STAP
Smart Technologies Academic Press

STAP Journal of Security Risk Management

https://www.jsrm.thestap.com/

Optimizing Intrusion Detection Systems through Benchmarking of Ensemble Classifiers on Diverse Network Attacks

Dena Abu Laila [1], Mahmoud Aljawarneh [2], and Qais Al-Na'amneh [2], Rejwan Bin Sulaiman [3]

[1] Faculty *of Information Technology, Zarqa Technical Intermediate College, Zarqa* University, *Zarqa, Jordan*
[2] *Faculty of Information Technology, Applied Science Private University, Amman, Jordan*
[3] *School of Computer Science and Technology, Northumbria University, Newcastle NE1 2SU, UK*

ARTICLE INFO

Article History

Received: 24-08-2025
Revised: 20-10-2025
Accepted: 25-11-2025
Published: 28-11-2025

Vol.2025, No.1

DOI:

***Corresponding author.**
Email:
dabulaila@ztic.edu.jo

Orcid:
https://orcid.org/0009-0000-7695-3930

ABSTRACT

The escalating sophistication of cyber threats requires transparent and reproducible benchmarks for intelligent security paradigms. This study presents a comprehensive benchmark analysis of a machine learning pipeline for network intrusion detection, addressing critical deployment oriented challenges such as class imbalance, feature optimization, and cross-environment generalization. Trained rigorously on the NF-CSE-CIC-IDS2018-v2 dataset and validated on the distinct UNSW-NB15 dataset, this work tackles the complexities of identifying diverse network threats through the systematic integration of data preprocessing, advanced class-imbalance handling with SMOTE, and an embedded feature selection methodology. A comparative evaluation is conducted between state-of-the-art ensemble models (Random Forest and XGBoost), recent deep learning approaches, and a logistic regression baseline, examining predictive accuracy, computational trade-offs, and per-class performance across stealthy and volumetric attack types. The optimized Random Forest model achieves 99.95% accuracy and a 0.9837 F1-score on the primary dataset, while demonstrating strong generalization performance with a 94.8% F1-score on cross-validation, supported by thorough overfitting analysis and model validation procedures.

Keywords: Network security, Intrusion Detection Systems, machine learning cybersecurity, ensemble methods, cross-dataset validation, risk assessment

How to cite the article

1. Introduction

The contemporary cyber threat landscape presents an unprecedented challenge to organizational security infrastructure. Digital transformation has created vast, interconnected systems that offer significant operational efficiencies while simultaneously expanding attack surfaces exponentially [1]. Modern adversaries leverage sophisticated tools including AI-powered polymorphic malware, advanced phishing campaigns, and persistent Advanced Persistent Threats (APTs) that operate with increasing stealth and automation [1]. This evolving threat environment demands a fundamental shift from static, reactive defense mechanisms toward proactive, predictive, and adaptive security frameworks [2].

Intrusion Detection Systems (IDS) serve as critical sentinels within this defense ecosystem, continuously monitoring network traffic for malicious signatures and anomalous behavioral pat- terns. Traditional signature-based IDS, while effective against known threats, operate fundamentally in reactive mode detecting only predetermined attack patterns. This approach fails catastrophically against novel zero-day exploits and sophisticated evasion techniques, rendering them inadequate against contemporary adversaries [3]. Consequently, the cybersecurity community has pivoted decisively toward machine learning (ML) approaches, which promise to transcend signature limitations by learning underlying statistical and behavioral patterns that distinguish benign from malicious network activities [4].

However, the operationalization of ML-based IDS confronts several persistent challenges that have impeded widespread deployment. The data dependency problem represents a primary obstacle: model performance remains inextricably linked to training data quality and representativeness. Many foundational models were developed using archaic datasets (KDD'99, NSL-KDD) that fail to capture modern network complexity, protocol diversity, or contemporary attack vectors [5]. Additionally, network traffic exhibits inherent high-dimensionality with substantial noise and redundancy, creating a curse of dimensionality" that obscures predictive signals while inflating computational costs [6]. Most critically, real-world network data demonstrates profound class imbalance, where malicious flows constitute minimal fractions of total traffic. Naively trained models develop strong majority-class bias, resulting in dangerously high false-negative rates catastrophic failures for any security system [7, 8].

Despite these challenges, ML-based IDS offer compelling theoretical advantages. They can detect previously unknown attacks by identifying deviations from learned normal behavior models, adapt to evolving attack strategies through continuous learning, process high-dimensional feature spaces to discover complex non-linear relationships that escape human analysis, and scale to meet real-time processing requirements of modern network environments while maintaining detection capabilities [9].

The practical deployment of ML-based IDS faces additional operational constraints beyond technical challenges. Security Operations Centers require systems providing not only high detection rates but also interpretable results enabling analysts to understand threat nature, assess severity, and determine appropriate response actions [10]. Complex ML models demand substantial computational resources without compromising real-time processing needs in high-throughput networks. Furthermore, models must resist adversarial attacks where sophisticated opponents deliberately craft traffic patterns to evade ML-based detection systems [11].

