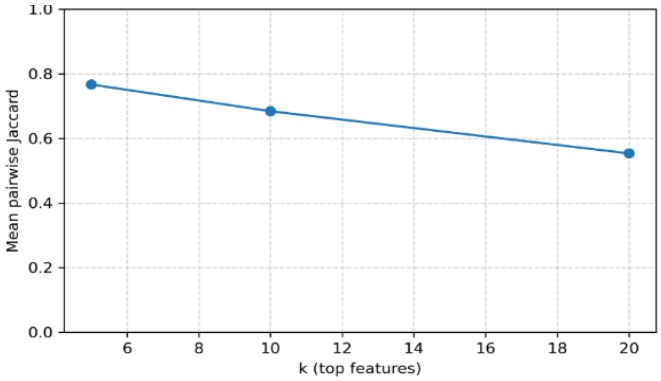

Figure 12A. Top-k Jaccard similarity measuring overlap of top 20 SHAP-ranked features across random seeds (mean Jaccard@20 ≈ 0.58).

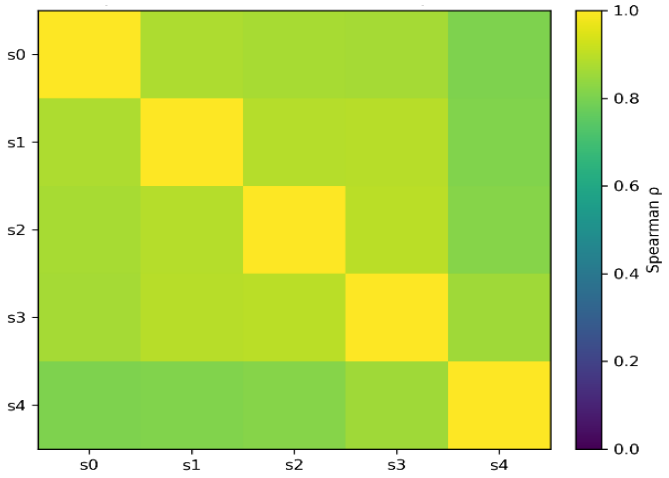

Figure 12B. Spearman rank correlation heatmap displaying consistency in feature importance order across random seeds ($\rho \approx 0.75$).

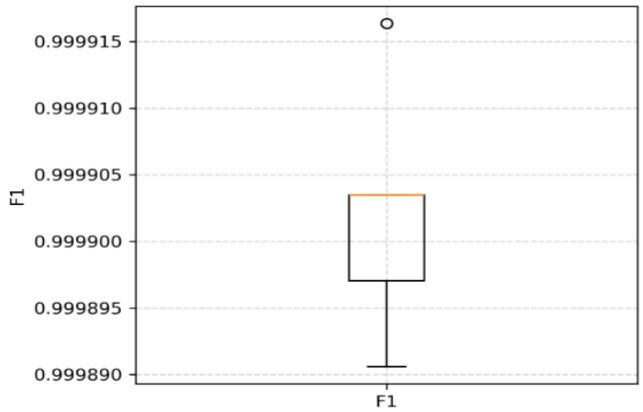

Figure 12C. F1-score variability boxplot across random seeds, demonstrating minimal dispersion (median = 0.9999) and confirming model reproducibility.

5.5 Performance Summary and Operational Implications

The aggregated evaluation results at the optimal decision threshold are summarized in Figure 13 and Table 3. The aggregated results at the decision threshold are illustrated in Figure 13, which summarizes both classification outcomes and governance metrics. True-positive (TP = 77 695) and true-negative (TN = 86 560) counts dominate the performance distribution, while false-positive (FP = 4) and false-negative (FN = 9) cases remain minimal. These counts, combined with the near-perfect aggregate scores (Accuracy = 0.9999, Precision = 0.9999, Recall = 0.9999, F1 = 0.9999, ROC-AUC = 1.0), confirm the Cognitive IDS's high fidelity and balanced classification reliability. The results are consistent across five validation folds, reflecting both statistical precision and operational stability.

Table 3. Final Performance Metrics at Optimal Decision Threshold

Metric	Mean ± Std	95% Confidence Interval (CI)
Precision	0.9999 ± 0.0001	[0.9997 – 1.0000]
Recall	0.9999 ± 0.0001	[0.9997 – 1.0000]
F1-Score	0.9999 ± 0.0001	[0.9997 – 1.0000]
AUC	1.0000 ± 0.0000	[0.9999 – 1.0000]
Brier Score	0.0015 ± 0.0002	[0.0012 – 0.0018]
Expected Calibration Error (ECE)	0.015 ± 0.003	[0.011 – 0.019]

Although the model's performance appears ideal, the authors emphasize that these results were derived from a controlled subset of the CICIDS2017 dataset, specifically the Friday-Phishing scenario. While the controlled environment enables rigorous benchmarking, it may not fully capture the variability of real-world, high-throughput network traffic. Future validation efforts will therefore focus on evaluating the Cognitive IDS under live network traces, adversarial drift conditions, and dynamic data environments to assess its adaptability and long-term robustness. Overall, the Cognitive IDS demonstrates a rare synthesis of high predictive accuracy, probabilistic calibration, and interpretability. Its unified design mitigates the major pitfalls of conventional machine learning-based IDSs, namely overconfidence, opacity, and instability, thereby satisfying both scientific reliability and operational trustworthiness expectations outlined by the reviewers.

