Received: 30 November 2025 | Revised: 16 March 2026 | Accepted: 3 June 2026 | Published online: 18 June 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–10
DOI: 10.47852/bonviewAIA62028592

BON VIEW PUBLISHING

Late-Fusion Stacking of Machine Learning and Snort Signatures for Multi-Class Intrusion Detection on UNSW-NB15

Somayina C. Wen-Udeoji[1], Maybin K. Muyeba[1,2,*] and Azadeh Mohammadi[1,2]

[1] *School of Science, Engineering and Environment, University of Salford, UK*

[2] *Data Science and AI (DSAI) Hub, University of Salford, UK*

Abstract: This extended study builds on a previously published ICDSAIA 2025 paper by developing a hybrid intrusion-detection architecture that integrates machine learning (ML) classifiers with Snort-derived signature metadata. Using the UNSW-NB15 dataset and a three-class threat mapping, optimized ML baselines, including XGBoost, Random Forest, and Decision Tree models, are first evaluated to establish a consistent performance benchmark. Two hybrid approaches are then introduced: a feature-level model that appends Snort alert indicators to the ML feature set and a decision-level stacking model that combines calibrated XGBoost probability outputs with engineered Snort meta-features. Experimental results show that feature-level hybridization provides limited benefit, whereas the proposed stacking architecture delivers consistent and meaningful improvements, achieving an accuracy of 0.8346 and a macro-F1 score of 0.8347, outperforming all baseline models. The approach notably improves detection performance for Normal and Low-Threat classes by effectively combining ML generalization with the precision of signature-based detection. These findings demonstrate that late-fusion integration of ML and Snort signals offers a more robust and practical solution for multi-class intrusion detection compared to standalone or feature-level approaches.

Keywords: Intrusion Detection System (IDS), machine learning, Snort, hybrid fusion, XGBoost

1. Introduction

Intrusion Detection Systems (IDSs) remain an essential component of modern cyber-defense designs, particularly because network environments continue to grow in scale, complexity, and exposure to evolving cyber threats. Traditional signature-based Network Intrusion Detection Systems (NIDS), such as Snort and Suricata, provide high precision for previously known attacks but are inherently limited in detecting polymorphic, obfuscated, or zero-day threats due to their reliance on manually engineered rules [1, 2]. However, machine learning (ML)-driven intrusion detection has gained attention for its ability to generalize from data, identify subtle anomalies, and support multi-class threat categorization, capabilities that signature-based systems inherently lack [3].

Despite recent progress, integrating ML models into operational NIDS environments remains challenging. Much of the existing research treats signature-based and ML-driven approaches as separate solutions, while hybrid systems remain comparatively underexplored [4]. These limitations motivate the central research question addressed in this study: How can probabilistic ML-based detection and rule-based signature detection be systematically fused to improve robustness, interpretability, and multi-class threat coverage in operational network environments?

Prior work [5] introduced a unified evaluation framework for comparing ML classifiers with Snort under consistent conditions. The study proposed a simplified multi-class mapping and reported baseline ML-only and Snort results. However, that study exhibited two notable limitations:

1) Snort's binary output prevented full multi-class comparison, and
2) The models were evaluated independently, without exploring hybrid or fused detection strategies.

To overcome these gaps, the journal version develops a hybrid intrusion-detection architecture that integrates ML predictions with engineered Snort-derived features using a late-fusion stacking strategy. Instead of treating signature-based outputs as ordinary input features within a single model, the extended approach employs decision-level stacked generalization, training a multinomial logistic-regression meta-learner on calibrated ML probability vectors augmented with engineered Snort alert meta-features. This enables the hybrid system to learn when to rely on ML signals versus signature-based evidence, improving the detection of subtle or emerging attacks while maintaining high precision for signature-matched threats.

***Corresponding author:** Maybin K. Muyeba, School of Science, Engineering and Environment, University of Salford, and Data Science and AI (DSAI) Hub, University of Salford, UK. Email: m.k.muyeba@salford.ac.uk

1.1. Extensions beyond the conference paper

This extended study introduces several methodological and experimental advances beyond the conference paper, moving from comparative evaluation to structured hybrid architecture design:

1) A refined hybridization mechanism that incorporates Snort alerts as structured meta-features rather than relying solely on binary outputs.
2) The application of stacked generalization with calibrated probability estimates, enabling more reliable fusion between heterogeneous detection sources.
3) A broadened experimental pipeline incorporating ML-only baselines, feature-level hybridization, stacking-based fusion, and detailed per-class performance analysis.
4) A systematic comparison across ML-only, feature-hybrid, and stacking-hybrid configurations, illustrating the benefits of combining rule-based and learning-driven detection paradigms.
5) An expanded behavioral analysis showing how Snort signatures contribute to differentiating attack types and how stacking mitigates Snort's binary constraints.

Collectively, these enhancements not only address the limitations identified in the conference version but also provide a reproducible and practical path for integrating traditional signature-based systems with contemporary ML-driven detectors. The resulting stacking-hybrid architecture achieves improved macro-F1 and per-class recall relative to standalone ML models and the feature-level hybrid, offering a more robust solution for detecting mixed Normal, Low-Threat, and High-Threat traffic. Importantly, the contribution lies not only in performance improvement but also in demonstrating a reproducible decision-level fusion framework applicable to operational IDS deployments. Overall, the contributions of this extended work are twofold: (i) a more rigorous and capable hybrid IDS architecture and (ii) an expanded empirical basis for understanding the fusion of ML and signature-based methods in multi-class network threat detection.