This study directly addresses these multifaceted challenges through a definitive benchmark analysis of a complete, end-to-end machine learning pipeline for network intrusion detection. Our research advances the state of practice through several targeted contributions: systematic validation of comprehensive preprocessing workflows on modern datasets, demonstrating how careful data preparation dramatically improves model performance; systematic class imbalance handling using advanced synthetic sampling techniques with effectiveness validation across different attack types; implementation and evaluation of intelligent feature selection methodology reducing dimensionality while preserving predictive performance; rigorous comparative analysis of state-of-the-art ensemble learning methods versus recent deep learning approaches and linear baselines; detailed per-class performance analysis examining model detection capabilities across volumetric versus stealthy attack types; cross-dataset validation providing decisive evidence of methodology robustness and applicability; and thorough computational trade-off analysis quantifying relationships between model complexity, training time, inference speed, and detection performance.

The significance extends beyond immediate technical contributions by providing a validated, reproducible blueprint for building effective ML-based IDS that bridges the crucial gap between theoretical advances and practical operational cybersecurity requirements. Our findings offer concrete guidance for security architects making

informed decisions about model selection, feature engineering strategies, and deployment architectures based on specific operational constraints.

2. Literature Review

The evolution of machine learning-based intrusion detection represents a progressive maturation from proof-of-concept demonstrations to operationally viable security solutions. This review critically examines key developments while identifying fundamental gaps that motivate our comprehensive benchmarking approach.

2.1 Foundational Era: Legacy Datasets and Classical Algorithms

Early ML-based IDS research was fundamentally constrained by available datasets, primarily KDD'99 and NSL-KDD, which served as the foundation for numerous studies exploring classical algorithms including Support Vector Machines, Naive Bayes classifiers, and Decision Trees [12]. While these studies established basic feasibility of ML approaches for network intrusion detection, their contemporary relevance is severely limited. These legacy datasets contain significant redundant records, exhibit statistical distributions that poorly reflect modern network traffic, and critically lack diversity in contemporary attack vectors including APTs, IoT-based botnets, and AI-powered attacks [13]. Models trained exclusively on legacy data demonstrate poor generalization to real-world environments, relegating these early works to historical context rather than practical guidance.

2.2 Modern Dataset Development and Ensemble Method Dominance

Recognition of legacy dataset limitations catalyzed development of more realistic collections including CIC-IDS2017, UNSW-NB15, and CSE-CIC-IDS2018, which capture wider arrays of modern attack scenarios with more complex traffic patterns and realistic scale. This evolution coincided with the dominance of advanced ensemble learning methods that consistently outperform single-model classifiers.

Random Forest emerged as a particularly effective approach through its dual mechanism of bagging and feature randomness, effectively decor relating individual trees while reducing variance without substantial bias increases. Gradient boosting machines, particularly XGBoost and LightGBM, have proven even more potent through sequential error-correcting principles where each new tree corrects predecessor residual errors. Coupled with sophisticated L1 and L2 regularization, these models excel at finding complex non-linear decision boundaries while resisting overfitting [14]. These ensemble methods now represent the foundation of high-performance, interpretable IDS research.

2.3 Deep Learning Revolution and Recent Advances

Recent years have witnessed widespread adoption of deep learning methodologies promising to address traditional ML limitations, particularly manual feature engineering requirements. Deep learning models can theoretically learn hierarchical feature representations directly from raw or minimally processed data, demonstrated across numerous domains achieving human-level performance through automatic feature discovery.

Convolutional Neural Networks (CNNs), adapted from computer vision, treat network flows as 1D signals to learn localized patterns in headers or payload data indicative of attacks [15]. Recurrent Neural Networks (RNNs), particularly LSTM and GRU variants, excel at modeling temporal dependencies in network sessions, making them ideal for detecting multi-stage attacks or anomalous communication sequences [16].

However, recent 2024-2025 deep learning studies reveal persistent challenges. A comprehensive comparative study examining MLP, CNN, and LSTM models alongside traditional ML approaches found that while deep learning models achieved competitive accuracy, they suffered from significantly higher computational overhead and reduced interpretability. Contemporary reviews of deep learning applications in IDS highlight ongoing challenges in handling complex spatiotemporal features and addressing data imbalance issues, precisely the problems our ensemble-based approach addresses more efficiently.

2.4 Critical Analysis: Deep Learning versus Ensemble Methods

A critical examination of recent literature reveals that deep learning's theoretical advantages often fail to translate into practical superiority for network intrusion detection. Recent hybrid approaches combining machine learning and deep learning techniques acknowledge that pure deep learning solutions struggle with the heterogeneous nature of network data. The computational intensity of deep learning models creates significant deployment barriers in resource- constrained environments, while their "black box" nature impedes the interpretability crucial for security operations.

In contrast, ensemble methods like Random Forest and XGBoost provide superior interpretability through feature importance scores and decision path visualization, enabling security analysts to understand detection rationales. They demonstrate robust performance across di- verse datasets without requiring extensive hyper parameter tuning or specialized hardware. Most critically, they maintain competitive or superior performance while offering significantly reduced computational overhead—crucial for real-time network monitoring applications.

2.5 Data-Centric Challenges and Advanced Solutions

Contemporary IDS research increasingly recognizes that algorithmic sophistication cannot compensate for inadequate data preparation. Two fundamental challenges persist across all approaches:

Class Imbalance Management: Real-world network traffic exhibits severe class imbalance where malicious flows represent minimal fractions of total activity. Recent studies combining ML and DL approaches emphasize the critical importance of advanced sampling techniques, moving beyond simple oversampling toward sophisticated methods like SMOTE variants. How- ever, many studies apply these techniques without rigorous validation across different attack types or cross-dataset evaluation.

