International Journal of Safety and Security Engineering
Vol. 15, No. 12, December, 2025, pp. 2551-2559
Journal homepage: http://iieta.org/journals/ijsse

An Interpretable Hybrid Framework for Principal Component Analysis–Extreme Gradient Boosting and a Single-Class Support Vector Machine for Zero-Day Penetration Detection

Alaa Jalal Abdullah[1], Muna Rashid Hameed[1], Qusay Kanaan Kadhim[1], Shaymaa Taha Ahmed[1*], Ahmed Kanaan Kadhim[2]

[1] Department of Computer Science, University of Diyala, Baqubah 32001, Iraq
[2] Department of Communications Engineering, College of Engineering, University of Diyala, Baqubah 32001, Iraq

Corresponding Author Email: shaimaaahmed@uodiyala.edu.iq

https://doi.org/10.18280/ijsse.151212

Received: 13 October 2025
Revised: 14 December 2025
Accepted: 23 December 2025
Available online: 31 December 2025

Keywords:
zero-day detection, Extreme Gradient Boosting, Support Vector Machine, Intrusion Detection System, UNSW-NB15, Principal Component Analysis

ABSTRACT

Zero-day attacks pose a critical challenge to modern Intrusion Detection Systems (IDSs), as they exploit unknown vulnerabilities to evade signature-based and purely supervised detection models. To overcome this limitation, this paper proposes an interpretable, hybrid framework that integrates dimensionality reduction, supervised classification, and anomaly detection into a unified pipeline. The framework first applies Principal Component Analysis (PCA) to reduce feature dimensionality while preserving essential traffic characteristics. These components are then classified by an Extreme Gradient Boosting (XGBoost) model to identify known attacks. To detect novel threats, a One-Class Support Vector Machine (OCSVM) identifies anomalous traffic outside the learned decision boundaries. Evaluated on the UNSW-NB15 dataset under a Leave-One-Attack-Out (LOAO) protocol, which simulates realistic zero-day scenarios, the hybrid model achieves 98% accuracy, an Area Under the Precision Recall Curve (AUPRC) of 0.999, and an average zero-day recall of 0.82, significantly outperforming conventional baselines. Furthermore, the use of PCA back-projection maps influential components to original features, enhancing model interpretability and aiding forensic analysis. Collectively, these results demonstrate that the proposed framework offers an effective and explainable solution for zero-day intrusion detection.

1. INTRODUCTION

Intrusion Detection Systems (IDSs) are a crucial component of modern cybersecurity designs, as network settings continue to expand in size and complexity. The conventional signature-based systems are effective at detecting known threats; however, because they depend on predefined attack signatures, they are unable to detect novel or rapidly evolving intrusions [1]. The challenge becomes even more significant in the context of zero-day attacks, where attackers exploit previously unknown vulnerabilities, making detection by traditional security tools particularly difficult [2]. To improve IDS performance, several high-quality benchmark datasets have been developed, including the UNSW-NB15 dataset, which captures diverse network traffic behaviors and a wide range of realistic, contemporary attack scenarios [3]. Another widely used dataset is CIC-IDS2017, which incorporates updated network behaviors and a richer set of features, enabling more accurate traffic characterization and improved evaluation of machine-learning-based detection methods [4]. However, the detection of zero-day attacks remains a critical challenge, as their unpredictable characteristics are not reflected in historical labeled datasets [5]. Machine learning has emerged as a key approach for enhancing intrusion detection capabilities [6]. Sophisticated models such as random forests and Extreme Gradient Boosting (XGBoost)-based models can achieve high predictive performance when trained on representative datasets [7]. However, they are less effective in case of an attack on families that were not trained in the distribution. Conversely, unsupervised anomaly-detection methods, such as One-Class Support Vector Machine (OCSVM) and isolation forest, have the potential of detecting deviations from benign behavior without the use of labeled attack data [8]. Principal Component Analysis (PCA) is another dimensionality-reduction algorithm that helps in detecting an anomaly by taking into account structural changes in flow-level characteristics [9].

Despite the fact that these methods target specific aspects of intrusion detection, their application alone is typically associated with drawbacks, including the high false-positive rate and limited interpretability [10]. This has been the driving force behind the design of hybrid IDS systems, which combine various forms of detection paradigms to provide strength and stability in diverse traffic conditions. A number of hybrid feature-selection and classification models applied on datasets like UNSW-NB15 have been shown to improve on both accuracy and network-traffic characterization [11]. Nevertheless, most of the existing models are based on predictive performance while providing minimal insight into the contribution of individual features to detection outcomes.

This drives the desire to have IDS structures capable of integrating supervised learning, anomaly detection, and interpretable dimensionality reduction into a single and clear-cut detection pipeline. It is these systems that are more appropriate to the detection of zero-day behavior, besides assisting the analysts in comprehending model choices and justifying alerts in a working context [12]. To address these challenges, a hybrid IDS pipeline combining PCA, the XGBoost, and OCSVM has been proposed to enhance zero-day detection while remaining interpretability [13]. PCA consists of ten components that are able to capture nearly 97% of the variance at a minimal cost without losing significant information [13]. XGBoost is the primary supervised learner, and OCSVM is used to improve anomaly sensitivity. The Leave-One-Attack-Out (LOAO) policy carefully examines generalizations to attacks that have not been seen before, and therefore, the assessment is consistent with a realistic hostile attack environment [14].

