A robust semi-supervised hybrid ensemble for Network Intrusion Detection System (NIDS)

Sofy Fitriani[1], and *Siti Dwi* Setiarini[1*]

[1]Department of Informatics Engineering, Politeknik Negeri Bandung, Bandung, Indonesia

Abstract. The increasing sophistication of cyber threats demands the development of robust Intrusion Detection Systems (IDS). However, creating an effective anomaly-based IDS is hindered by critical challenges, including data labeling scarcity, inherent class imbalance, and a severe vulnerability to adversarial attacks. This research proposes a novel semi-supervised hybrid ensemble architecture designed to holistically address these issues. The methodology leverages a rigorous 5-Fold Cross-Validation framework on the NSL-KDD dataset. Our proposed model uniquely integrates three distinct learning paradigms: a Self-Training classifier, an unsupervised Autoencoder, and an Isolation Forest, aggregating their scores via an F1-score-optimized threshold. This design inherently handles class imbalance without explicit data resampling. The primary contribution of this work is a two-fold evaluation. First, the framework's baseline performance is established, achieving a state-of-the-art average F1-score of 99.24%, a detection recall of 99.12%, and a remarkably low False Positive Rate of 0.58%. Second, its robustness is quantified by subjecting the model to white-box adversarial attacks. While performance was impacted, the model demonstrated significant resilience, maintaining a high detection recall of 88.11%. This validates that the synergistic combination of diverse learning models not only provides a highly effective and balanced solution for standard detection but also offers a quantifiable degree of inherent robustness against adversarial threats, a crucial attribute for real-world deployment.

1 Introduction

The modern digital world faces increasing cyber threats, with the average cost of a data breach rising to $4.45 million in 2023 and ransomware attacks affecting all sectors [1, 2]. In this environment, Network Intrusion Detection Systems (NIDS) have become essential [3]. Traditional signature-based IDSs can detect known threats but fail against zero-day attacks, driving the adoption of anomaly-based ML systems that detect deviations from normal behavior [4, 5].

Implementing anomaly-based IDSs presents several challenges. First, labeling large volumes of network data is costly, limiting fully supervised approaches and motivating semi-supervised methods that leverage a small labeled subset and a larger unlabeled set to learn

* Corresponding author: siti.dwi@polban.ac.id

more effective decision boundaries [6–8]. Second, severe class imbalance—rare malicious events versus abundant normal traffic—affects training. Techniques like SMOTE help but may introduce unrealistic samples that reduce real-world performance [9–11]. Third, ML models are vulnerable to adversarial attacks; subtle input manipulations can bypass detection, highlighting that high accuracy on imbalanced data does not guarantee reliable real-world performance [12, 13].

These overlapping challenges—limited labels, imbalanced classes, and adversarial vulnerability—remain largely addressed separately in current research, which often assumes fully supervised scenarios or overlooks practical constraints [14, 15].

To address this gap, this paper proposes a hybrid ensemble framework combining a Self-Training Classifier, an unsupervised Isolation Forest, and a self-supervised Autoencoder. We first assess its handling of class imbalance, with and without SMOTE, and then evaluate robustness against white-box adversarial attacks. Using the NSL-KDD benchmark dataset, this two-step evaluation examines both design effectiveness and operational resilience, targeting reliable defenses against costly cyber threats.

2 Research methodology

This study proposes a novel semi-supervised hybrid ensemble architecture for anomaly-based intrusion detection. The research process is systematically structured into data preparation, model implementation within a cross-validation framework, and a two-stage performance evaluation, as illustrated in Fig 1.

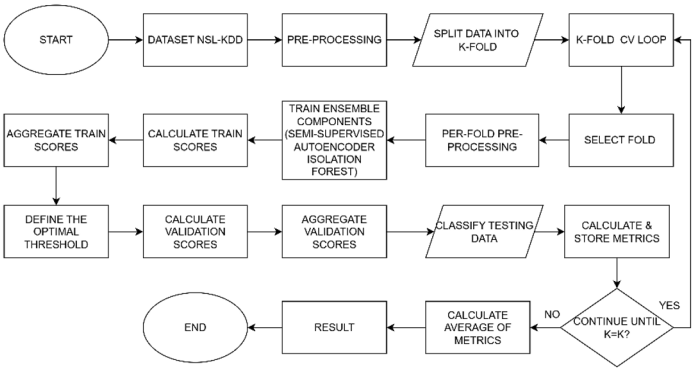

Fig. 1 Flowchart of the proposed hybrid ensemble methodology.

