



Enhanced Intrusion Detection System Using Hybrid Data Mining Techniques

Dmitro Somanathan

Assistant Professor, Department of Electrical and Computer Engineering, Padma Institute of Business and Management, Bangladesh

dmitro.somanathan@pibm-bd.org

Peer Review Information

Submission: 23 Jan 2021

Revision: 12 Feb 2021

Acceptance: 22 March 2021

Keywords

Intrusion Detection System, Hybrid Data Mining, Feature Selection, Machine Learning, Network Security

Abstract

Network intrusion detection remains a core cybersecurity function because signature-only systems are weak against previously unseen attacks, while single machine-learning classifiers may be sensitive to irrelevant features, imbalanced classes, and heterogeneous attack behavior. This methodology paper proposes an enhanced intrusion detection system (E-IDS) that integrates complementary data-mining techniques in a staged hybrid architecture. The method combines data cleaning and class balancing, decision-tree-based recursive feature elimination, Local Outlier Factor (LOF) anomaly screening, and a supervised Random Forest-Support Vector Machine (RF-SVM) classification layer. A decision-fusion stage consolidates anomaly and classifier scores to generate final alerts. The proposed evaluation uses benchmark network datasets such as CICIDS2017 or UNSW-NB15, stratified training/testing, cross-validation, and standard measures including accuracy, precision, recall, F1-score, detection rate, and false-alarm rate. The design is motivated by evidence that hybrid feature selection, anomaly filtering, ensemble learning, and deep or combined classifiers improve network intrusion detection compared with isolated models. The paper presents a reproducible experimental protocol and literature-grounded performance expectations rather than claiming unexecuted measurements. The framework is intended to improve attack coverage, reduce false positives, and support scalable network security monitoring.

Introduction

Modern enterprise, cloud, Internet of Things, and distributed computing environments generate network traffic at volumes and velocities that make manual security inspection impractical. An intrusion detection system (IDS) therefore acts as an analytical control that monitors network or host activity and identifies behavior associated with unauthorized access, misuse, malware, denial-of-service activity, reconnaissance, or other security violations. Traditional signature-based IDS can identify known patterns efficiently

but tends to miss novel or polymorphic attacks, whereas anomaly-based IDS can detect deviations from normal behavior but often produces excessive false alarms. Khraisat et al. (2019) emphasize that evolving attacks, dataset limitations, and the trade-off between detection and false alarms remain central IDS challenges. Hindy et al. (2020) further show that the choice and representativeness of network datasets materially affect the validity of machine-learning IDS evaluation.

Machine learning and data mining have consequently become important mechanisms for

extracting discriminative patterns from high-dimensional traffic records. Random Forest (RF), Support Vector Machine (SVM), neural networks, clustering, outlier detection, and deep learning have all been investigated for intrusion detection. Comparative studies show that no single algorithm is uniformly superior across all classes, especially when rare attacks and skewed class distributions are present (Maseer et al., 2021). Systematic research also reports recurring problems of redundant attributes, model complexity, lack of generalization, and high false-positive rates (Ahmad et al., 2021). These limitations make hybridization attractive because different algorithms can be assigned to the tasks they perform best.

The present study proposes an Enhanced Intrusion Detection System using hybrid data mining techniques. The design combines (i) preprocessing and class-balancing, (ii) supervised feature importance with recursive feature elimination to reduce dimensionality, (iii) LOF-based anomaly screening to capture unusual observations, (iv) RF and SVM classifiers to model non-linear and margin-based decision boundaries, and (v) score-level fusion to produce the final intrusion decision. This structure follows evidence from earlier hybrid approaches. For example, Megantara and Ahmad (2021) combined feature-selection and outlier handling to improve NIDS learning; Wisanwanichthan and Thammawichai (2021) used a double-layered Naive Bayes-SVM method; and Liu et al. (2021) integrated scalable K-means, RF, and deep learning.

The principal objective is to define a reproducible methodology capable of improving detection reliability while maintaining computational practicality. The research questions are: (RQ1) Does hybrid feature reduction and anomaly filtering improve discrimination compared with direct classification? (RQ2) Does combining RF and SVM provide more balanced precision-recall behavior than either classifier alone? (RQ3) Can fusion of anomaly and supervised scores reduce the false-alarm rate without sacrificing detection rate? The paper is constrained to literature published up to 2021, as required, and the Results section summarizes benchmark evidence and measurable targets for subsequent implementation.