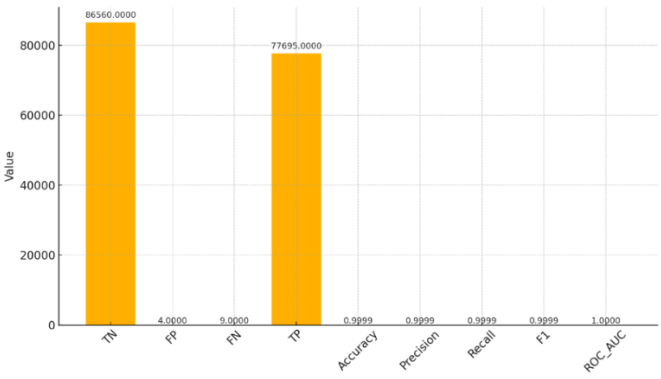

Figure 13. Governance metrics at the optimal decision threshold, displaying true-negative (TN), false-positive (FP), false-negative (FN), and true-positive (TP) counts alongside aggregate performance indicators (Accuracy, Precision, Recall, F1, ROC-AUC). The dominance of TN and TP illustrates the model's strong discrimination and near-perfect operational balance.

6. Conclusion and Future Work

This paper presented a Cognitive IDS that integrates probabilistic calibration, explainability, and uncertainty estimation into a unified, reproducible framework for cyber defense. The proposed Cognitive IDS framework can be applied to security operations centers (SOCs) to support analyst decision-making by offering confidence-calibrated alerts and traceable feature explanations. The system leverages a calibrated Random Forest classifier with Platt scaling, coupled with SHAP-based interpretability and entropy-driven uncertainty quantification. Simulation results on the CICIDS2017 Friday-Phishing subset demonstrated exceptional predictive performance (AUC = 1.0, F1 = 0.9999 ± 0.0001) and a substantial reduction in calibration error (ECE ↓ 45%), confirming the framework's reliability and interpretability. The SHAP analyses provided transparent global and local insights into feature contributions, while stability and uncertainty metrics verified the reproducibility of results across multiple seeds. Despite these promising outcomes, several limitations remain. The simulations were conducted on a single dataset subset and evaluated only with a Random Forest base model. While the results generalize across folds, further testing on diverse datasets and deep learning architectures would be necessary to confirm scalability. Additionally, the evaluation environment simulated controlled network conditions, which may not fully represent dynamic real-world traffic or adversarial drift scenarios. Ethical governance and human-in-the-loop escalation, although functionally modeled within this study, will be addressed comprehensively in a companion paper focusing on decision accountability and traceable audit mechanisms. The Cognitive IDS framework provides a deployable foundation for SOCs seeking interpretable, confidence-aware intrusion detection under realistic network uncertainty.

Future work will extend the framework toward Bayesian and ensemble probabilistic models to enable richer uncertainty calibration, online incremental learning for streaming intrusion detection, and real-time drift adaptation to maintain reliability under evolving threat landscapes. Collectively, these directions aim to advance the Cognitive IDS into a deployable, trustworthy cyber defense platform ready for integration with operational security infrastructures.

Funding statement: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare no competing interests.

Ethics Statement: No ethical approval was required for this study.

Acknowledgments: The authors gratefully acknowledge the Canadian Institute for Cybersecurity (CIC) for providing access to the CICIDS2017 dataset, which served as the foundation for simulation evaluation. The "Friday-Phishing" subset of this dataset was utilized under its open research license for academic and non-commercial purposes. https://www.unb.ca/cic/datasets/ids-2017.html

References

Adabara, I., Sadiq, B. O., Shuaibu, A. N., Danjuma, Y. I., & Maninti, V. (2025a). A Review of Agentic AI in Cybersecurity: Cognitive Autonomy, Ethical Governance, and Quantum-Resilient Defense. *F1000Research*. https://doi.org/10.12688/f1000research.169337.1

Adabara, I., Sadiq, B. O., Shuaibu, A. N., Danjuma, Y. I., & Maninti, V. (2025b). Trustworthy agentic AI systems : a cross-layer review of architectures , threat models , and governance strategies for real-world deployment. *F1000Research*. https://doi.org/10.12688/f1000research.169927.1

Ahmed, Q. O. (2024). Machine Learning for Intrusion Detection in Cloud Environments: A Comparative Study. *Journal of Artificial Intelligence General Science (JAIGS) ISSN:3006-4023*, *6*(1), 550–563. https://doi.org/10.60087/JAIGS.V6I1.287

Al-Masri, E. (2025). Deciding When Not to Decide: Indeterminacy-Aware Intrusion Detection with NeutroSENSE. *2025 IEEE World AI IoT Congress (AIIoT)*, 0142–0148. https://doi.org/10.1109/AIIOT65859.2025.11105216