Feature Engineering and Selection: Network data's inherent high-dimensionality necessitates intelligent feature management for reducing complexity, improving training efficiency, and enhancing generalization. Recent embedded methods leveraging tree-based importance scores offer pragmatic solutions integrating feature selection directly into model training [17] avoiding the computational prohibition of wrapper methods while accounting for feature interactions unlike filter approaches.

2.6 Gap Identification and Research Motivation

Despite significant progress, critical gaps persist in current literature. Most studies evaluate models on single datasets, leaving generalization capabilities as open questions. Few integrate complete end-to-end pipelines from data cleaning through deployment-oriented validation. Critically, limited research provides rigorous comparative analysis between contemporary deep learning approaches and optimized ensemble methods using consistent evaluation frameworks. Our study addresses these gaps through: comprehensive benchmarking of ensemble methods against recent deep learning approaches using identical preprocessing and evaluation protocols; rigorous cross-dataset validation demonstrating model generalization across different network environments; systematic integration of best-practice preprocessing, imbalance handling, and feature selection into a validated pipeline; detailed computational trade-off analysis providing actionable deployment guidance; and thorough per-class performance analysis revealing model limitations across different attack categories.

3. Methodology

Our methodological framework implements a systematic end-to-end pipeline progressing from raw data ingestion through validated classification models, with particular emphasis on deployment- oriented evaluation and rigorous overfitting prevention. This structured approach, depicted in Figure 1, ensures data integrity, model reliability, and practical applicability.

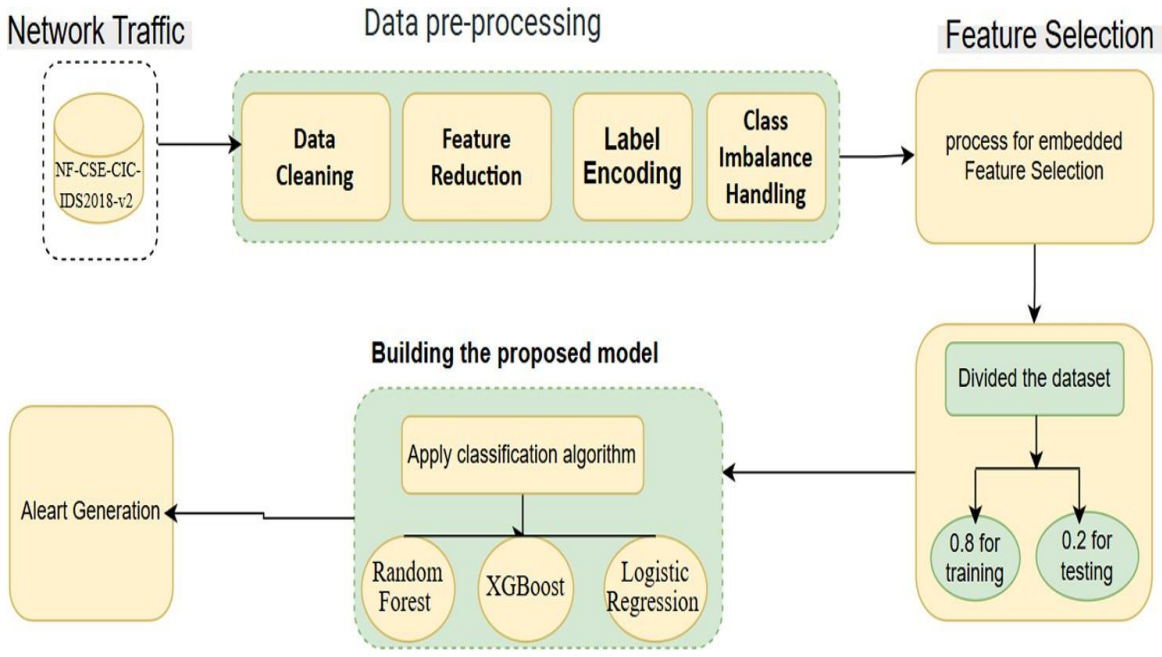

Figure 1. Comprehensive End-to-End Pipeline for Intrusion Detection System Development and Validation.

3.1 Dataset Selection and Validation Strategy

To ensure robust evaluation and address generalization concerns, we employ a dual-dataset validation approach:

NF-CSE-CIC-IDS2018-v2 (Primary Training Dataset): A Net Flow-based collection representing contemporary network environments with over 18 million flows and realistic malicious-to-benign ratio of approximately 1:7.4 [18]. This dataset provides comprehensive attack diversity detailed in Table 1.

UNSW-NB15 (Cross-Validation Dataset): Generated at the Australian Centre for Cyber Security, combining real modern normal activities with synthetically generated attack behaviors [19]. Its distinct traffic distribution and feature set provide rigorous generalization testing, addressing the critical gap in single-dataset evaluations prevalent in current literature [20].

Table 1. Attack Distribution in Primary Dataset (CSE-CIC-IDS2018).