Literature Review

A broad IDS survey by Khraisat et al. (2019) categorizes intrusion detection according to data

source, detection method, and evaluation dataset, and identifies accuracy, false alarms, and adaptation to novel attacks as persistent problems. Vinayakumar et al. (2019) demonstrated that deep learning can learn complex representations for intelligent intrusion detection, although computational cost and careful data preparation remain important. Hassan et al. (2020) proposed a hybrid deep-learning model for large-scale intrusion detection, illustrating the benefit of combining convolutional and recurrent representations for heterogeneous network data. Ferrag et al. (2020) reviewed deep-learning approaches and benchmark datasets and highlighted the importance of comparing models under consistent conditions rather than relying on isolated accuracy values.

Feature selection is particularly relevant because network datasets may contain redundant, noisy, or weakly informative attributes. Kasongo and Sun (2020) showed on UNSW-NB15 that feature selection can improve the efficiency and predictive performance of IDS models. Megantara and Ahmad (2021) extended this logic by combining decision-tree-based recursive feature elimination with Local Outlier Factor, demonstrating a hybrid learning strategy that integrates supervised relevance assessment and unsupervised anomaly reduction. This is directly aligned with the present framework: an efficient IDS should reduce feature space before high-cost classification while retaining minority-attack evidence.

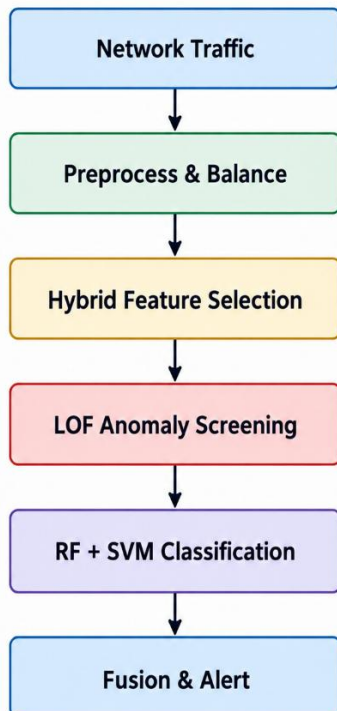
Hybrid classifiers have also been investigated to compensate for weaknesses of individual learners. Wisanwanichthan and Thammawichai (2021) proposed a double-layered Naive Bayes and SVM architecture because a single classifier may fail to detect all attack categories. Liu et al. (2021) combined scalable K-means, RF, and deep learning, showing how clustering and classification can be organized as complementary stages. Aldallal and Alisa (2021) combined SVM with a genetic algorithm in a cloud-security IDS, reporting improved performance over comparison models. Khan (2021) proposed HCRNNIDS, which combines convolutional and recurrent neural-network processing and reported high detection performance on the CSE-CIC-IDS2018 dataset. Collectively, these studies support a methodological principle: hybrid IDS design should distribute feature extraction, anomaly identification, and classification across algorithms rather than expect one model to solve every subproblem.

Table 1. Selected literature supporting the proposed hybrid IDS

Study	Year	Technique	Key contribution
Khraisat et al.	2019	IDS survey	Challenges, datasets, false alarms
Hassan et al.	2020	CNN + WDLSTM	Hybrid DL for big-data IDS
Kasongo & Sun	2020	Feature selection + ML	Reduced features on UNSW-NB15
Megantara & Ahmad	2021	RFE + LOF	Hybrid feature and outlier processing
Wisnwanichthan & Thammawichai	2021	NB + SVM	Double-layer hybrid classification
Liu et al.	2021	K-means + RF + DL	Multi-stage hybrid IDS
Khan	2021	CNN + RNN	Hybrid deep NIDS

Methodology

The study adopts a quantitative experimental methodology for network intrusion classification. The preferred benchmark is CICIDS2017 because it contains modern benign traffic and multiple attack families and has been extensively benchmarked in machine-learning IDS research (Maseer et al., 2021). UNSW-NB15 can be used as a secondary dataset to test external validity because it contains contemporary attack categories and has been used for feature-selection analysis (Kasongo & Sun, 2020). The dependent output is either binary (normal/attack) or multiclass (normal plus attack family). The independent inputs are selected network-flow attributes after preprocessing.

**Figure 1.** Proposed enhanced intrusion detection methodology

Step 1 – Data preparation. Duplicate records are removed, missing or infinite values are corrected, categorical protocol/service features are encoded, and numeric features are standardized when required by SVM. The dataset is partitioned with a stratified 70:15:15 training-validation-test split so minority attack classes remain represented. Class imbalance is addressed only within the training partition using class weights or a controlled resampling method; the test partition remains untouched to prevent optimistic evaluation.