2.1 Dataset description

The problem in this research is how the proposed method can detect anomalies in the network, especially due to a high FPR [16]. There are many factors that cause this condition, one of which is dataset complexity. From this dataset, a map of the problems that occur can be seen, especially in solving anomaly detection using the Network Security Layer-Knowledge Discovery in Database (NSL-KDD) dataset. This dataset is the updated version of KDD Cup'99, which aims to overcome the problems of data redundancy and bias in classification models. Typically, the components in NSL-KDD are used to evaluate the performance of various intrusion detection algorithms and methods in a network. The dataset contains a diverse range of intrusions categorized into four main attack types, such as Denial of Service (DoS), Probe, User to Root (U2R), and Remote to Local (R2L), along with normal traffic records as shown in Table 1.

Table 1. Distribution of attacks in NSL-KDD.

Attack Category	Number of Attack Type	Total Sample
DoS	10	45,927
Probe	4	11,656
R2L	17	995
U2R	8	52
Normal	1	67,343
Total	**40 (39 Attack + 1 Normal)**	**125,973**

The imbalance across categories is evident: DoS and Probe attacks dominate the dataset, while R2L and U2R are highly underrepresented. This characteristic significantly impacts model performance, particularly in the detection of minority attack types such as U2R.

2.2 Data preprocessing

The pre-processing pipeline consists of several key steps applied consistently within each fold of the cross-validation: (1) Each record is assigned a binary label, where 'normal' connections are mapped to class 0 and all distinct attack types are mapped to class 1 (anomaly), (2) Categorical features such as protocol_type, service, and flag are transformed into a numerical format using One-Hot Encoding and (3) Feature Scaling: All numerical features are standardized using StandardScaler, which scales the data to have a mean of 0 and a standard deviation of 1. This step is critical to prevent features with large numerical ranges from disproportionately influencing the models. The scaler is fit exclusively on the training data of each fold to prevent data leakage.

2.3 K-Fold cross-validation framework

Beyond the 5-Fold Cross-Validation process on the training set, the system was also evaluated using the unseen NSL-KDD test set (KDDTest+), which contains instances not included in training. This additional evaluation ensures that the proposed hybrid approach is not overfitting to the training distribution and can generalize to previously unseen data.

To avoid the bias of a single train-test split and to provide a stable estimate of model performance, a 5-Fold Stratified Cross-Validation methodology was employed. The complete dataset was partitioned into five equally sized folds, maintaining the same percentage of samples for each class as in the complete set. In each of the five iterations, one fold is reserved as the validation set, while the remaining four folds are used for training the proposed ensemble.

2.4 The proposed semi-supervised hybrid ensemble framework

To reflect real-world scenarios where labeled attack data is limited, we designed a semi-supervised approach. In this phase, only 10% of the training samples were treated as labeled data, while the remaining 90% were considered unlabeled. This design explicitly models the data scarcity challenge, a key limitation in real-world security systems.

The core of our proposed method is a tri-paradigm hybrid ensemble, engineered to overcome the limitations of single-model systems. This architecture is designed to provide comprehensive threat coverage by synergizing the strengths of three fundamentally different learning approaches. First, a Self-Training classifier with a LightGBM base is employed to address data scarcity; this semi-supervised component leverages the small labeled set to iteratively generate pseudo-labels for the large unlabeled pool. Concurrently, a reconstruction-based Autoencoder is trained to build a robust model of normality, enabling

the detection of complex contextual anomalies. Finally, an Isolation Forest is integrated to efficiently identify statistically rare threats or outliers by partitioning the data. The inclusion of these three components ensures the framework is sensitive to a wide spectrum of threats.