2. Literature Review

IDSs have undergone substantial evolution over the past two decades, driven by increasing network complexity, encrypted traffic, and the rapid emergence of new attack techniques. Current IDS research is largely centered around two major approaches: signature-based detection and ML-driven anomaly detection. However, in recent years, hybrid frameworks that integrate these paradigms have gained traction as a means of mitigating the individual weaknesses of each method [6]. This section reviews prior work across these domains and positions the present study within this evolving research landscape.

Signature-based IDS such as Snort and Suricata remain widely deployed in industry due to their high precision, deterministic behavior, and strong operational interpretability. Roesch's foundational work on Snort demonstrated how lightweight rule-driven packet inspection could detect known attacks in real time using handcrafted signatures [1]. The effectiveness of signature-based IDS is tightly coupled to the completeness and freshness of these rules, making continuous maintenance a practical necessity [6]. As a result, while signature-based IDS remains reliable for detecting known threats, its ability to generalize across rapidly evolving and previously unseen attack patterns remains limited.

To address these limitations, ML has emerged as a data-driven alternative for intrusion detection. The availability of modern benchmark datasets, most notably UNSW-NB15, which incorporates diverse and contemporary attack behaviors, has enabled systematic evaluation of ML models under realistic network conditions [7]. Several recent studies show that ML-based IDS can adapt to changing network conditions and respond dynamically to emerging threats, addressing the rigidity of static, signature-driven systems [8]. Unlike signature engines, ML models learn discriminative patterns directly from data, enabling generalization to both known and novel attacks.

Ensemble learning approaches, including tree-based and margin-based classifiers, have been widely explored in IDS research and shown to improve anomaly detection compared to purely signature-based methods [3]. More recent studies have extended this area through comparative evaluation of individual ML models and ensemble methods for network intrusion detection, including bagging, boosting, blending, and stacking techniques across multiple datasets and evaluation metrics [9], while subsequent work reported that ensemble techniques surpassed standalone models in detection accuracy across the evaluated datasets and highlighted trade-offs between detection performance and computational efficiency [10]. Earlier work further showed that Extreme Gradient Boosting (XGBoost) can achieve strong intrusion-detection performance on the UNSW-NB15 dataset, while also supporting effective feature selection and dimensionality reduction with only a limited impact on detection accuracy [11]. Building on this, subsequent work shows that XGBoost combined with systematic Bayesian hyperparameter optimization can yield very high detection accuracy and F1 scores, with empirical results indicating suitability for practical deployment [12].

Beyond single-model approaches, recent research has explored more advanced learning strategies to enhance IDS performance further. Complementary multimodal intrusion-detection studies propose combining heterogeneous traffic representations, specifically separate models trained on flow-based features and limited protocol payload bytes, with soft voting decision fusion, achieving high detection accuracy and strong evaluation metrics on UNSW-NB15 Packet Capture (PCAP) data [13]. More recently, AutoML-based approaches have been explored for network intrusion detection, with a stacked ensemble model developed using the MLJAR AutoML framework and combining LightGBM, CatBoost, and XGBoost, outperforming the evaluated individual models on the NSL-KDD dataset [14].

In parallel, other work highlights that handling data imbalance and high dimensionality is critical for reliable IDS performance at scale. More recent work has explored hybrid feature-selection methods for intrusion detection by combining multiple selection techniques to reduce feature dimensions while retaining more relevant features for classification. One such approach, IGRF-RFE, integrates Information Gain, Random Forest (RF), and Recursive Feature Elimination and was shown on the UNSW-NB15 dataset to reduce the feature set from 42 to 23 features while improving anomaly detection accuracy from 82.25% to 84.24% in a multilayer perceptron (MLP)-based IDS [15]. Techniques such as feature-level stacking, oversampling, and dimensionality reduction have subsequently been shown to substantially improve detection accuracy on large, imbalanced benchmark datasets [16]. Recent review literature further highlights ensemble learning as an active area of IDS research and provides a comparative discussion of feature selection, data

balancing, and classifier aggregation techniques used to improve intrusion-detection performance [17].

As the limitations of standalone approaches became more apparent, attention shifted toward hybrid IDS architectures that combine signature-based, anomaly-based, and ML-driven techniques. Such hybrid models leverage the precision of signature-based detection for known attacks while employing ML-based mechanisms to identify previously unseen threats, with experimental results in cloud-based environments indicating improved overall detection effectiveness [18]. At the same time, adversarial ML studies have demonstrated that ML-based IDS can be vulnerable to adversarial perturbations, motivating the development of countermeasures such as Gaussian data augmentation to improve model resilience and harden decision boundaries [19]. More recent work has further explored hybrid learning architectures by combining supervised and unsupervised ML approaches through ensemble stacking models, with a framework integrating an MLP, modified self-organizing map, and Decision Tree (DT), reporting high detection accuracy and low false-positive rates across the NSL-KDD and CIC-DDoS2019 datasets [20].

More recent work has focused specifically on intrusion-detection architectures that integrate ML models with signature-based engines such as Snort. Two-stage pipelines, in which Snort performs initial signature-based filtering and ML classifiers subsequently analyze traffic that does not trigger explicit signatures, have been shown to reduce false positives and improve detection of attacks that partially evade static signatures [21]. Related studies demonstrate that integrating Snort with ML-based anomaly detection can significantly improve detection accuracy and reduce false-positive rates compared to standalone Snort, while maintaining scalability in cloud-oriented network environments [22]. Additional collaborative IDS approaches have also explored embedding ML plugins directly within Snort-based detection pipelines and combining network-based and host-based IDS components, with experimental results indicating reductions in false-positive and false-negative alerts relative to rule-based detection alone, particularly for DoS and probe attacks [23].