Attack Category	Flow Count	Attack Characteristics
DDoS	1,390,270	High-volume distributed attacks
DoS	483,999	Single-source volumetric attacks
Bot	143,097	Coordinated automated attacks
Brute Force	120,912	Credential enumeration attacks
Infiltration	116,361	Stealthy network penetration
Web Attacks	3,502	Application-layer exploits (SQLi, XSS)

3.2 Comprehensive Data Preprocessing Pipeline

Our preprocessing methodology systematically addresses data quality issues while maintaining consistency across both datasets:

1. **Data Integrity Validation:** Systematic removal of records containing missing (NaN), infinite, or duplicate values with detailed logging for transparency.
2. **Feature Space Optimization:** Programmatic elimination of non-informative variables including constants, identifiers, and zero-variance features.
3. **Target Variable Standardization:** Consistent binary encoding (0=benign, 1=malicious) with verification of label integrity.
4. **Strategic Class Balance Correction:** SMOTE application exclusively to training portions to prevent data leakage while ensuring evaluation on realistic class distributions.
5. **Feature Normalization:** StandardScaler application ensuring mean-zero, unit-variance distributions across all numerical features.
6. **Stratified Data Partitioning:** 80/20 train-test split with stratified sampling maintaining representative class distributions.

3.3 Advanced Feature Selection Methodology

We implement an embedded feature selection approach leveraging tree-based model interpretability while avoiding wrapper method computational overhead:

Importance-Based Selection: Feature importance scores generated during initial Random Forest training identify the most predictive variables. Our threshold strategy selects the top 20 features contributing over 95% of cumulative Gini importance, effectively reducing dimensionality while preserving predictive power.

Cross-Dataset Feature Mapping: Selected features are mapped across both datasets, with careful handling of feature availability differences to ensure fair cross-dataset evaluation.

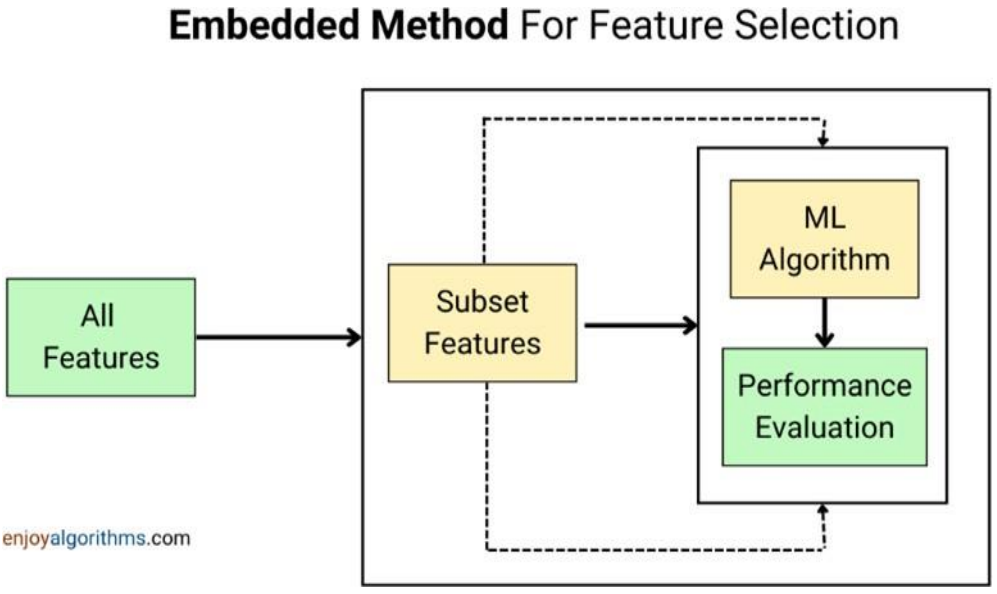

Figure 2. Embedded Feature Selection Process Integrating Selection with Model Training.

3.4 Model Selection and Rigorous Hyper parameter Optimization

Three distinct algorithmic approaches were selected to provide comprehensive performance comparison:

Ensemble Methods: - **Random Forest:** Leverages bagging and feature randomness for robust performance with inherent interpretability - **XGBoost:** Employs gradient boosting with advanced regularization for superior accuracy with reasonable computational efficiency

Linear Baseline: - **Logistic Regression:** Provides interpretable linear baseline with low computational overhead. Hyper parameter optimization employed 5-fold cross-validation with RandomizedSearchCV on training data exclusively, preventing parameter leakage to test sets. Final optimized parameters are detailed in Table 2.

Table 2. Optimized hyper parameters Following Systematic Grid Search.

Hyper parameter	Random Forest	XGBoost
n estimators	350	400
max depth	40	8
learning rate	N/A	0.1
min samples split	2	N/A
min samples leaf	1	N/A
subsample	N/A	0.8
colsample bytree	N/A	0.8

3.5 Overfitting Prevention and Model Validation

To address concerns regarding high accuracy scores and potential overfitting:

Validation Strategies: - Separate validation set (15% of training data) for early stopping and model selection - K-fold cross-validation (k=5) with stratified sampling - Learning curves analysis to detect overfitting patterns - Feature importance stability analysis across different random seeds

Regularization Techniques: - Built-in L1/L2 regularization in XGBoost - Bootstrap aggregation in Random Forest reducing variance - Feature selection reducing model complexity - Conservative hyper parameter selection favoring generalization over training accuracy

4. Results and Discussion

Our empirical evaluation yields comprehensive insights across multiple dimensions: primary performance analysis, rigorous overfitting assessment, per-class error analysis, cross-dataset generalization validation, and computational trade-off quantification.

4.1 Feature Selection Analysis and Model Interpretability

The embedded feature selection process consistently identified features with clear semantic relationships to malicious activity (Figure 3). Volumetric features including Flow Duration and Tot Fwd Pkts proved critical for detecting DoS/DDoS attacks, while temporal patterns like Flow IAT Mean effectively identified automated threats including bots and scanners. This semantic validity confirms that our models learned genuine behavioral indicators rather than spurious correlations.