2.5 Ensemble decision and threshold optimization

The system combines anomaly scores from the Self-Training classifier, Autoencoder, and Isolation Forest. Scores are first normalized to [0, 1] and then averaged into a composite score. A dynamic threshold, optimized per training fold to maximize F1-score, is used for classification. Averaging leverages algorithmic diversity: it balances the strengths and weaknesses of individual models—such as Isolation Forest's high recall versus Self-Training's low FPR—producing a more robust final decision. Normalization ensures each model contributes equally, preventing any single model from dominating the result.

2.6 Experimental design for robustness evaluation

To evaluate the resilience of our proposed architecture, we tested its robustness against white-box adversarial attacks, which aim to misclassify malicious inputs as 'normal.' Adversarial samples were generated using an untargeted white-box method from the NSL-KDD test set, ensuring the same inputs were used across the hybrid ensemble and its three components (Self-Training, Autoencoder, Isolation Forest). This setup simulates a worst-case scenario where the attacker has full model knowledge. Robustness was measured by the F1-score degradation between clean and adversarial data, with lower degradation indicating better resilience.

2.7 Evaluation metrics

The performance of the proposed architecture is evaluated using a suite of metrics suitable for imbalanced classification tasks: F1-Score, Recall (Detection Rate) for the anomaly class, and FPR. The final reported performance is the mean and standard deviation of these metrics across all five folds. To further validate the model's generalizability, we also report its performance on the truly unseen KDDTest+ dataset, which was not used in any training or validation phase.

3 Result and discussion

This section presents the empirical results of our experiments. The performance of the proposed hybrid ensemble framework is benchmarked against its three individual components: an Autoencoder-only model, an Isolation Forest-only model, and a Semi-Supervised-only model. The evaluation is conducted in two phases: first on the clean validation set to establish a baseline, and second on the adversarially perturbed set to assess robustness.

3.1 Baseline performance on clean data

The initial evaluation on the clean, non-adversarial data serves to validate the fundamental effectiveness of each approach. The comprehensive results are summarized in Table 2, with a visual comparison provided in Fig 2.

Table 2. Evaluation metrics.

Method	F1-Score	Recall	FPR
Autoencoder	88.41%	86.41%	8.4%
Isolation Forest	78.32%	99.78%	51.02%
Self-Training	99.15%	98.96%	0.61%
Proposed Hybrid Ensemble	**99.24%**	**99.12%**	**0.58%**

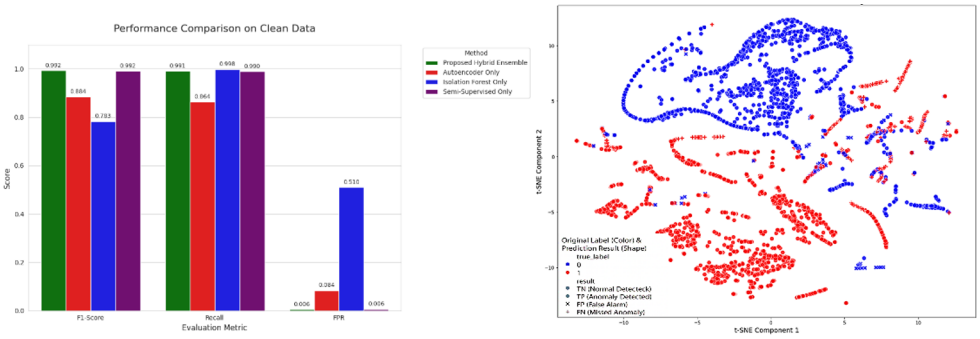

Fig 2. Performance comparison all model on clean data and visualization the classification results of proposed method.

As shown in Table 2 and Fig. 2, the baseline results highlight the strengths and weaknesses of each model. The Proposed Hybrid Ensemble achieved the best balance, with an F1-Score of 99.25%, high Recall of 99.12%, and very low FPR of 0.57%, showing that combining diverse models maximizes detection while minimizing false alarms. The Semi-Supervised model also performed strongly, with 98.96% Recall and only 0.61% FPR, confirming the effectiveness of self-training with limited labeled data. In contrast, unsupervised methods showed trade-offs: the Autoencoder reached 86.41% Recall but a higher FPR of 8.40%, while the Isolation Forest achieved 99.78% Recall but an extremely high FPR of 51.02%. Overall, these results confirm that although individual models have specific strengths, the hybrid ensemble provides the most optimal and balanced performance.