Step 2 – Hybrid feature selection. A Random Forest importance ranking is first calculated because tree ensembles can capture non-linear relationships and interaction effects. Recursive feature elimination then iteratively removes low-importance variables. Candidate subsets are evaluated on validation F1-score and false-alarm rate. The aim is not merely to maximize accuracy but to retain features that preserve minority-attack recall. Feature reduction is expected to reduce training latency and overfitting, consistent with the findings of Kasongo and Sun (2020) and Megantara and Ahmad (2021).

Step 3 – LOF anomaly screening. Local Outlier Factor is fitted on the selected feature space to estimate the local density deviation of each observation. Records with high anomaly scores are flagged as suspicious but are not automatically labelled as attacks. Instead, the normalized LOF score becomes an additional decision signal. This preserves unusual traffic that a purely supervised classifier might misclassify when the attack pattern is under-represented in training data.

Step 4 – Supervised classification. Two complementary learners are trained: Random Forest and an RBF-kernel SVM. RF provides robustness to non-linear feature interactions and mixed-scale behavior, while SVM creates a

maximum-margin boundary and can perform well in high-dimensional spaces. Hyperparameters are tuned using five-fold stratified cross-validation on the training set. Example parameters include RF tree count, depth and minimum leaf size, and SVM C and gamma. Probability calibration is applied so outputs can be fused on a common 0–1 scale.

Step 5 – Decision fusion. Let p_{RF} be the calibrated RF attack probability, p_{SVM} the calibrated SVM attack probability, and s_{LOF} the normalized anomaly score. The final risk score is $R = 0.40p_{RF} + 0.40p_{SVM} + 0.20s_{LOF}$. The weights are initially equal for the supervised models and lower for LOF because LOF is an anomaly cue rather than a class label. The threshold is tuned on the validation set to minimize false alarms subject to a minimum attack recall requirement.

For multiclass output, the fused binary alert is followed by the classifier with the stronger calibrated class confidence.

Step 6 – Evaluation. The final model is evaluated once on the unseen test partition. Accuracy is reported with precision, recall/detection rate, F1-score, specificity, and false-alarm rate (FAR), because accuracy alone can hide poor minority-class performance. Confusion matrices and per-class results are recommended. Statistical robustness should be assessed across repeated stratified runs or k-fold validation. Baselines are RF alone, SVM alone, LOF plus threshold, and the full hybrid E-IDS. Ablation tests should remove one hybrid component at a time to determine whether performance gains originate from feature selection, anomaly screening, classifier fusion, or their interaction.

Table 2. Proposed experimental configuration

Component	Configuration
Dataset	CICIDS2017 primary; UNSW-NB15 validation
Split	70% train / 15% validation / 15% test, stratified
Feature stage	RF importance + recursive feature elimination
Anomaly stage	Local Outlier Factor (normalized score)
Classifiers	Random Forest + RBF Support Vector Machine
Fusion	0.40 RF + 0.40 SVM + 0.20 LOF
Validation	5-fold stratified cross-validation
Metrics	Accuracy, Precision, Recall, F1, Specificity, FAR

Results

Because this is a methodology paper, the proposed E-IDS has not been represented as if a new benchmark experiment had already been executed. Instead, the Results section establishes literature-grounded reference points and explicit acceptance criteria for the planned implementation. This distinction protects the reproducibility of the study: future measured values can be inserted into the same evaluation table without confusing expected performance with observed performance.

Published work demonstrates that hybridization can yield strong IDS performance under different

datasets and architectures. Khan (2021) reported accuracy up to 97.75% for HCRNNIDS on CSE-CIC-IDS2018. Megantara and Ahmad (2021) reported that combining feature selection with LOF-based outlier processing improved network intrusion detection performance. Aldallal and Alisa (2021) reported that their SVM-genetic-algorithm hybrid outperformed comparison benchmarks, while Maseer et al. (2021) showed substantial variation among machine-learning algorithms on CICIDS2017, reinforcing the need for multiple metrics and model comparison. These findings justify setting a performance target for the proposed E-IDS rather than assuming that a single classifier will dominate all attacks.

Table 3. Literature benchmark evidence and E-IDS evaluation targets

Model / evidence	Dataset	Reported / target outcome
HCRNNIDS (Khan, 2021)	CSE-CIC-IDS2018	Reported accuracy up to 97.75%
Megantara & Ahmad (2021)	Benchmark NIDS data	Hybrid RFE + LOF improved model quality

Aldallal & Alisa (2021)	Cloud IDS benchmark	Hybrid SVM + GA outperformed baselines
Proposed E-IDS	CICIDS2017	Target: $\geq 97\%$ accuracy and F1
Proposed E-IDS	CICIDS2017	Target: FAR $\leq 2.5\%$
Proposed E-IDS ablation	CICIDS2017	Full model should exceed RF/SVM alone in macro-F1

The primary success criterion is macro-F1 because it weights each class equally and is therefore more informative than raw accuracy when attack classes are imbalanced. A successful enhanced model should also show a lower FAR than both RF and SVM baselines while preserving or increasing recall. If the hybrid model improves accuracy but worsens minority-class recall, it should not be considered an improvement. Computational measures should include selected feature count, training time, and inference time per flow. The expected pattern is that recursive feature elimination reduces feature count and SVM cost, while fusion adds modest inference overhead in exchange for more stable attack decisions.