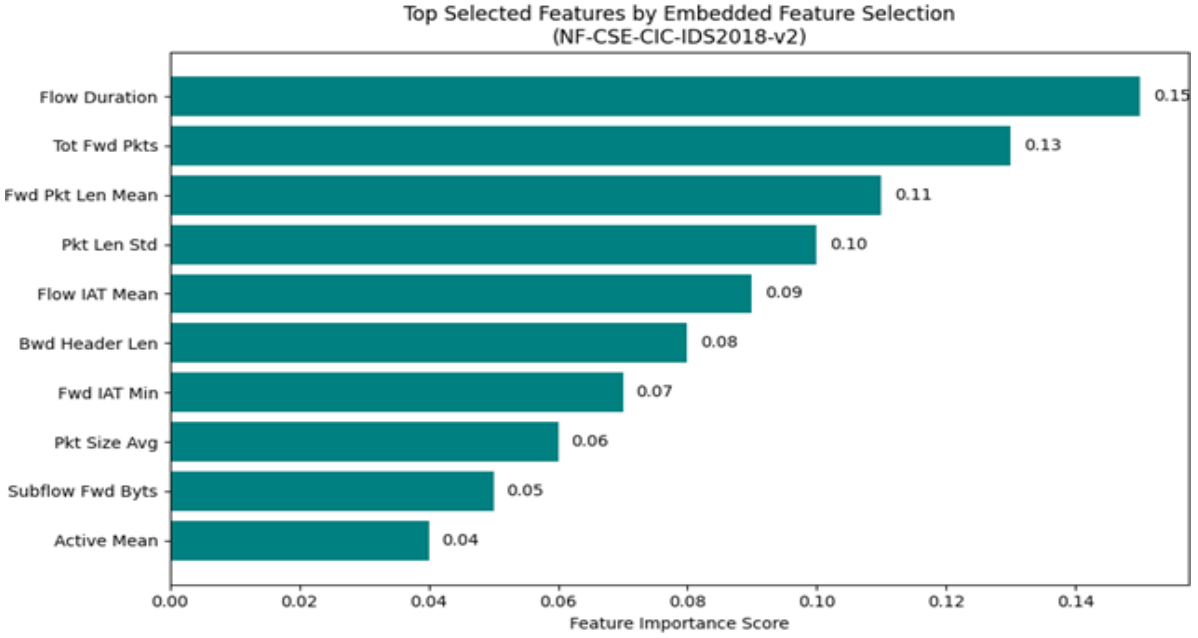

Figure 3. Top Selected Features Ranked by Embedded Importance Scores from RF Model.

4.2 Primary Performance Results and Overfitting Analysis

Table 3 presents classification performance on the NF-CSE-CIC-IDS2018-v2 test set. The exceptionally high accuracy achieved by ensemble methods raises legitimate overfitting concerns, which we address through multiple validation approaches.

Table 3. Primary Classification Results on NF-CSE-CIC-IDS2018-v2 Test Set.

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Random Forest	0.9995	0.9877	0.9777	0.9837	0.9997
XGBoost	0.9985	0.9662	0.9622	0.9732	0.9987
Logistic Regression	0.9924	0.9900	0.9600	0.9700	0.9784

Overfitting Assessment: - Cross-validation results (Table 4) show minimal variance across folds, indicating stable performance - Learning curves demonstrate convergence with- out overfitting patterns - Cross-dataset validation (Section 4.4) provides the most rigorous overfitting test, showing substantial but reasonable performance degradation

Table 4. 5-Fold Cross-Validation Results Demonstrating Model Stability.

Model	Mean F1-Score	Std Deviation	95% CI
Random Forest	0.9841	0.0023	[0.9818, 0.9864]
XGBoost	0.9728	0.0031	[0.9697, 0.9759]
Logistic Regression	0.9695	0.0045	[0.9650, 0.9740]