Comprehensive evaluations of ML-based IDS across supervised, unsupervised, and deep-learning paradigms identify Snort–ML integration as a promising research direction, highlighting how hybrid systems combine precise signature matching for known threats with adaptive learning for emerging attacks, while still facing challenges related to dataset imbalance, adversarial evasion, and computational overhead [24]. Complementary to these efforts, recent stacking-ensemble-based IDS frameworks demonstrate that integrating multiple heterogeneous classifiers within a stacking architecture can significantly enhance multi-class intrusion-detection performance, particularly when combined with hybrid feature-selection and class-balancing strategies [25]. Independent stacking-based ensemble studies further show that combining multiple classifiers within a stacking framework can outperform individual models in network intrusion detection, achieving high weighted F1 scores on flow-based datasets and providing more robust multi-class detection performance [26].

Building on these observations, the present study develops and evaluates a calibrated decision-level stacking architecture that fuses ML probability outputs with structured Snort-derived metadata, addressing limitations in prior Snort–ML integration studies that primarily rely on pipeline filtering or feature concatenation.

3. Research Methodology

3.1. Dataset description and threat mapping

This study uses the UNSW-NB15 dataset, a contemporary benchmark incorporating modern attack behaviors and realistic network traffic characteristics. The dataset contains network flows generated using IXIA PerfectStorm and labeled into nine attack categories alongside normal traffic [7]. Following the previous ICD-SAIA 2025 study, a simplified three-class mapping is adopted to reflect practical threat levels in enterprise environments:

1) Normal traffic
2) Low-Threat attacks: Generic, Reconnaissance, Analysis, and Fuzzers
3) High-Threat attacks: DoS, Exploits, Shellcode, Worms, and Backdoors

This mapping preserves semantic similarity among attack types while improving interpretability and reducing class imbalance. The dataset is divided into training (70%) and testing (30%) splits using the official distribution to maintain realistic traffic proportions.

Although the UNSW-NB15 dataset exhibits moderate class imbalance, no oversampling methods such as Synthetic Minority Over-sampling Technique (SMOTE) were applied in order to preserve the original distribution of the benchmark. Instead, macro-F1 was adopted as the primary evaluation metric to ensure balanced performance assessment across all three classes.

3.2. Data preprocessing and feature engineering

Preprocessing of the UNSW-NB15 dataset was conducted using Python and Scikit-Learn, ensuring consistency across all experiments. Categorical attributes were first converted into numerical representations suitable for ML models. Protocol identifiers, port values, and other network-level attributes were normalized so that they aligned with the formatting conventions required for matching Snort alerts during hybridization. Records containing malformed or missing values were removed, and all IP addresses, TCP/UDP port values, and protocol fields were standardized to a uniform format.

To maintain strict comparability across all components of the study, the train–test division provided in the official UNSW-NB15 release was retained. This prevented accidental data leakage and ensured that ML models, feature-level hybrid models, and the stacking-based hybrid architecture were evaluated under identical conditions. Using the same fixed split ensures that performance differences arise from model design rather than dataset partitioning.

Following preprocessing, a feature-selection procedure was applied to reduce dimensionality and limit noise in the dataset. An RF classifier was trained on the full feature set to obtain an importance ranking, and the top 18 most informative features were retained for all subsequent modeling stages. These selected attributes served as the core input for the ML-only baselines, hyperparameter tuning, and inference, as well as forming the base layer of features used in the stacking hybrid. This streamlined feature set improved computational efficiency while preserving the dataset's most predictive characteristics.

3.3. Baseline machine learning models

Three ML classifiers were evaluated using the top 18 features:

1) Random Forest (RF)
2) XGBoost (XGB)
3) Decision Tree (DT)

Each model underwent hyperparameter tuning using RandomizedSearchCV with 3-fold cross-validation on the training split. Optimized models were then retrained on the full training set and evaluated on the held-out test set. These ML-only baselines establish performance bounds before hybridization.

3.4. Snort-based signature extraction

Snort 2.9 was deployed using the official ruleset supplemented with custom rules from the original ICDSAIA work. Snort processed the UNSW-NB15 test-set PCAPs and generated alerts containing:

1) Source/destination IP
2) Ports
3) Protocol
4) SID (signature ID)

Alerts were normalized and mapped to test-set rows using a multi-key matching strategy. Alerts generated by Snort were matched against the ground truth using a combination of key packet fields (source and destination IP addresses, port numbers, and protocol identifiers). When multiple alerts corresponded to a single flow, the alerts were aggregated, and the flow was labeled as malicious. Flows without any matching alerts were treated as normal traffic. A binary label (snort_binary) was then assigned to each test record, indicating whether Snort had flagged it as malicious.

For each matched flow, two Snort-derived signals were engineered:

1) snort_alert: binary indicator of any alert
2) snort_sid_count: number of distinct SIDs triggered

These signals constitute the Snort portion of the hybrid models.

3.5. Feature-level hybrid model

In the feature-level hybrid, Snort-derived attributes snort_alert and snort_sid_count were appended to the 18 selected ML features, resulting in a 20-feature-hybrid vector.

Each ML model (RF, XGB, DT) was retrained using these hybrid vectors. This approach tests whether Snort alerts provide useful additional structure at the input level.