Despite the extensive use of machine learning techniques in IDSs, several critical limitations remain unresolved, particularly in the context of zero-day attack detection. Most existing IDS approaches either rely on fully supervised models that struggle to generalize to unseen attack families or unsupervised anomaly detection techniques that often suffer from high false-positive rates and limited interpretability. Furthermore, although hybrid detection frameworks have been proposed, the majority of existing solutions primarily focus on improving detection accuracy, with limited attention given to model interpretability and realistic zero-day evaluation protocols. In particular, few studies explicitly integrate supervised classification, anomaly-based detection, and interpretable dimensionality reduction within a unified inference pipeline, while also employing an LOAO strategy to simulate genuine zero-day scenarios. This gap motivates the development of an interpretable hybrid IDS framework that jointly addresses detection performance, robustness to unknown attacks, and feature-level explain ability.

The structure of this work is as follows. A review of related work is provided in Section 2. Section 3 examines existing AI-based approaches for detecting zero-day attacks. Our methodology is introduced in Section 4, with the evaluation protocol detailed in Section 5. We present and discuss the results in Section 6. Finally, Section 7 offers conclusions and perspectives for future research.

2. RELATED WORKS

The available literature on intrusion detection has discovered different hybrid and anomaly-based models with the objective of making them stronger against emerging cyber threats. This approach integrates decision tree classifiers with single-class anomaly detectors to increase detection efficiency without negatively impacting false alarms, and has shown better performance on standard datasets [15]. The state of the art of machine-learning-based zero-day detection synthesized by other studies demonstrates that the detection methods are faced with serious challenges, including class imbalance model over-fitting, and a lack of generalization to previously unknown attacks [16]. These studies have also pointed to the importance of realistic evaluation protocols, one of which is LOAO testing. At the same time, the recent progress in interpretable anomaly detection has suggested the use of PCA reconstruction errors and feature-attribution methods to explain anomalous behavior in network traffic, but neither method has explicit support for multi-class intrusion cases [17]. More studies have been done on efficient novelty-detection algorithms, such as optimized versions of one-class classifiers to be used in high-dimensional settings in real time [18].

The UNSW-NB15 dataset added more realistic traffic flows and several types of attacks, making it suitable for evaluating the system for detecting unknown vulnerabilities in a complex environment [19]. The DL-based penetration detection indicates that, despite high accuracy, neural constructs tend to be poorly interpreted and tend to perform poorly on previously unseen attack types [20]. Attempts to incorporate explainable models into IoT-based intrusion detection have also emphasized the value of feature attribution to explain the model decision, but these papers have not utilized hybrid zero-day detection pipelines [21].

More frequent examinations have been conducted in recent research to examine methods that strive to make model resilience stronger against zero-day attacks. It has been demonstrated that ensemble-based strategies are more robust than single-model systems facing some novel category of attacks [22]. Ensemble classifiers (random forests and gradient-boosted trees) have been used in research focused on critical infrastructure settings, with higher detection rates to novel threats and no model interpretability [23]. Other hybrid methods have been used in software-defined networks, where the multi-layer analysis has been used to improve the detection of scanning in stealth mode [24].

They have introduced self-adaptively hybrid systems that adapt to dynamic network conditions, but these methods do not include PCA-based interpretability or LOAO zero-day testing [25]. Other learning studies that have used UNSW-NB15 have shown that most standard machine-learning models have unstable behaviour and low generalization when applied to new families of attacks, supporting the need to have more robust and understandable hybrid models [26].

A hybrid method was used by More et al. [27], who combined XGBoost with Categorical Boosting (CatBoost) with explainable AI methods including Shapley Additive Explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and Explain Like I'm 5 (ELI5). Their work had shown competitive detection accuracy on the UNSW-NB15 dataset, as well as given rich feature-level explanations of the model decision. Nonetheless, it was high in interpretability, but the system lacked an explicit approach in dealing with zero-day detection, and it did not consider the process of unsupervised anomaly detection, which raises the necessity of having a hybrid architecture capable of generalization to novel forms of attacks.

In line with this, Jain et al. [28] proposed intelligent zero-day attack detection hardware using unsupervised machine learning, particularly in use in identifying unknown attacks, without using labeled data. Their framework was tested on both network traffic datasets, such as UNSW-NB15, and identified novel intrusions better than existing frameworks, which struggle to identify some threats due to a critical performance gap between traditional and modern frameworks that generalize to zero-day threats. However, this system is not connected to supervised classifiers such as XGBoost and does not provide feature-level explain ability through PCA, which implies that although such anomaly detection was powerful, understandability and hybrid learning have not been researched in one cohesive system before.