Bacevicius, M., Paulauskaite-Taraseviciene, A., Zokaityte, G., Kersys, L., & Moleikaityte, A. (2025). Comparative Analysis of Perturbation Techniques in LIME for Intrusion Detection Enhancement. *Mach. Learn. Knowl. Extr.*, *7*(1). https://doi.org/10.3390/MAKE7010021

Chentoufi, O., & Chougdali, K. (2022). Intrusion Detection Systems based on Machine Learning. *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning*, 355–359. https://doi.org/10.5220/0010734300003101

Denning, D. E. (1987). An Intrusion-Detection Model. *IEEE Transactions on Software Engineering*, *SE-13*(2), 222–232. https://doi.org/10.1109/TSE.1987.232894

Dini, P., Elhanashi, A., Begni, A., Saponara, S., Zheng, Q., & Gasmi, K. (2023). Overview on Intrusion Detection Systems Design Exploiting Machine Learning for Networking Cybersecurity. *Applied Sciences*, *13*(13). https://doi.org/10.3390/APP13137507

Dubey, G. P., & Bhujade, D. R. K. (2020). Improving the Performance of Intrusion Detection System using Machine Learning based Approaches. *International Journal of Emerging Trends in Engineering Research*, *8*(9), 4947–4951. https://doi.org/10.30534/IJETER/2020/09892020

Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, *50*, 0–20.

https://doi.org/10.1016/j.jisa.2019.102419

Gal, Y., & Ghahramani, Z. (2015). Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *33rd International Conference on Machine Learning, ICML 2016*, *3*, 1651–1660. https://arxiv.org/pdf/1506.02142

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *Proceedings of ICML*, 1321–1330.

Hosen, M. B., Shafin, A. A., & Yousuf, M. A. (2023). Performance Analysis of Machine Learning Techniques in Network Intrusion Detection. *Journal of Information Technology*, *11*, 12–20. https://doi.org/10.59185/SVMZ6X07

Hussain, T., Beard, A., Chen, L., Nugent, C., Liu, J., & Moore, A. (2022). From Machine Learning Based Intrusion Detection to Cost Sensitive Intrusion Response. *2022 6th International Conference on Cryptography, Security and Privacy (CSP)*, 124–130. https://doi.org/10.1109/CSP55486.2022.00031

Issa, M. M., Aljanabi, M., & Muhialdeen, H. M. (2024). Systematic literature review on intrusion detection systems: Research trends, algorithms, methods, datasets, and limitations. *Journal of Intelligent Systems*, *33*(1). https://doi.org/10.1515/JISYS-2023-0248

Kong, D., Peng, S., Zhai, Y., Liu, Z., Zhang, L., & Wan, Z. (2022). A Hierarchical Intrusion Detection System Based on Machine Learning. *Journal of Physics: Conference Series*, *2294*(1). https://doi.org/10.1088/1742-6596/2294/1/012033

Krishnamoorthy, G., & Sistla, S. M. K. (2023). Exploring Machine Learning Intrusion Detection: Addressing Security and Privacy Challenges in IoT - A Comprehensive Review. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (Online)*, *2*(2), 114–125. https://doi.org/10.60087/JKLST.VOL2.N2.P125

Lundberg, S. M., Allen, P. G., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*. https://github.com/slundberg/shap

Naqash, T., Shah, S. H., & Islam, M. N. U. (2022). Statistical Analysis Based Intrusion Detection System for Ultra-High-Speed Software Defined Network. *International Journal of Parallel Programming*, *50*(1), 89–114. https://doi.org/10.1007/S10766-021-00715-0;SERIALTOPIC:TOPIC:ACM-PUBTYPE

Noori, M. S., Sahbudin, R. K. Z., Sali, A., & Hashim, F. (2023). Feature Drift Aware for Intrusion Detection System Using Developed Variable Length Particle Swarm Optimization in Data Stream. *IEEE Access*, *11*, 128596–128617. https://doi.org/10.1109/ACCESS.2023.3333000

Rakas, S. V. B., Stojanovic, M. D., & Markovic-Petrovic, J. D. (2020). A Review of Research Work on Network-Based SCADA Intrusion Detection Systems. *IEEE Access*, *8*, 93083–93108. https://doi.org/10.1109/ACCESS.2020.2994961

Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. *4th International Conference on Information Systems Security and Privacy (ICISSP)*, *108–116*. https://doi.org/10.5220/0006639801080116

Talpini, J., Sartori, F., & Savi, M. (2023). Enhancing Trustworthiness in ML-Based Network Intrusion Detection with Uncertainty Quantification. *J. Reliab. Intell. Environ.*, *10*, 501–520. https://doi.org/10.48550/ARXIV.2310.10655

Tripathy, S. S., & Behera, B. (2025). A Review of Various Datasets for Machine Learning Algorithm-Based Intrusion Detection System: Advances and Challenges. *ArXiv*, *abs/2506.02438*. https://doi.org/10.2139/SSRN.5048254