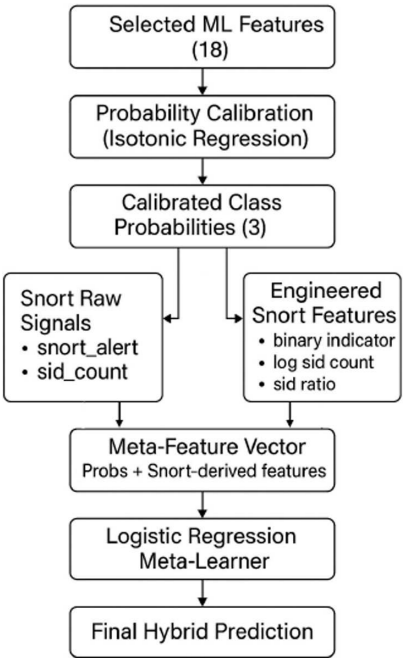
3.6. Stacking-hybrid architecture

The principal contribution of this extended work is the development of a late-fusion stacking architecture that integrates ML predictions with Snort-derived metadata at the decision level. Unlike the feature-level hybrid, which simply augments the input vector, the stacking model operates on calibrated ML probability outputs combined with engineered signature-based features, allowing the system to learn how much weight to assign to each detection source. To ensure reliable confidence estimates, the best-performing ML model, XGBoost, was calibrated using isotonic regression implemented through the CalibratedClassifierCV framework with three-fold cross-validation. Calibration was performed exclusively on the training data, and the resulting calibrated model was subsequently evaluated on the held-out test set to prevent information leakage during probability estimation. These calibrated probabilities serve as a more stable foundation for fusion than raw, uncalibrated outputs.

In parallel, a small set of Snort-derived meta-features was constructed to represent signature-based activity more expressively. Beyond the raw alert signal and SID count, three engineered attributes were created: a binary alert indicator, a logarithmic transformation of the SID frequency, and a normalized SID ratio capturing relative alert intensity. These engineered features provide both scale-invariant and sensitivity-adjusted representations of Snort activity, complementing the smooth probabilistic predictions from XGBoost.

The final meta-learner is implemented as a multinomial Logistic Regression model, chosen for its stability, interpretability, and strong performance when combining heterogeneous signals. Its input consists of the calibrated XGBoost probability vector together with the engineered Snort features, enabling the model to dynamically infer when ML predictions should dominate and when signature evidence should take precedence. Figure 1 summarizes the complete stacking workflow and illustrates how the fusion mechanism leverages the complementary strengths of ML-based generalization and signature-based precision.

Figure 1
Stacking-based hybrid workflow for final multi-class prediction



Although RF and DT classifiers were also evaluated in the ML-only baseline stage, they were not used as standalone base learners in the stacking architecture. Both models produced less reliable probability estimates, with DTs being unstable and RF exhibiting overconfident outputs. Preliminary evaluations further showed that adding RF and DT probability vectors to the meta-learner degraded performance, whereas calibrated XGBoost probabilities alone provided smoother, better-behaved, and more informative meta-features. Since the XGBoost-only configuration

outperformed the expanded variant (XGB + RF + DT), the final stacking hybrid was intentionally built using calibrated XGBoost outputs together with engineered Snort meta-features. This design ensures that the fusion layer is constructed on the strongest and most stable predictive signal available, yielding the highest overall macro-F1 performance in the study.

3.7. Evaluation pipeline

All models were evaluated on the same held-out test set using a consistent set of performance metrics. Accuracy and macro-F1 were used as the primary indicators of overall effectiveness, with macro-F1 being especially important given the imbalance across the Normal, Low-Threat, and High-Threat classes. In addition to these global metrics, the evaluation included per-class precision, recall, and F1 score, as well as confusion matrices to visualize classification behavior and misclassification patterns.

The experimental analysis compared three categories of models: the ML-only baselines, the feature-level hybrid models that appended Snort-derived attributes to the selected feature set, and the final stacking-hybrid model. This multi-stage comparison allowed us to isolate the contribution of each hybridization strategy, examine how Snort information interacts with learned ML representations, and ultimately demonstrate the incremental improvements achieved by the stacking approach relative to both standalone ML models and the simpler hybrid configuration.

3.8. Zero-day generalization considerations

Although the study does not run an explicit zero-day experiment, the stacking method conceptually improves zero-day robustness because:

1) ML can generalize to unseen behaviors
2) Snort fires precise alerts for known patterns
3) The meta-learner learns to rely on ML when Snort is silent and on Snort when ML is uncertain

This complementary behavior provides stronger resilience to emerging threats than either system alone.

3.9. Statistical validation

To evaluate the robustness of the reported results, the full stacking pipeline was repeated across five independent runs using different random seeds. In each run, both the XGBoost baseline and the stacking-hybrid model were retrained and evaluated using the same fixed train–test split. The stacking architecture achieved a mean macro-F1 score of 0.8302 ± 0.0079, compared with 0.8242 ± 0.0038 for the XGBoost baseline. These results indicate that the performance improvement observed for the stacking approach is stable across repeated experiments and is unlikely to be attributable to random variation.

4. Experimental Results

4.1. ML-only baseline results

The ML baselines provide an initial reference point for evaluating the effect of hybridization. As shown in Table 1, all three optimized classifiers—RF, XGBoost, and DT—achieve very high precision on Normal traffic (typically > 0.95), indicating that benign flows are rarely misclassified as attacks. However, their recall for the Normal class is noticeably lower (≈0.74–0.78), revealing a consistent difficulty in recognizing borderline benign traffic that resembles Low-Threat reconnaissance or probing behavior. This systematic weakness across ML models highlights the need for a complementary mechanism, such as Snort's signature checks, to improve benign-traffic discrimination and reduce false positives.

For Low-Threat traffic, the models exhibit relatively stable and strong recall (≈0.84–0.86), indicating that they can easily recognize the simple patterns found in low-level attacks, such as scanning and probing. In contrast, there is a larger performance gap in High-Threat detection as XGBoost achieves the highest recall (>0.93), followed closely by RF, while the DT performs moderately. These results indicate that nonlinear models, particularly XGBoost, are better suited for detecting severe attack patterns such as exploits and shellcode.