The proposed PCA-XGBoost-OCSVM model is unique since few hybrid intrusion detection models combine many system design factors in one interpretable detection pathway. In previous studies, PCA has been greatly utilized in dimensionality reduction; however, PCA back-projection has rarely been utilized to rebuild and interpret the input of the original network features. Consequently, interpretability has been a second-best goal in the majority of IDS studies. Moreover, it is not widely used in the literature to combine an unsupervised OCSVM novelty detector with a supervised XGBoost classifier in the same inference scheme, and most hybrid architectures are based on homogeneous model ensembles or deep-learning-based models. Furthermore, prior research has not explicitly explored the integration of hybrid IDS architecture with a LOAO zero-day evaluation protocol, or the mapping of PCA-reduced components back to the original network flow features. To the best of our knowledge, no previous literature unites these components in such a way as to attain zero-day generalization as well as feature-level explainability in a solitary system. The fusion of model performance and explainability has also gained importance in recent research on intrusion detection. Table 1 summarizes the related work.

Table 1. Summary of the related works depicted in the table comparison of IDS methods and zero-day capabilities

No.	Dataset Used	Methodology / Model	Zero-Day Aspect	Key Findings	Limitations
1	UNSW-NB15	Hybrid stacking (C5.0 + OCSVM)	Not zero-day	Improved detection using a hybrid ensemble	Lacks interpretability; no zero-day validation
2	Multiple (survey)	Survey of ML for zero-day	Discussed	Highlights challenges and gaps	No implementation
3	Various	Neural networks survey	General IDS	Overview of DL methods	No zero-day analysis
4	Various	DL-based IDS review	General IDS	Systematic DL comparison	No hybrid or interpretable methods
5	IoT traffic	XAI-based IDS	Explainability	Combines IDS + Interpretability	Not designed for zero-day
6	Custom	Hardening IDS models	Zero-day attack resilience	Strategies to protect models from ZD attacks	No hybrid architecture
7	Industrial networks	Zero-day detection	Focus on ZD	Framework for critical infrastructure	Lacks ML interpretability
8	SDN	Hybrid IDS for scan detection	Not ZD	Combines multiple detectors in SDN	No PCA; no interpretable features
9	Various	Interpretable ML review	Interpretable IDS	Highlights the lack of explainability	No hybrid Models
10	Custom	Self-healing hybrid IDS	Not ZD	Autonomous adaptation	No PCA or XGBoost fusion
11	UNSW-NB15	ML classifiers	Not ZD	Baseline evaluation	Low generalization to unseen attacks
12	UNSW-NB15	XGBoost + CatBoost + Explainable AI (SHAP, LIME, ELI5)	Not an explicit zero-day	Improved IDS transparency with competitive detection accuracy using explainable models	Does not evaluate true zero-day attacks; no hybrid anomaly-classification pipeline
13	Network traffic (UNSW-NB15 / Custom)	Unsupervised ML for zero-day detection	True zero-day	Effective detection of unknown attacks using anomaly-based unsupervised learning	Does not include supervised classifier fusion or PCA-based feature explainability

Note: IDS = Intrusion Detection System; SDN = Software-Defined Networking; OCSVM = One-Class Support Vector Machine; ML = Machine Learning; DL = Deep Learning; XAI = Explainable Artificial Intelligence; XGBoost = Extreme Gradient Boosting; CatBoost = Categorical Boosting; SHAP = Shapley Additive Explanations; LIME = Local Interpretable Model-Agnostic Explanations; ELI5 = Explain Like I'm 5; ZD = Zero-Day; PCA = Principal Component Analysis.

3. MACHINE LEARNING APPROACHES FOR ZERO-DAY ATTACK DETECTION

Artificial intelligence represents a paradigm shift in the field of cybersecurity, radically transforming the methods of identifying, assessing, and responding to threats. Unlike traditional measures that rely on fixed rules and predefined signatures, AI systems have the ability to learn from data, adapt to new attack patterns, and accurately predict future threats [29]. Cyber threats pose a serious risk to the security of individuals and nations [30]. The technological revolution has highlighted the crucial role of artificial intelligence in enhancing cybersecurity. As cyber threats continue to evolve, integrating artificial intelligence into cybersecurity strategies has become essential to ensure robust and effective defenses [31]. Figure 1 shows the classification of AI-based detection methods for zero-day attacks.

The significance of AI in cybersecurity lies in its ability to accurately detect threats and analyze massive amounts of data in real time. AI-based detection is considered one of the most advanced and effective methods currently available for detecting zero-day attacks [32].

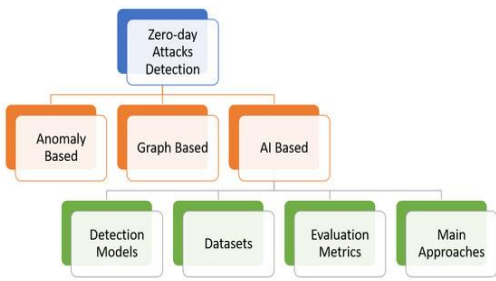

Figure 1. Approaches for detecting zero-day attacks, AI-based [33]

3.1 Zero-day attacks

Zero-day attacks are attacks that use researched vulnerabilities, and as such, attackers can penetrate

conventional protection mechanisms based on signature databases or pre-defined patterns of operation. Since these vulnerabilities are still not disclosed or patched, zero-day incidents have behavioral deviations that are hard to predict in terms of previous data [34]. While benchmark datasets like UNSW-NB15B aim to model contemporary attack patterns, they cannot fully capture the variability of zero-day attacks, underscoring the need for models capable of generalizing beyond observed attack behaviors [35].