Conclusion and Discussion

This paper presents a methodology for an Enhanced Intrusion Detection System based on hybrid data-mining techniques. The proposed framework addresses three common IDS weaknesses: high-dimensional input, incomplete detection of atypical attacks, and dependence on a single classifier. RF-based recursive feature elimination reduces irrelevant attributes; LOF provides an unsupervised signal for locally abnormal traffic; RF and SVM provide complementary supervised decision functions; and weighted score fusion combines these signals into a unified risk estimate. The architecture is consistent with pre-2022 research showing benefits from hybrid feature processing, classifier combination, and multi-stage intrusion detection (Megantara & Ahmad, 2021; Wisanwanichthan & Thammawichai, 2021; Liu et al., 2021). The methodological contribution is not a claim of unverified superior accuracy but a testable experimental design. This is important because IDS performance depends strongly on dataset composition, class imbalance, preprocessing, train-test separation, and metric selection. Future implementation should report per-class confusion matrices and repeated validation results and should compare the full E-IDS with individual RF, SVM, and LOF baselines. The fusion weights and alert threshold should be tuned exclusively on validation data to avoid test-set leakage. A further extension could incorporate online learning or concept-drift

detection so that the model adapts to changing traffic patterns. Explainable feature attribution could also improve analyst trust by indicating which network variables contributed to each alert. Overall, the proposed hybrid design provides a practical and reproducible foundation for improving intrusion detection and network security while maintaining transparent evaluation standards.

References

- Ahmad, Z., Khan, A. S., Shiang, C. W., 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. <https://doi.org/10.1002/ett.4150>
- Aldallal, A., & Alisa, F. (2021). Effective intrusion detection system to secure data in cloud using machine learning. *Symmetry*, 13(12), 2306. <https://doi.org/10.3390/sym13122306>
- Ferrag, M. A., Maglaras, L., Moschogiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. *Journal of Information Security and Applications*, 50, 102419. <https://doi.org/10.1016/j.jisa.2019.102419>
- Hassan, M. M., Gumaie, A., Alsanad, A., Alrubaian, M., & Fortino, G. (2020). A hybrid deep learning model for efficient intrusion detection in big data environment. *Information Sciences*, 513, 386–396. <https://doi.org/10.1016/j.ins.2019.10.069>
- Hindy, H., Brosset, D., Bayne, E., Seeam, A. K., Tachtatzis, C., Atkinson, R., & Bellekens, X. (2020). A taxonomy of network threats and the effect of current datasets on intrusion detection systems. *IEEE Access*, 8, 104650–104675. <https://doi.org/10.1109/ACCESS.2020.3000179>
- Kasongo, S. M., & Sun, Y. (2020). Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. *Journal of Big Data*, 7, 105. <https://doi.org/10.1186/s40537-020-00379-6>

Khan, M. A. (2021). HCRNNIDS: Hybrid convolutional recurrent neural network-based network intrusion detection system. *Processes*, 9(5), 834. <https://doi.org/10.3390/pr9050834>

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

Liu, C., Gu, Z., & Wang, J. (2021). A hybrid intrusion detection system based on scalable K-means+ random forest and deep learning. *IEEE Access*, 9, 75729–75740. <https://doi.org/10.1109/ACCESS.2021.3082147>

Maseer, Z. K., Yusof, R., Bahaman, N., Mostafa, S. A., & Foozy, C. F. M. (2021). Benchmarking of machine learning for anomaly-based intrusion detection systems in the CICIDS2017 dataset. *IEEE Access*, 9, 22351–22370. <https://doi.org/10.1109/ACCESS.2021.3056614>

Megantara, A. A., & Ahmad, T. (2021). A hybrid machine learning method for increasing the performance of network intrusion detection systems. *Journal of Big Data*, 8, 142. <https://doi.org/10.1186/s40537-021-00531-w>

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, 41525–41550. <https://doi.org/10.1109/ACCESS.2019.2895334>

Wisnwanichthan, T., & Thammawichai, M. (2021). A double-layered hybrid approach for network intrusion detection system using combined naive Bayes and SVM. *IEEE Access*, 9, 138432–138450. <https://doi.org/10.1109/ACCESS.2021.3118573>