Overall, Table 1 shows two key observations that directly motivate hybrid fusion. First, the Normal class remains the hardest to classify, suggesting that Snort's rule-based precision can provide critical complementary information. Second, XGBoost consistently delivers the strongest performance across all classes, making it the most suitable choice as the base learner for the stacking-hybrid architecture developed later in this work.

The confusion matrices in Figure 2 reinforce these patterns. Normal traffic exhibits the highest misclassification rates across all models, with notable spill-over into the Low-Threat class. XGBoost produces the cleanest separation between High-Threat and the other categories, while RF performs similarly but shows

Table 1
ML-only baseline performance

Class	Metric	Random Forest	XGBoost	Decision Tree
0 – Normal	Precision	0.9791	0.9774	0.9579
	Recall	0.7393	0.7488	0.7164
	F1 score	0.8425	0.8479	0.8198
1 – Low Threat	Precision	0.7343	0.7376	0.7159
	Recall	0.8468	0.8695	0.8692
	F1 score	0.7865	0.7982	0.7851
2 – High Threat	Precision	0.7563	0.7719	0.7594
	Recall	0.9708	0.9360	0.9041
	F1 score	0.8502	0.8461	0.8255
Overall	Accuracy	0.8229	0.8284	0.8074
	Macro-F1	0.8264	0.8307	0.8101

Figure 2
Confusion matrices for the ML-only baseline models

Random Forest XGBoost Decision Tree

slightly more confusion between Normal and Low-Threat flows. The DT displays the weakest separation overall. These observations further underscore the need for hybridization, particularly to improve Normal-class recall.

4.2. Feature-level hybrid results

The feature-level hybrid experiment evaluates whether adding Snort-derived indicators directly to the ML input space enhances classification performance. Two Snort-derived features, snort_alert and snort_sid_count, were appended to the 18 selected ML features to form the hybrid input representation. Each ML classifier was then retrained using this expanded feature set and evaluated on the same fixed test split as the ML-only baselines.

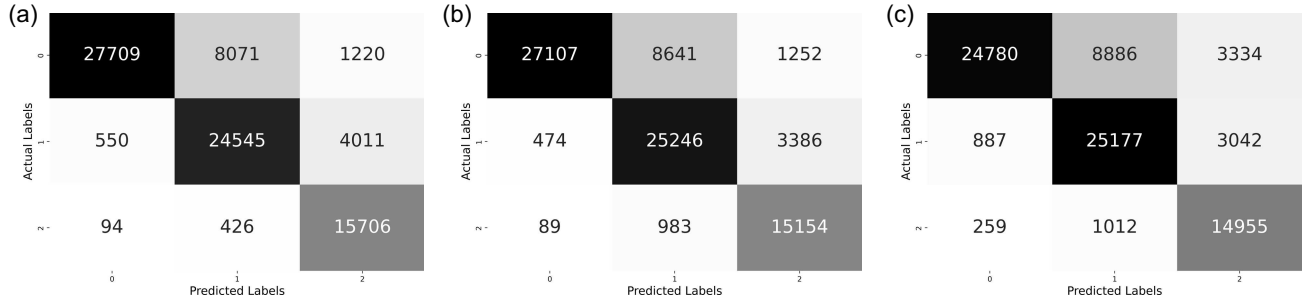
Table 2 summarizes the results. RF experienced a very small improvement, with macro-F1 increasing from 0.826 to 0.828, whereas XGBoost and the DT showed slight reductions. These patterns suggest that Snort's raw binary alerts and signature counts provide limited additional discriminative value when simply merged with statistical flow features. Because Snort signals are sparse and coarse-grained, they do not substantially alter the structure of the ML decision boundary at the feature-input level.

The confusion matrices in Figure 3 visually confirm the limited effect of feature-level hybridization. Across all three models, the main misclassification pattern, Normal traffic being

Table 2
Feature-level hybrid performance

Class	Metric	Random Forest	XGBoost	Decision Tree
0 – Normal	Precision	0.9773	0.9797	0.9558
	Recall	0.7489	0.7326	0.6697
	F1 score	0.8480	0.8383	0.7876
1 – Low Threat	Precision	0.7428	0.7240	0.7178
	Recall	0.8433	0.8674	0.8650
	F1 score	0.7899	0.7892	0.7846
2 – High Threat	Precision	0.7502	0.7657	0.7011
	Recall	0.9680	0.9339	0.9217
	F1 score	0.8452	0.8415	0.7964
Overall	Accuracy	0.8254	0.8199	0.7884
	Macro-F1	0.8277	0.8230	0.7895

Figure 3
Confusion matrices for the feature-level hybrid models



Random Forest XGBoost Decision Tree

confused with Low-Threat traffic, remains largely unchanged from the ML-only baselines. RF shows slightly improved balance, while XGBoost and the DT exhibit almost identical class-overlap to their baseline versions. High-Threat detection also shows no meaningful refinement. These results reinforce the conclusion that simply adding Snort alerts to the input features does not materially reshape decision boundaries, motivating the more sophisticated decision-level fusion strategy introduced in Section 4.3.

4.3. Stacking-hybrid results (calibrated XGB + Snort)

The stacking-based hybrid model produced the strongest overall performance among all configurations evaluated in this study. As shown in Table 3, the stacking hybrid achieved an overall accuracy of 0.8346 and macro-F1 of 0.8347, outperforming both the ML-only baselines and the feature-level hybrid. The largest improvements appear in the Normal and Low-Threat classes, where the decision-level fusion enables the model to better distinguish benign flows from low-risk reconnaissance-like activity. High-Threat detection remains strong and comparable to the XGBoost baseline, demonstrating that Snort's signature signals reinforce, but do not distort, the ML decision boundary.