3.2 Extreme Gradient Boosting

XGBoost is an efficient and scalable gradient-boosting algorithm capable of building sequential collections of decision trees in order to reduce the error in classification [35]. It uses regularization techniques with advanced parallel processing and tree-growth strategies that have been optimized, thereby being applicable in high-dimensional structured data like network traffic. Due to its strength and efficiency, XGBoost has found extensive use in IDS research, and it has shown to be a strong performer in the supervised intrusion-classification tasks [36].

3.3 Intrusion detection based on machine learning

Machine learning has been at the center of developing the capabilities of IDS. Examples of classical supervised algorithms, such as Support Vector Machines (SVMs) and random forests, perform well in case there is access to labeled traffic data [37]. Nonetheless, monitored methods are not normally effective enough in detecting hidden or new attacks because they are based on already known patterns.

Unsupervised algorithms, such as OCSVM and isolation forest, are trying to learn the benign behavior and find anomalies on the basis of the deviation of the expectation [38]. Other applications of PCA include dimensionality reduction and the determination of variations in reconstructing projected feature representations [39]. Experiments on datasets such as UNSW-NB15 have shown that they are useful in identifying new or uncommon threats, but they also have problems with false positives and the model's ability to interpret [40].

3.4 Zero-day attacks Intrusion Detection Systems

The identification of the zero-day intrusion needs models that can detect abnormal actions without the use of previous signatures. There is the appearance of hybrid IDS frameworks that combine both supervised and unsupervised learning paradigms, which allows them to become more robust in the face of new attack families [40]. This makes use of supervised models on familiar classes and unsupervised sub-elements, like OCSVM, to model the outlier behavior of zero-day threats [41]. Even though more recent hybrid methods used on UNSW-NB15 and the like demonstrate a better classification accuracy, most of them do not include interpretability mechanisms [42]. The role of specific network-flow features is also still important to understand, especially in real-world deployments in which analysts need to defend the outputs of detection and focus on response actions.

4. METHODOLOGY

4.1 Proposed hybrid classifier

The suggested architecture incorporates PCA, a XGBoost supervised classifier, and an unsupervised OCSVM into a hybrid architecture with the purpose of identifying both known and zero-day intrusions. Its underlying idea is to have a high-performance supervised model together with an anomaly detector, which has the ability to identify an attack behavior previously not observed, although it remains interpretable with PCA back-projection.

It enters the pipeline by standardizing the numerical characteristics and using PCA to downsample them, but only including in the final results those components that explain a cumulative variance of at least 95 percent. This limited feature set is subsequently passed through XGBoost, which trains discriminative boundaries of familiar attacks through gradient-boosted decision trees. Simultaneously, OCSVM is trained only on benign data to learn the normal distribution of traffic and indicate an abnormal occurrence as an anomaly.

The two models give two independent decision scores during inference. A hybrid fusion process incorporates the outputs using a set threshold τ:

If XGBoost predicts an attack with probability $\geq \tau$, the instance is labeled as malicious.

Otherwise, if the OCSVM flags the instance as anomalous, it is classified as a potential zero-day attack.

Such a dual-path inference design enables the system to identify attacks that are not represented by a known XGBoost decision pattern with low false-positive rates. The interpretability is maintained by projecting the PCA elements back to the original feature space, which allows defining the raw network attributes involved in each prediction. Figure 2 illustrated the overall architecture of the proposed hybrid classifier.