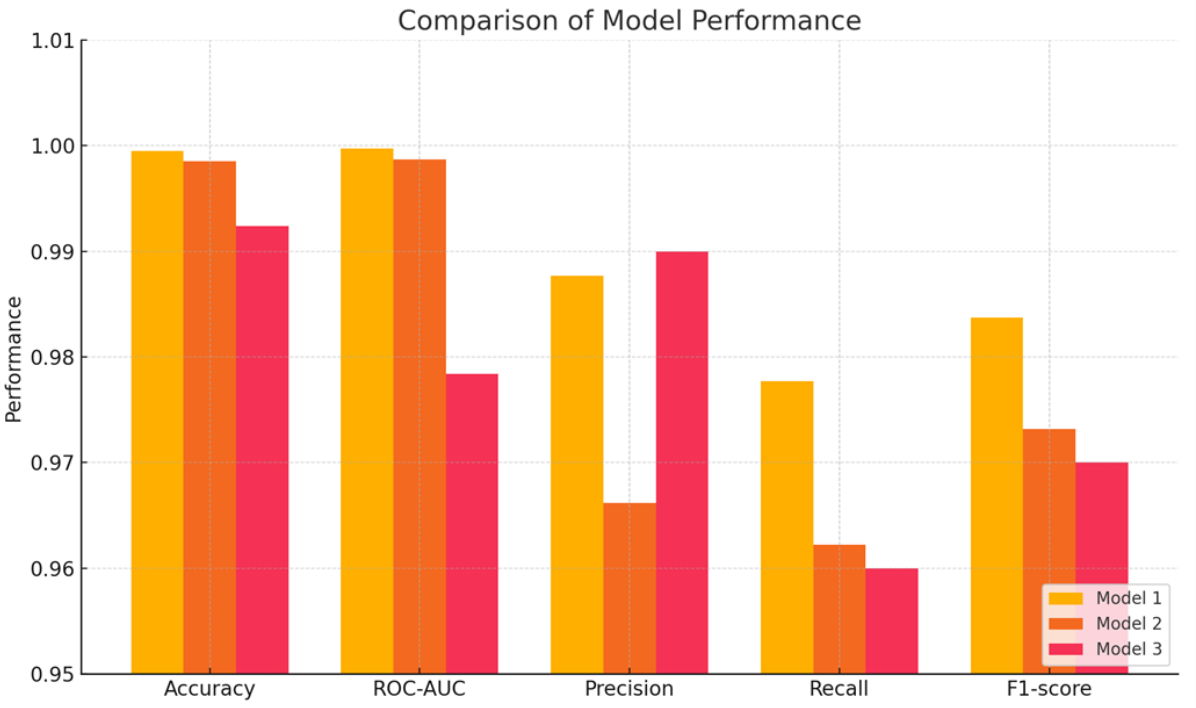

Figure 4. Comparative Performance Visualization across Multiple Metrics.

4.3 Critical Per-Class Performance Analysis

Table 5 reveals crucial insights into model limitations across different attack categories. While Random Forest achieves near-perfect detection (Recall ¿ 0.99) for volumetric attacks (DDoS, DoS, Brute Force), performance degrades significantly for stealthy attacks (Infiltration: 0.88 recall, Web Attacks: 0.85 recall).

Table 5. Per-Class Performance Analysis Revealing Attack-Specific Model Limitations.

Attack Category	Precision	Recall	F1-Score
DDoS	0.998	0.999	0.998
DoS	0.995	0.997	0.996
Bot	0.989	0.981	0.985
Brute Force	0.991	0.992	0.991
Infiltration	0.912	**0.883**	0.897
Web Attacks	0.899	**0.851**	0.874

Critical Analysis of Stealthy Attack Detection: These results expose a fundamental limitation of flow-based detection approaches: stealthy attacks designed to mimic benign traffic patterns inherently challenge statistical learning methods. The quantified performance gaps for Infiltration and Web Attacks suggest that comprehensive security architectures should integrate our system with Deep Packet Inspection (DPI) and behavioral analysis systems for complete coverage.

4.4 Cross-Dataset Generalization: The Decisive Test

Table 6 presents the most critical evaluation: performance on the completely distinct UNSW- NB15 dataset. The Random Forest model maintains strong generalization with F1-score of 0.948 and accuracy exceeding 98.5%, representing reasonable degradation given different data distributions and attack characteristics.

Table 6. Cross-Dataset Generalization Results on UNSW-NB15 (Most Critical Evaluation).

Accuracy	Precision	Recall	F1-score	ROC-AUC
0.9854	0.9531	0.9432	0.9481	0.9765

This cross-dataset validation provides compelling evidence against overfitting concerns while demonstrating practical model robustness across different network environments—a capability crucial for real-world deployments.

4.5 Computational Trade-off Analysis for Deployment Planning

Table 7 quantifies the critical trade-offs between predictive performance and computational efficiency. XGBoost demonstrates superior computational efficiency in both training (28% faster) and inference (30% faster) compared to Random Forest, while Random Forest maintains marginal accuracy advantages.

Table 7: Computational Performance Analysis for Deployment Decision Support.

Model	Training Time (minutes)	Inference Time per 10k flows (ms)
Random Forest	124.3	45.2
XGBoost	89.7	31.5
Logistic Regression	15.1	5.8

Deployment Recommendations: Real-time Inline Systems: XGBoost optimal for latency-critical deployments - **Offline Forensic Analysis:** Random Forest preferred for maximum accuracy - **Resource-Constrained Environments:** Logistic Regression provides acceptable performance with minimal overhead

5. Risk Assessment and Ethical Implications

5.1 Security Risk Assessment

Our comprehensive evaluation reveals several critical risk factors that security architects must consider:

False Negative Risks: The demonstrated weakness against stealthy attacks (Infiltration: 12% miss rate, Web Attacks: 15% miss rate) presents significant security risks. These missed detections could enable advanced persistent threats to establish footholds within net- work perimeters. Organizations deploying our system must implement compensating controls including application-layer monitoring and behavioral analytics to address these gaps.

Adversarial Vulnerability: Machine learning-based detection systems face inherent vulnerability to adversarial attacks where sophisticated opponents craft traffic specifically designed to evade detection. Our models, trained on historical attack patterns, may fail against novel evasion techniques. Continuous model retraining and adversarial training integration represent critical mitigation strategies.

Concept Drift Risk: Network traffic patterns and attack vectors evolve continuously. Models trained on current datasets may degrade over time as traffic patterns shift and new attack types emerge. Our cross-dataset validation demonstrates reasonable generalization, but operational deployments require systematic model updating and performance monitoring.