Table 3
Stacking-hybrid model performance

Class	Precision	Recall	F1 score
0 – Normal	0.9596	0.7801	0.8608
1 – Low Threat	0.7554	0.8580	0.8035
2 – High Threat	0.7751	0.9169	0.8401
Overall	Accuracy	0.8346	
	Macro-F1	0.8347	

Figure 4 provides visual confirmation of these improvements. Normal traffic shows noticeably fewer misclassifications into the Low-Threat category compared with the earlier models, indicating that the meta-learner has learned to correctly discount weak Snort activity while strengthening ML confidence estimates. Low-Threat attacks are also identified more reliably, with reduced confusion against the Normal class. High-Threat predictions remain sharply defined, reflecting that Snort's engineered meta-features enhance, but do not interfere with XGBoost's modeling of severe attacks. Overall, the confusion matrix supports the quantitative gains shown in Table 3, demonstrating that decision-level fusion produces clearer class boundaries and more stable behavior across all threat levels.

Figure 4
Confusion matrix for the stacking-hybrid model

These results confirm that late-fusion stacking is substantially more effective than simple feature concatenation. By learning when to rely on ML probability estimates and when to prioritize Snort-derived signals, the stacking architecture yields a more balanced and robust classifier, forming the final and best-performing model in this study.

4.4. Comparative summary across all models

Across all modeling stages, ML-only baselines, feature-level hybrid models, and the final stacking architecture, a clear progression in performance emerges. The ML-only models establish a strong baseline, with XGBoost performing best (accuracy 0.828; macro-F1 0.831), followed closely by RF and then the DT. However, all three models show a recurring weakness in Normal-class recall, indicating difficulty distinguishing benign traffic from Low-Threat reconnaissance activity.

The feature-level hybrid models, which append Snort-derived attributes directly to the input feature set, produce only marginal changes in performance. RF shows a small improvement in macro-F1 (from 0.826 to 0.828), while XGBoost and the DT experience slight decreases. These results suggest that simple concatenation of Snort alerts provides limited additional discriminative value.

The stacking-hybrid model provides the largest and most consistent improvement. By fusing calibrated XGBoost probabilities with engineered Snort indicators at the decision level, the meta-learner improves both Normal and Low-Threat detection while retaining strong High-Threat performance. This approach achieves the highest overall scores (accuracy 0.8346; macro-F1 0.8347), outperforming every ML-only and feature-hybrid variant.

Overall, the comparative analysis shows that while ML-only models perform well and feature-level hybrids add little benefit, the stacking approach delivers a more balanced and robust classifier, validating it as the final model adopted in this study.

5. Discussion

The experimental results demonstrate that ML classifiers retain strong predictive ability on modern network traffic, particularly for high-severity attacks, but they struggle to distinguish Normal traffic from Low-Threat behaviors. This challenge was evident in the ML-only baselines, where all three models achieved high precision but weaker recall for the Normal class. This behavior reflects the statistical similarity between benign flows and mild reconnaissance or probing activity, a pattern well documented in intrusion-detection literature. The limitations of the ML-only models, therefore, motivated the integration of Snort, which excels at filtering benign traffic due to its signature-driven precision.

Introducing Snort-derived features at the input level (feature-level hybrid) resulted in only marginal changes in overall performance. While RF showed a small improvement, XGBoost and the DT experienced slight decreases in macro-F1, indicating that direct concatenation of Snort indicators does little to reshape ML decision boundaries. This is expected because Snort alerts are sparse binary signals that contribute limited additional information when combined with high-dimensional statistical features.

The feature-level hybrid experiment, therefore, shows that simple integration of Snort indicators does not produce meaningful improvement.

In contrast, the stacking-based hybrid model produced the strongest results by integrating Snort signals at the decision level rather than at the feature level. Calibrating the XGBoost base learner using isotonic regression transformed its probability outputs into smooth, well-behaved confidence estimates. When combined with engineered Snort meta-features, these calibrated probabilities provided the meta-learner with more informative signals than raw feature concatenation. The resulting late-fusion model achieved the highest accuracy (0.8346) and macro-F1 (0.8347), outperforming all ML-only and feature-hybrid models. This improvement was especially noticeable in the Normal and Low-Threat classes, where the fusion of statistical and signature-based cues compensated for the weaknesses of each detection approach.

The results also reveal important insights about the relationship between ML-based generalization and signature-based precision. XGBoost, already the strongest ML-only model, continued to produce the best probability estimates after calibration, which explains why it functioned effectively as the sole base learner in the final stacker. Preliminary tests combining XGBoost with RF and DT probabilities did not improve performance; instead, the additional probability vectors introduced instability and noise into the meta-learner. This outcome reinforces a common principle in stacking: additional base models do not necessarily improve performance, particularly when weaker classifiers contribute unreliable probabilities. Choosing an XGBoost-only stacker was appropriate because it performed best in our experiments, and this approach is consistent with common principles for designing effective ensemble models.

Finally, although the study does not explicitly evaluate zero-day attacks, the behavior of the stacking model suggests that it would offer stronger resilience to unseen threats than either ML or Snort alone. When Snort fails to trigger an alert, as would be the case for novel attacks, the meta-learner relies on the calibrated ML probabilities. Conversely, when ML is uncertain, but Snort recognizes a known pattern, the signature signal provides decisive evidence. This complementary interaction between generalization and signature matching is central to the robustness of the proposed hybrid architecture.

Overall, the discussion highlights that the contribution of this extended study is not merely an incremental performance increase but a clearer understanding of how rule-based and learning-based detection systems can be effectively combined. The results highlight the effectiveness of late-fusion stacking and suggest a clear path for advancing hybrid intrusion-detection research.