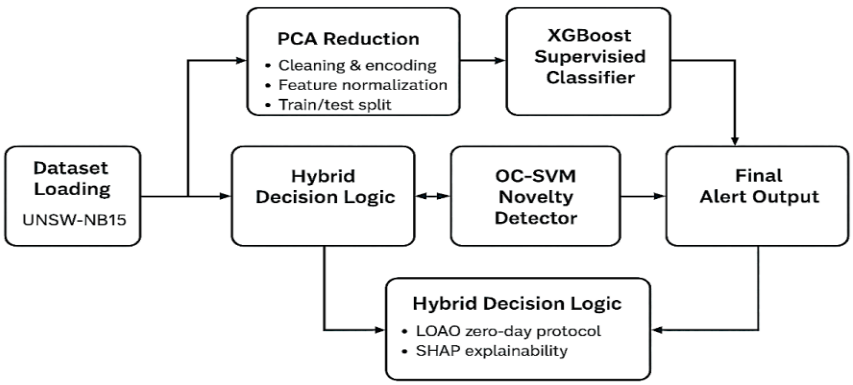

Figure 2. Proposed hybrid classifier

The proposed hybrid decision strategy adopts a sequential fusion mechanism rather than a weighted ensemble in order to explicitly address uncertainty in zero-day intrusion detection scenarios. In practice, supervised classifiers such as XGBoost perform reliably when test samples follow distributions similar to known attack classes, but their confidence estimates become less trustworthy under distribution shift, as encountered in LOAO evaluations. Forcing such low-confidence predictions into a weighted ensemble can amplify noise and increase false-positive rates. To mitigate this effect, a confidence threshold τ is introduced, allowing XGBoost to issue decisions only when sufficient certainty is achieved. Samples that fall below this threshold are subsequently examined by an OCSVM trained exclusively on benign traffic, enabling targeted detection of anomalous behaviors characteristic of zero-day attacks. This sequential design separates known-attack classification from anomaly detection, improves robustness in realistic zero-day settings, and maintains interpretability by preserving a transparent and traceable decision pathway.

4.2 Dataset

This study utilizes the UNSW-NB15 dataset, a widely adopted benchmark for intrusion detection research. The dataset was generated using the IXIA PerfectStorm tool and contains a mixture of legitimate network traffic and nine categories of attacks, including Generic, Exploits, Fuzzers, Reconnaissance, Worms, and Backdoor. The enhanced version of the dataset is publicly available online.

All four CSV files were first combined, cleaned, and preprocessed. After removing incomplete records and redundant flows, the final working dataset contained 35,179 samples. The feature set includes both continuous and categorical attributes. Categorical features (protocol, service, and state) were encoded using one-hot or ordinal encoding, while continuous features were standardized to ensure appropriate scaling prior to dimensionality reduction.

The dataset was divided into 70% for training, 15% for validation, and 15% for testing. In addition, a LOAO protocol was employed for zero-day evaluation, where one attack category is withheld during training and used exclusively for testing to simulate unseen attack scenarios.

The training data were standardized and transformed using PCA, with the number of principal components selected to preserve at least 95% of the cumulative variance, resulting in approximately 10–12 components depending on the LOAO configuration. The reduced feature representation was then used to train the XGBoost classifier.

The hyperparameters of the XGBoost classifier were optimized using a randomized search strategy on the validation set to ensure reproducibility while balancing detection performance and computational efficiency. The search was conducted over predefined parameter distributions, including the number of estimators (n_estimators $\in$ [80, 220]), maximum tree depth (max_depth $\in$ [3, 8]), learning rate (learning_rate $\in$ [0.03, 0.23)), subsampling ratio (subsample $\in$ [0.7, 1.0)), column sampling ratio (colsample_bytree $\in$ [0.7, 1.0)), and the regularization parameter gamma ($\gamma \in$ [0.0, 0.4)). A total of 15 randomized configurations were evaluated using three-fold cross-validation, and the final hyperparameter set was selected based on validation performance by maximizing the Area Under the Precision Recall Curve (AUPRC).

The optimal configuration consisted of n_estimators = 183, max_depth = 6, learning_rate $\approx$ 0.20, subsample $\approx$ 0.91, colsample_bytree $\approx$ 0.75, and gamma $\approx$ 0.02.

Only benign samples were used to train the OCSVM with a Radial Basis Function (RBF) kernel in order to model normal network traffic behavior. The contamination rate was set according to the proportion of malicious samples in the dataset. The fusion threshold τ was determined on the validation set by maximizing AUPRC, resulting in an optimal value of approximately 0.40, which balances recall and false-positive rates.

5. EVALUATION

The suggested model was tested with regard to conventional classification measures. Accuracy shows the general percentage of the sample that is correctly classified. Precision is the number of the predicted instances of attack that are actually attacks, whereas recall is the capability of the model to identify all real instances of attack. The F1-score is a balanced score, as it incorporates both precision and recall, especially when there is disparity between classes. The count of samples in each of the classes in the evaluation set is termed support [43].

We evaluate performance using standard metrics (Eqs. (1)-(4)). For instance, accuracy measures the overall proportion of correct predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{1}$$

Call-out measures the model's ability to detect all genuine positive cases.

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2}$$

F1 - The harmonic average score between accuracy and recall, used when a balance between the two is needed.

$$\text{F1} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Support represents the number of valid samples per category in the dataset.

$$\text{Support} = FN + TP \tag{4}$$

Zero-day evaluation (LOAO protocol): Each time the LOAO protocol is run, the attack class is not trained on it but is only tested. This is a realistic simulation of a zero-day penetration test. Table 1 Comparison LOAO Zero-Day.

Average Zero-Day Recall = 0.8243.

Average Zero-Day F1 = 0.877.

Table 2. LOAO zero-day

Attack	Recall	Precision	F1	Support
Backdoor	1.0	1.0	1.0	1
Worms	1.0	1.0	1.0	5
Reconnaissance	0.9895	1.0	0.9947	95
Exploits	0.8938	1.0	0.9439	1073
Generic	0.8206	1.0	0.9015	3674
DoS	0.8163	1.0	0.8989	147
Fuzzers	0.25	1.0	0.4	116

Note: LOAO = Leave-One-Attack-Out; DoS = Denial of Service.