3.2 Robustness analysis under adversarial attack

The core of this research was to evaluate how each model withstands a targeted adversarial attack designed to evade detection by camouflaging malicious samples to appear more like normal data. The results, visualized in Fig 3, revealed a striking divergence in resilience among the individual components.

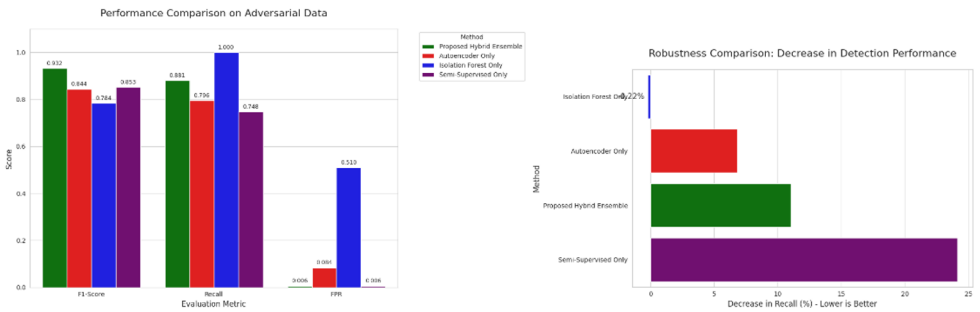

Fig 3. Robustness comparison: Decrease in detection rate under adversarial attack.

The Semi-Supervised model, despite its high accuracy on clean data, proved to be the most fragile, with its Detection Rate plummeting from 98.96% to 74.84% (a ~24% decrease), critically demonstrating the attack's high transferability to tree-based classifiers. The Autoencoder, as the direct target, also confirmed its vulnerability with a significant ~7% drop in detection. In a remarkable contrast, the Isolation Forest exhibited complete immunity to this attack, with its detection rate remaining at 100%, suggesting that perturbations designed for reconstruction-based models may paradoxically enhance the isolability of anomalies. This is where the synergistic benefit of the Proposed Hybrid Ensemble was proven. It successfully mitigated these extreme outcomes, with its Detection Rate dropping by only ~11% (from 99.12% to 88.11%). This degradation is less than half that of its most vulnerable component, demonstrating a superior and more balanced robustness that no single model could achieve alone.

3.3 Generalization on unseen data (KDDTest+)

Generalization on unseen data, specifically on the KDDTest+ dataset, is a critical evaluation to measure how well a model can apply its knowledge to data it has never seen before. With an accuracy level of 75%, your model demonstrates a decent capability in distinguishing between normal and anomalous network connections in a realistic environment. Although this figure indicates the model successfully classifies three out of every four data points correctly, there is still room for improvement. A 75% accuracy is a solid achievement, but it underscores the importance of further efforts to refine the model to reduce the error rate and enhance the reliability of anomaly detection on dynamic data.

3.4 Discussion: The power of algorithmic diversity

The evaluation on KDDTest+ confirms the strength of our proposed hybrid ensemble. Individual models showed varied vulnerabilities to adversarial attacks: the Semi-Supervised model experienced the highest F1-score degradation, the Autoencoder showed moderate degradation, and the Isolation Forest, while maintaining high recall, proved more resilient. By combining these models, the ensemble reduced overall F1-score degradation to less than half that of its strongest single component, demonstrating superior robustness. The Isolation Forest acted as a “safety net,” compensating for the weaknesses of the other models. These results confirm that integrating models with fundamentally different paradigms—learned decision boundaries, reconstruction error, and isolability—creates a more complex and resilient decision landscape. Consequently, attacks that easily bypass individual models are much less effective against the ensemble, resulting in a more reliable and robust security solution on truly unseen data.