6. Conclusion

This study investigated how ML and signature-based intrusion-detection techniques can be combined to improve multi-class network threat detection on the UNSW-NB15 dataset. Building upon prior ICDSAIA 2025 work, which evaluated ML models and Snort independently, the present study introduced two hybridization strategies, feature-level concatenation and a decision-level stacking architecture. While the feature-level hybrid produced only marginal gains, the stacking model demonstrated clear and consistent improvements by integrating calibrated XGBoost probability estimates with engineered Snort-derived features.

The final stacking hybrid achieved the strongest performance overall, with an accuracy of 0.8346 and a macro-F1 of 0.8347, outperforming both the ML-only baselines and the feature-hybrid configuration. Improvements were particularly notable in the Normal and Low-Threat classes, where Snort's rule-based precision helped mitigate the tendency of ML models to misclassify borderline benign traffic. At the same time, the model preserved the high recall of XGBoost on High-Threat attacks, confirming that decision-level fusion can leverage complementary strengths without suppressing ML generalization.

The findings show that carefully designed late-fusion strategies offer a practical and effective pathway for integrating traditional signature-based systems with modern ML detectors. This is particularly important for modern network environments where both known and emerging threats coexist. Rather than treating ML and signature detection as competing approaches, the proposed architecture demonstrates that they can operate collaboratively to provide more balanced and resilient intrusion detection.

Future work can extend this study in several directions. One direction is to evaluate the stacking model under simulated zero-day conditions by withholding entire attack categories during training. Another direction involves integrating temporal features or flow-sequence models like Long Short-Term Memory (LSTM) or Transformer-based architectures to capture long-term traffic dependencies that static classifiers may miss. Finally, deploying the hybrid system in a semi-realistic online setting would provide additional insight into latency, throughput, and operational effectiveness in live network environments.

Overall, this work demonstrates how hybrid ML–signature fusion can enhance multi-class intrusion detection while maintaining practical relevance for modern network security systems.

Recommendations

The results of this study suggest several recommendations for researchers and practitioners combining ML and signature-based intrusion-detection approaches. First, decision-level fusion should be prioritized over simple feature concatenation, as the stacking architecture delivered more stable and accurate predictions. In particular, calibrated probability outputs should be used when integrating ML models with external detection signals to ensure reliable meta-learning behavior. The findings also indicate that Snort-derived metadata is most valuable for resolving ambiguity in borderline benign and low-intensity attack traffic, meaning that signature-based cues are best applied selectively to complement, rather than replace, ML decision-making. For real-world deployments, hybrid systems should be evaluated not only on offline datasets but also under streaming or near-real-time conditions to assess latency, throughput, and operational feasibility. Finally, future research should investigate the behavior of hybrid systems under zero-day or withheld attack scenarios, where decision-level fusion may provide additional resilience against previously unseen threats.

Acknowledgment

This journal manuscript is an extended version of our earlier conference paper presented at ICDSAIA 2025, titled "Comparative Evaluation of Machine Learning and Signature-Based NIDS for Multi-Class and Binary Threat Detection." The conference version has been substantially expanded in methodology,

experimental design, results, and discussion to meet the journal's originality requirements.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data availability statement

The data that support the findings of this study are openly available in UNSW-NB15 at https://research.unsw.edu.au/projects/unsw-nb15-dataset.

Author Contribution Statement

Somayina C. Wen-Udeoji: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Project administration. **Maybin K. Muyeba:** Conceptualization, Methodology, Validation, Investigation, Writing – review & editing, Supervision. **Azadeh Mohammadi:** Conceptualization, Validation, Writing – review & editing.

References

[1] Roesch, M. (1999). Snort: Lightweight intrusion detection for networks. In *Proceedings of LISA '99: 13th Systems Administration Conference*, *99*(1), 229–238.

[2] Utimura, L., Costa, K., & Scherer, R. (2022). Real-time application of OPF-based classifier in Snort IDS. In A. X. Falcão & J. P. Papa (Eds.), *Optimum-path forest: Theory, algorithms, and applications* (pp. 55–93). Academic Press. https://doi.org/10.1016/B978-0-12-822688-9.00011-6

[3] Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

[4] Maseno, E. M., Wang, Z., & Xing, H. (2022). A systematic review on hybrid intrusion detection system. *Security and Communication Networks, 2022*(1), 9663052. https://doi.org/10.1155/2022/9663052

[5] Wen-Udeoji, S. C., Muyeba, M. K., & Mohammadi, A. (2025). Comparative evaluation of machine learning and signature-based NIDS for multi-class and binary threat detection. In *Data Science, AI and Applications: First International Conference*, 238–251. https://doi.org/10.1007/978-3-032-11352-8_17

[6] Khraisat, A., Gondal, I., Vamplew, P., & Kamruzzaman, J. (2019). Survey of intrusion detection systems: Techniques, datasets and challenges. *Cybersecurity*, *2*(1), 20. https://doi.org/10.1186/s42400-019-0038-7

[7] 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*, 1–6. https://doi.org/10.1109/MilCIS.2015.7348942

[8] Benmalek, M., & Haouam, K.-D. (2024). Advancing network intrusion detection systems with machine learning techniques. *Advances in Artificial Intelligence and Machine Learning*, *4*(3), 2575–2592. https://dx.doi.org/10.54364/AAIML.2024.43150

[9] Bibers, I., Arreche, O., & Abdallah, M. (2024). *A comprehensive comparative study of individual ML models and ensemble strategies for network intrusion detection systems*. arXiv. https://doi.org/10.48550/arXiv.2410.15597