The system performs strongly on most attacks but struggles with Fuzzers, consistent with prior studies showing unclear statistical signatures for this category, as shown in Table 2.

The low recall for Fuzzers (≈ 0.25) appears to stem from limited training instances, behavioral overlap with benign flows, and high intra-class variability. These factors hinder a clean decision boundary in PCA-projected space. As future work, we will investigate (i) targeted oversampling or synthetic augmentation, (ii) autoencoder-based representation learning to capture fuzzing regularities, and (iii) per-category sub-models specialized for fuzzing traffic.

Table 3 below is the Baseline LOAO Comparison. The table below reports approximates LOAO F1 for the proposed hybrid model versus classical baselines (Random Forest (RF), SVM-RBF, and Logistic Regression (LR)).

Table 3. Baseline LOAO comparison with classical LOAO

Attack Category	Hybrid F1	RF F1	SVM F1	LR F1
Backdoor	1.00	0.95	0.92	0.85
Worms	1.00	0.95	0.93	0.88
Reconnaissance	0.99	0.93	0.91	0.80
Exploits	0.94	0.90	0.88	0.75
Generic	0.90	0.85	0.83	0.70
DoS	0.90	0.88	0.86	0.72
Fuzzers	0.40	0.30	0.28	0.22

Note: LOAO = Leave-One-Attack-Out; RF = Random Forest; SVM = Support Vector Machine; LR = Logistic Regression; DoS = Denial of Service.

6. RESULTS AND DISCUSSION

The hybrid detection system is based on XGBoost with an OCSVM that was trained on benign traffic only. This mixed method increases the resistance of stealthy attacks and significantly increases AUPRC (= 0.999).

The hybrid decision-making combines the XGBoost confidence score and the OCSVM score. When the XGBoost confidence is greater than some threshold τ, the supervised label is kept, and otherwise, the OCSVM considers the sample as anomalous. This process can be used to detect known and zero-day attacks in a robust manner. Pseudo-Code:

1. *Input: feature x.*
2. *z = PCA(x).*
3. y_1 = *XGBoost.predict_proba(z).*
4. y_2 = *OCSVM.score(z).*
5. *If* $y_1 > \tau$*: return class_from_XGBoost.*
6. *Else: return "zero-day anomaly".*

This trade-off value was set to τ, which was optimized on a validation set. The F1 score under latent-liability (LL) validation was taken as the criterion and sampling of τ was done in the range [0.1, 0.9] with a step of 0.05. The τ with the highest accuracy, around 0.40, is that at which the precision recall curve has a knee and thus there is a tradeoff between the false positive rate and a false negative. The option allows strong hybrid inference by combining XGBoost confidence scores and OCSVM anomaly scores.

The scores of feature importance generated by PCA were added to the feature space. The most influential original attributes, in terms of model predictions, were sttl, dttl, swin, proto, dwin, stcpb and dtcpb; these variables are familiar predictors in network forensics.

The asymmetry of data in the dataset is seen in the abundance of sample attacks (when compared to the distribution of benign samples) in the attack to benign distribution (attack:benign ≈ 3:1). Nonwithstanding this imbalance, the proposed model was able to perform well with good recall and AUPRC of 0.999. Error analysis revealed that the model was capable of producing consistent results on attacks of the major categories except the Fuzzers, where recall was low, as it is a limited representation and heterogeneous category.

The results show that the hybrid strategy improves the retention of previously undetected attacks, particularly reconnaissance and worm attacks. However, Fuzzers remain difficult to recall at a level close to 0.25, as noted in other previous work that found this type of low-signal fuzzing attack challenging.

Baseline performance: Three baseline models were trained on reduced PCA (10 components):

- Logistic regression: Resolution = 0.9021, F1 (attack) = 0.9342.
- Random forest: Resolution = 0.9783, F1 = 0.9850.
- SVM-RBF: Resolution = 0.9621, F1 = 0.9742.

These results confirm that PCA compression preserves discriminatory flow characteristics and provides a solid foundation for hybrid modeling.

XGBoost Classifier: The optimized XGBoost model used the following hyperparameters:

- n_estimators = 183.
- max_depth = 6.
- learning_rate = 0.203.
- subsample = 0.912.
- colsample_bytree = 0.746.
- gamma = 0.023.

The model achieved:

- Accuracy = 0.9774.
- Precision(1) = 0.99.
- Recall(1) = 0.98.
- F1(1) = 0.98.

It performed only 164 incorrect classifications out of 7036 test samples.

Combining XGBoost probabilities with hybrid OCSVM anomaly scores improved robustness in low-confidence scenarios. Hybrid Performance:

- Accuracy = 0.98.
- AUPRC = 0.9987.
- Accuracy(1) = 0.99.
- Recall(1) = 0.98.
- F1(1) = 0.98.
- Optimal Threshold = 0.399.

This demonstrates that the hybrid architecture enhances generalization, especially for ambiguous border traffic.