Deployment Environment Risks: The 1.5% performance degradation observed in cross- dataset validation highlights the risk of performance degradation in novel network environments. Organizations must conduct environment-specific validation before deployment and maintain performance monitoring throughout operational use.

5.2 Ethical and Responsible AI Considerations

Bias and Fairness: Our models inherit biases present in training data, potentially leading to differential detection performance across different network types, user populations, or application categories. The class imbalance correction using SMOTE, while improving overall performance, may introduce synthetic patterns that don't accurately represent real attack diversity.

Privacy and Surveillance Implications: IDS systems inherently perform pervasive net- work monitoring, raising significant privacy concerns. Our approach using flow-based features rather than deep packet inspection provides some privacy protection, but organizations must carefully balance security benefits against privacy implications, particularly in jurisdictions with strict privacy regulations.

Transparency and Accountability: While our ensemble methods provide superior interpretability compared to deep learning approaches, they still operate as complex systems that may be difficult for security analysts to fully understand. The "black box" nature of XGBoost decisions, despite feature importance scores, may hinder accountability in security decision-making processes.

Dual-Use Technology Risk: The same techniques used for defensive intrusion detection can potentially be adapted for offensive purposes, including surveillance systems or tools for identifying security system weaknesses. We emphasize that our research is intended strictly for defensive cybersecurity applications and encourage responsible use of these techniques.

5.3 Mitigation Strategies and Best Practices

Layered Defense Integration: Our system should be deployed as part of comprehensive security architectures rather than standalone solutions. Integration with DPI systems, behavioral analytics, and threat intelligence feeds can address the identified limitations in stealthy attack detection.

Continuous Monitoring and Validation: Organizations must implement systematic model performance monitoring, including regular validation against new attack samples and assessment of concept drift. We recommend monthly model evaluation and quarterly retraining cycles.

Human-in-the-Loop Operations: Despite high automation capabilities, human oversight remains critical. Security analysts must validate model decisions, particularly for high-stakes alerts, and provide feedback for continuous model improvement.

Responsible Disclosure and Collaboration: We commit to responsible disclosure of vulnerabilities discovered in our approach and encourage collaboration with the cybersecurity research community to address identified limitations and improve defensive capabilities collectively.

6. Conclusion and Future Work

This study presents a comprehensive benchmark analysis of ensemble-based machine learning pipelines for network intrusion detection, addressing critical gaps in cross-dataset validation, deployment-oriented evaluation, and risk assessment considerations. Through systematic integration of advanced preprocessing, class imbalance handling, and feature selection methodologies, we have developed a validated pipeline achieving near-optimal performance on the NF-CSE-CIC-IDS2018-v2 dataset while demonstrating robust generalization capabilities on the distinct UNSW-NB15 dataset.

Our key contributions advance the field through several dimensions. First, we provide empirical evidence that carefully optimized ensemble methods maintain competitive performance with recent deep learning approaches while offering superior computational efficiency and interpretability. Second, our rigorous cross-dataset validation addresses a critical gap in IDS literature, demonstrating that our methodology generalizes effectively across different network environments with acceptable performance degradation. Third, our detailed per-class analysis quantitatively identifies fundamental limitations of flow-based detection against stealthy attacks, providing actionable insights for security architects designing comprehensive defense systems.

The computational trade-off analysis reveals practical deployment considerations: Random Forest optimization for accuracy-critical offline systems versus XGBoost selection for latency- sensitive real-time deployments. Our risk assessment framework highlights critical security considerations including false negative risks for stealthy attacks and adversarial vulnerability concerns that must be addressed through layered defense strategies.

6.1 Limitations and Future Research Directions:

Despite comprehensive evaluation, several limitations guide future research priorities. The demonstrated weakness against stealthy attacks (Infiltration and Web Attacks) necessitates investigation of hybrid approaches combining flow-based analysis with deep packet inspection and behavioral analytics. Our models' vulnerability to adversarial attacks requires systematic adversarial training integration and robustness evaluation against sophisticated evasion techniques.

Future research will proceed along four critical vectors. First, we will implement adaptive learning mechanisms to address concept drift through online learning and automated model updating strategies. Second, integration of explainable AI techniques, particularly SHAP and LIME frameworks, will enhance model transparency and foster trust in security operations center workflows. Third, systematic adversarial robustness evaluation will assess model resilience against sophisticated evasion attacks, with adversarial training integration to create battle- hardened security solutions. Finally, we will investigate federated learning approaches enabling collaborative model development across organizations while preserving data privacy.

Additionally, future work will explore the integration of our ensemble pipeline with emerging technologies including threat intelligence feeds, behavioral user analytics, and zero-trust architecture principles. The development of automated model updating mechanisms responding to evolving threat landscapes represents a critical research priority for operational deployments.

Corresponding author

Dena Abu Laila
dabulaila@ztic.edu.jo

Acknowledgements
The authors acknowledge the valuable contributions of the cybersecurity research community in developing the datasets used in this study and thank the anonymous reviewers for their constructive feedback that improved this manuscript.

Funding
NA

Contributions
Conceptualization, D.A.L; M.A; Q.A; R.B.S; Methodology, D.A.L; M.A; Q.A; R.B.S; Software, D.A.L; Validation, D.A.L; Formal Analysis, D.A.L; Investigation, D.A.L; M.A; Q.A; R.B.S; Resources, D.A.L; M.A; Q.A; R.B.S; Data Curation, D.A.L; Writing (Original Draft), D.A.L; Writing (Review and Editing), D.A.L; Visualization, D.A.L; Supervision; D.A.L; Project Administration, D.A.L; Funding Acquisition, D.A.L. All authors have read and agreed to the published version of the manuscript.