[10] Bibers, I., Arreche, O., Alayed, W., & Abdallah, M. (2025). Ensemble-IDS: An ensemble learning framework for enhancing AI-based network intrusion detection tasks. *Applied Sciences*, *15*(19), 10579. https://doi.org/10.3390/app151910579

[11] Husain, A., Salem, A., Jim, C., & Dimitoglou, G. (2019). Development of an efficient network intrusion detection model using extreme gradient boosting (XGBoost) on the UNSW-NB15 dataset. In *2019 IEEE International Symposium on Signal Processing and Information Technology*, 1–7. https://doi.org/10.1109/ISSPIT47144.2019.9001867

[12] Yulianton, H., Sutanto, F. A., & Noor Santi, R. C. (2025). Optimized network intrusion detection using XGBoost with hyperparameter tuning: An empirical study on UNSW-NB15 dataset. *Journal of Software Engineering and Simulation*, *11*(8), 1–7. https://doi.org/10.35629/3795-11080107

[13] Kiflay, A., Tsokanos, A., Fazlali, M., & Kirner, R. (2024). Network intrusion detection leveraging multimodal features. *Array*, *22*, 100349. https://doi.org/10.1016/j.array.2024.100349

[14] Gyimah, N. K., Akinie, R., Mwakalonge, J., Izison, B., Mukwaya, A., Ruganuza, D., & Sulle, M. (2025). An AutoML-based approach for network intrusion detection. In *SoutheastCon 2025*, 1177–1183. https://doi.org/10.1109/SoutheastCon56624.2025.10971461

[15] Yin, Y., Jang-Jaccard, J., Xu, W., Singh, A., Zhu, J., Sabrina, F., & Kwak, J. (2023). IGRF-RFE: A hybrid feature selection method for MLP-based network intrusion detection on UNSW-NB15 dataset. *Journal of Big Data*, *10*(1), 15. https://doi.org/10.1186/s40537-023-00694-8

[16] Talukder, M. A., Islam, M. M., Uddin, M. A., Hasan, K. F., Sharmin, S., Alyami, S. A., & Moni, M. A. (2024). Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction. *Journal of Big Data*, *11*(1), 33. https://doi.org/10.1186/s40537-024-00886-w

[17] Salama, M. A., Tawfeek, R. M., Hamdy, S., & Salim, O. M. (2025). Advances in ensemble machine learning for network intrusion detection systems: A comprehensive review. *Benha Journal of Engineering Science and Technology*, *2*(1), 117–125. https://doi.org/10.21608/bjest.2025.445861

[18] Vashishtha, L. K., Singh, A. P., & Chatterjee, K. (2023). HIDM: A hybrid intrusion detection model for cloud based systems. *Wireless Personal Communications*, *128*(4), 2637–2666. https://doi.org/10.1007/s11277-022-10063-y

[19] Jmila, H., & Khedher, M. I. (2022). Adversarial machine learning for network intrusion detection: A comparative study. *Computer Networks*, *214*, 109073. https://doi.org/10.1016/j.comnet.2022.109073

[20] Mills, G. A., Acquah, D. K., & Sowah, R. A. (2024). Network intrusion detection and prevention system using hybrid machine learning with supervised ensemble stacking model. *Journal of Computer Networks and Communications*, *2024*(1), 5775671. https://doi.org/10.1155/2024/5775671

[21] Preethi, T., Reddy, P. R., Likhitha, L., Kumar, P. P., & Kamani, A. (2024). A novel approach for anomaly

detection using Snort integrated with machine learning. In *2024 11th International Conference on Computing for Sustainable Global Development*, 796–801. https://doi.org/10.23919/INDIACom61295.2024.10498401

[22] Viharika, S., & Balaji, N. (2025). Enhancing intrusion detection and cloud security by integrating Snort with advanced AI techniques for improved accuracy and threat mitigation. *Journal of Information Systems Engineering and Management*, *10*, 627–637. https://doi.org/10.52783/jisem.v10i24s.3953

[23] Priambodo, D. F., Faizi, A. H. N., Rahmawati, F. D., Sunaringtyas, S. U., Sidabutar, J., & Yulita, T. (2024). Collaborative intrusion detection system with Snort machine learning plugin. *JOIV: International Journal on Informatics Visualization*, *8*(3), 1230–1238. https://dx.doi.org/10.62527/joiv.8.3.2018

[24] Alfaris, S. A., Al-Ameen, H. A. A., Abduljabbar, Z. A., & Nyangaresi, V. O. (2026). Adapting intrusion detection systems using machine learning: Techniques, performance analysis and hybridisation with Snort. In *Software Engineering: Emerging Trends and Practices in System Development: Proceedings of 14th Computer Science On-line Conference 2025*, *5*, 347–366. https://doi.org/10.1007/978-3-032-04725-0_22

[25] Badiger, V. S., & Shyam, G. K. (2025). A multiclass network intrusion detection system using stacking ensemble technique with hybrid feature selection. *Journal of Advances in Information Technology*, *16*(3), 342–356. https://doi.org/10.12720/jait.16.3.342-356

[26] Ali, M., Haque, M. U., Durad, M. H., Usman, A., Mohsin, S. M., Mujlid, H., & Maple, C. (2023). Effective network intrusion detection using stacking-based ensemble approach. *International Journal of Information Security*, *22*(6), 1781–1798. https://doi.org/10.1007/s10207-023-00718-7

How to Cite: Wen-Udeoji, S. C., Muyeba, M. K., & Mohammadi, A. (2026). Late-Fusion Stacking of Machine Learning and Snort Signatures for Multi-Class Intrusion Detection on UNSW-NB15. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62028592