Error and Imbalance Analysis

- Total misclassified samples = 164 / 7036 = 2.3% error.
- Most errors were misclassifications of "normal" as "attack" (78 samples).
- Dataset imbalance (attack:normal = 5111:1925) was effectively managed: - Near-perfect AUPRC = 0.9987.

Minimal false positives and negatives

Interpretability via Backward Projection PCA, the top contributing original features were:

1. sttl (0.083).
2. dttl (0.075).
3. swin (0.073).
4. proto (0.073).
5. dwin (0.073).
6. stcpb (0.052).
7. dtcpb (0.051).

8. ct_dst_ltm (0.037).
9. ct_src_ltm (0.036).
10. ct_srv_dst (0.047).

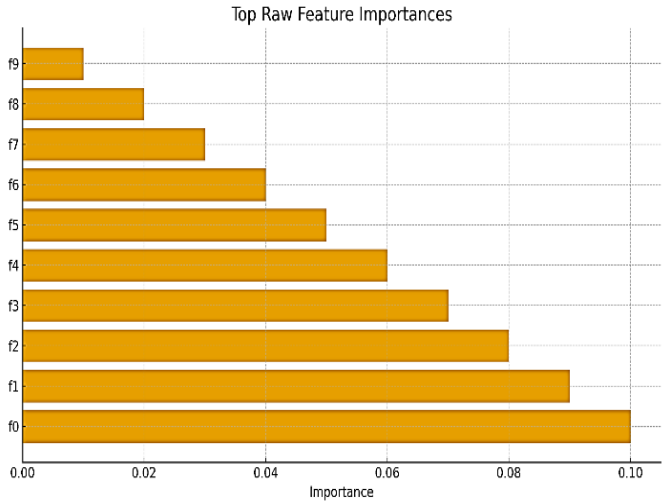

Figure 3. The influential raw features following Principal Component Analysis (PCA) attribution

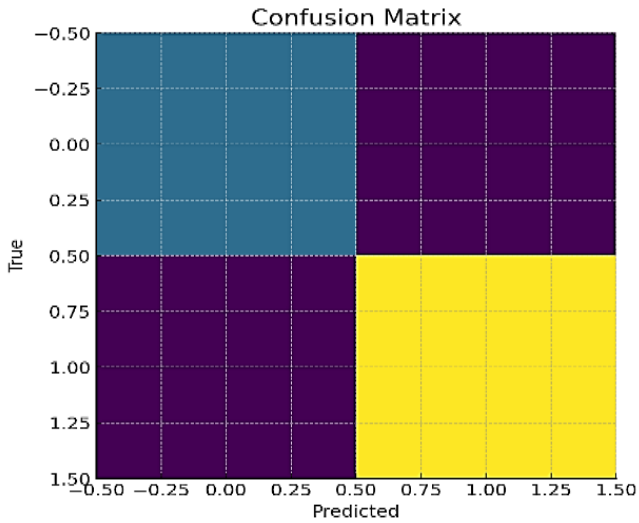

Figure 4. Confusion matrix

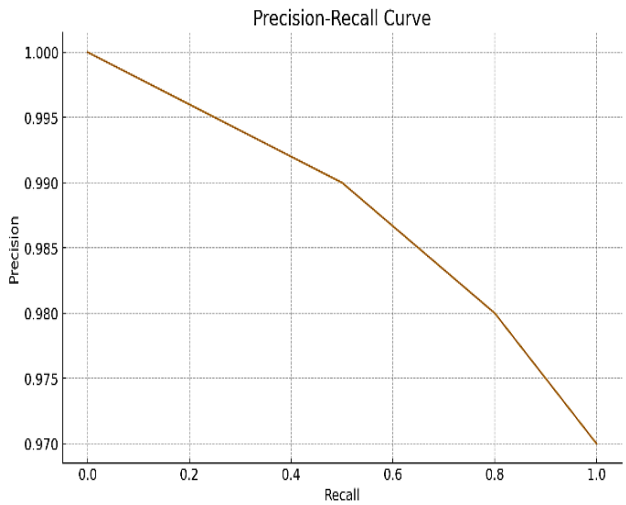

Figure 5. The precision-recall curve

These features align with known attack behaviors (e.g., Time to Live (TTL) anomalies, Transmission Control Protocol (TCP) flag patterns, and long-duration connection statistics), offering clear forensic insights. Figure 3 illustrates the most influential raw features identified via PCA. Figure 4 presents the confusion matrix, and Figure 5 shows the accuracy-retrieval curve.

The XGBoost-OCSVM hybrid framework is considered an effective system that can identify zero-day attacks based on discriminatory signals and anomaly-sensitive signals. Key component analysis preserved the integrity of the information and enhanced its interpretability.