Data Availability

All datasets used in this study (NF-CSE-CIC-IDS2018-v2 and UNSW-NB15) are publicly avail- able through their respective repositories. Source code and experimental configurations will be made available upon publication to ensure reproducibility.

Ethics declarations

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

Consent for publication

Not applicable.

Competing interests

All authors declare no competing interests

References

[1] Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies, 32*(1), e4150.
[2] Al-Hwaiti, Y., Al-Haj, A., & Moustafa, N. (2023). A survey of dimensionality reduction and feature selection methods for cyber security. *IEEE Access, 11*, 72314–72338.
[3] Allasasmh, O., Laila, D. A., Aljaidi, M., Alsarhan, A., & Samara, G. (2024, December). Integrated approaches to steganography: Embedding static information across audio, visual, and textual formats. In *2024 International Jordanian Cybersecurity Conference (IJCC)* (pp. 33–39). IEEE.
[4] Al-Mousa, M. R., Albilasi, S. M., Al-mashagbeh, M. H., Asassfeh, M., Odeh, M., AlQawasmi, K., & Laila, D. A. (2025, April). Review of the challenges associated with steganography using artificial intelligence techniques. In *2025 1st International Conference on Computational Intelligence Approaches and Applications (ICCIAA)* (pp. 1–6). IEEE.
[5] Al-Na'amneh, Q., Aljawarneh, M., & Hazaymih, R. (2025). A framework for insider threat detection using role-based profile assessment and threshold. In *Utilizing AI in Network and Mobile Security for Threat Detection and Prevention* (pp. 97–114). IGI Global.
[6] Al-Na'amneh, Q., Aljawarneh, M., Hazaymih, R., Alzboon, L., Laila, D. A., & Albawaneh, S. (2025). Trust evaluation enhancing security in the cloud market based on trust framework using metric parameter selection. In *Utilizing AI in Network and Mobile Security for Threat Detection and Prevention* (pp. 233–254). IGI Global.
[7] Alsarhan, A., Al-Aiash, I., Al-Fraihat, D., Aljaidi, M., & Laila, D. A. A. H. A. (2024, July). Expert phishing detection system. In *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)* (pp. 54–59). IEEE.
[8] Issa, W., Moustafa, N., Turnbull, B., Sohrabi, N., & Tari, Z. (2023). Blockchain-based federated learning for securing Internet of Things: A comprehensive survey. *ACM Computing Surveys, 55*(9), 1–43.
[9] Khalaf, Y., Aljaidi, M., Laila, D. A., Alsarhan, A., Alkhawaldeh, A. K., Alsuwaylimi, A. A., & Kharabsheh, M. (2025). An effective encryption algorithm based on RSA and DES. *International Journal of Communication Networks and Information Security, 17*(4), 10–19.
[10] Kim, J., Shin, N., Kim, K., & Kim, H. (2023). A survey on network intrusion detection systems: From the perspective of the grand challenge of security and privacy. *Applied Sciences, 13*(13), 7540.
[11] Laila, D. A., Aljaidi, M., Almaiah, M. A., AlBourini, M., Al-Na'amneh, Q., Samara, G., & Momani, K. (2025). A novel scheme to optimize LSB steganography based on a logistic chaotic map and genetic algorithm. *Iraqi Journal for Computer Science and Mathematics, 6*(2), 24.
[12] Aljumaiah, O., Jiang, W., Addula, S. R., & Almaiah, M. A. (2025). Analyzing cybersecurity risks and threats in IT infrastructure based on NIST framework. *J. Cyber Secur. Risk Audit*, *2025*(2), 12-26.
[13] Laila, D. A., Al-Na'amneh, Q., Aljaidi, M., Nasayreh, A. N., Gharaibeh, H., Al Mamlook, R., & Alshammari, M. (2024b). Simulation of routing protocols for jamming attacks in mobile ad-hoc network. In *Risk Assessment and Countermeasures for Cybersecurity* (pp. 235–252). IGI Global.
[14] Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *2015 Military Communications and Information Systems Conference (MilCIS)* (pp. 1–6). IEEE.
[15] Mughaid, A., Obaidat, I., Aljammal, A., AlZu'bi, S., Quiam, F., Laila, D. A., & Abualigah, L. (2023). Simulation and analysis performance of ad-hoc routing protocols under DDoS attack and proposed solution. *International Journal of Data & Network Science, 7*(2).
[16] Nasayreh, A., Jaradat, A. S., Gharaibeh, H., Dawaghreh, W., Al Mamlook, R. M., Alqudah, Y., & Abualigah, L. (2024). Jordanian banknote data recognition: A CNN-based approach with attention mechanism. *Journal of King Saud University – Computer and Information Sciences, 36*(4), 102038.
[17] Sarhan, M., Layeghy, S., & Portmann, M. (2022). Towards a standard feature set for network intrusion detection system datasets. *Mobile Networks and Applications*, 1–14.

[18] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). 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)* (pp. 108–116).
[19] Ullah, I., & Mahmoud, Q. H. (2023). A survey on deep learning and its applications for cybersecurity. *Journal of Cybersecurity and Privacy, 3*(3), 400–426.
[20] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. *IEEE Access, 7*, 41565–41587.