4 Conclusion

This research successfully addressed the critical challenges of data scarcity, class imbalance, and adversarial vulnerability in Network Intrusion Detection Systems. We proposed a novel semi-supervised hybrid ensemble combining a Self-Training Classifier, an Isolation Forest, and an Autoencoder, and rigorously evaluated it on both clean and adversarial data, including the KDDTest+ set. Experimental results showed that the ensemble achieves state-of-the-art performance on clean data, handling class imbalance effectively even without artificial over-sampling. More importantly, the robustness evaluation on KDDTest+ demonstrated that, while individual components suffered severe performance degradation under adversarial attacks, the ensemble’s F1-score decreased substantially less, highlighting its superior

resilience. Overall, the key achievement is that algorithmic diversity enables a multi-paradigm system where the strengths of one model compensate for the weaknesses of others, resulting in a more reliable and trustworthy IDS capable of handling real-world, hostile environments. Future work could further enhance robustness through techniques such as adversarial training or dynamic model weighting.

References

1. IBM Security, Cost of a Data Breach Report 2023, (IBM Corporation, Armonk, NY, 2023).
2. Verizon, 2023 Data Breach Investigations Report (DBIR), (Verizon, Basking Ridge, NJ, 2023).
3. A. Thakkar, R. Lohiya, A Review of the Deep Learning and Machine Learning Methods for Cyber Security. Arch. Comput. Methods Eng. **29**, 2379–2403 (2022). https://doi.org/10.1007/s11831-021-09641-x
4. S. H. H. Madni, M. S. A. Latiff, S. M. M. S. H. Coulibaly, Y., A survey on threats, attacks and counter-measures on DNS. Int. J. Comput. Sci. Inf. Secur. **14**, 906 (2016).
5. A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, A. Alazab, A survey of intrusion detection systems: techniques, datasets and challenges. Cybersecurity **2**, 1 (2019). https://doi.org/10.1186/s42400-019-0038-7
6. I. Sharafaldin, A.H. Lashkari, A.A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), pp. 108-116 (2018). https://doi.org/10.5220/0006639801080116
7. A.S. Sikeridis, E.P. Karyda, L.G. Papageorgiou, A semi-supervised approach for network intrusion detection. Inf. Sci. **546**, 392-407 (2021). https://doi.org/10.1016/j.ins.2020.08.067
8. J.E. Van Engelen, H.H. Hoos, A survey on semi-supervised learning. Machine Learning **109**, 373–440 (2020). https://doi.org/10.1007/s10994-019-05855-6
9. M. A. Ambusaidi, X. He, P. Nanda, Z. Tan, Building an intrusion detection system using a filter-based feature selection algorithm. IEEE Trans. Comput. **65**, 2986-2998 (2016). https://doi.org/10.1109/TC.2016.2549132
10. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique. J. Artif. Intell. Res. **16**, 321-357 (2002). https://doi.org/10.1613/jair.953
11. N.V. Kovács, A. Kertész, The forgotten problem of high-dimensional imbalanced data sets. Inf. Sci. **611**, 456-479 (2022). https://doi.org/10.1016/j.ins.2022.08.069
12. I. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014). https://arxiv.org/abs/1412.6572
13. D. B. Rawat, C. C. Sankey, C. A. Kamhoua, K. K. R. Bentum, K. P. Dahal, Characterizing Adversarial Attacks Against Deep Learning Models for Network Intrusion detection. IEEE Access **9**, 34555-34567 (2021). https://doi.org/10.1109/ACCESS.2021.3061690
14. S. M. Kasongo, Y. Sun, A deep learning method with filter-based feature engineering for wireless intrusion detection system. IEEE Access **8**, 38561-38576 (2020). https://doi.org/10.1109/ACCESS.2020.2975635

15. Y. Xin, L. Kong, Z. Liu, Y. Chen, Y. Li, H. Zhu, M. Zhang, A. V. Vasilakos, Machine learning and deep learning methods for cybersecurity. IEEE Access **6**, 35365-35381 (2018). https://doi.org/10.1109/ACCESS.2018.2836950
16. S. Fitriani, S. Mandala, M. A. Murti, Review of semi-supervised method for intrusion detection system. 2016 Asia Pacific Conference on Multimedia and Broadcasting, (2016).