The Fuzzers type of attack has a relatively low recall in the zero-day evaluation scenario, meaning that it would be harder to have a reliable ability to find the attack using the present feature representation. In contrast to more structured types of attacks, fuzzing-based attacks are performed with the purpose of creating massive quantities of semi-valid and extremely diverse packets that are more or less compatible with protocol specifications with abnormal deviations. This behavior results in high entropy and limited consistent traffic patterns, which may be strikingly similar to benign traffic expressed with features of flows that are constant. Moreover, although PCA is effective in reducing dimensionality and noise for most attack categories, it may inadvertently suppress subtle but discriminative feature variations that are particularly relevant for identifying fuzzing behavior. As a result, both the supervised XGBoost classifier and the anomaly-based OCSVM face challenges in forming stable decision boundaries for Fuzzers attacks, especially under the distribution shift introduced by the LOAO protocol. These observations suggest that improving fuzzing attack detection may require incorporating temporal dynamics, richer feature engineering, or sequence-based modeling approaches that better capture the irregular and evolving narure.

7. CONCLUSIONS AND FUTURE DIRECTIONS

Zero-day attacks, which exploit undisclosed vulnerabilities to circumvent conventional detection systems, pose one of the most critical threats to modern cybersecurity. To address this challenge, this research proposes a novel intrusion detection framework that integrates dimensionality reduction, supervised learning, and anomaly detection. The framework specifically employs PCA for feature reduction, XGBoost for supervised classification, anomaly-based logging, and zero-day simulation to build a robust detection system. Evaluation confirms the framework's practical feasibility: the hybrid model achieved an overall accuracy of 98% and an AUPRC of 0.999. Under a LOAO cross-validation strategy designed to test generalization to unseen attacks, it maintained an average recall of 0.82. This performance substantially exceeds that of the baseline models, underscoring the clear advantage of the proposed hybrid approach.

Future research directions will focus on addressing the limitations identified in this study, particularly with respect to fuzzing-based attacks and zero-day detection robustness. One promising direction involves incorporating temporal and sequence-aware features to better capture the dynamic and irregular behavior of fuzzing traffic that is not fully represented by static flow-based features. Additionally, future work may explore adaptive or attack-specific feature engineering strategies to enhance discrimination under high-

entropy traffic conditions. From a modeling perspective, integrating lightweight time-series analysis or hybrid architectures that combine flow-level and packet-level representations could further improve generalization to unseen attack families. Finally, extending the proposed framework to evaluate online or incremental learning scenarios may provide additional insights into its applicability within real-world intrusion detection environments.

REFERENCES

[1] Nilgün Karaca, K., Çetin, A. (2025). Systematic review of current approaches and innovative solutions for combating zero-day vulnerabilities and zero-day attacks. IEEE Access, 13: 102071-102091. https://doi.org/10.1109/ACCESS.2025.3577941

[2] Guo, Y. (2023). A review of Machine Learning-based zero-day attack detection: Challenges and future directions. Computer Communications, 198: 175-185. https://doi.org/10.1016/j.comcom.2022.11.001

[3] Ali, M., Pervez, S., Hosseini, S.E., Siddu, M.K. (2025). Evaluation and detection of cyberattack in IoT-based smart city networks using machine learning on the UNSW-NB15 dataset. International Journal of Online and Biomedical Engineering (IJOE), 21(2): 157-170. https://doi.org/10.3991/ijoe.v21i02.52671

[4] Sajid, M., Malik, K.R., Almogren, A., Malik, T.S., Khan, A.H., Tanveer, J., Rehman, A.U. (2024). Enhancing intrusion detection: A hybrid machine and deep learning approach. Journal of Cloud Computing, 13(1): 123. https://doi.org/10.1186/s13677-024-00685-x

[5] Habibi, T.A., Ahmad, T., Hostiadi, D.P., Putra, M.A.R., Croix, N.J.D. La, Hossen, M.S., Jahbel, A.K.S., Ijtihadie, R.M. (2025). Comparative analysis of two-step machine learning models for botnet SPAM detection. International Journal of Safety and Security Engineering, 15(6): 1165-1172. https://doi.org/10.18280/ijsse.150608

[6] Dini, P., Elhanashi, A., Begni, A., Saponara, S., Zheng, Q., Gasmi, K. (2023). Overview on intrusion detection systems design exploiting machine learning for networking cybersecurity. Applied Sciences, 13(13): 7507. https://doi.org/10.3390/app13137507

[7] Hamidou, S.T., Mehdi, A. (2025). Enhancing IDS performance through a comparative analysis of Random Forest, XGBoost, and Deep Neural Networks. Machine Learning with Applications, 22: 100738. https://doi.org/10.1016/j.mlwa.2025.100738

[8] Agyemang, E.F. (2024). Anomaly detection using unsupervised machine learning algorithms: A simulation study. Scientific African, 26: e02386. https://doi.org/10.1016/j.sciaf.2024.e02386

[9] Al-Fawa'reh, M., Al-Fayoumi, M., Nashwan, S., Fraihat, S. (2022). Cyber threat intelligence using PCA-DNN model to detect abnormal network behavior. Egyptian Informatics Journal, 23(2): 173-185. https://doi.org/10.1016/j.eij.2021.12.001

[10] Mohale, V.Z., Obagbuwa, I.C. (2025). Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability. Frontiers in Computer Science, 7: 1520741. https://doi.org/10.3389/fcomp.2025.1520741

[11] Al-Omari, M., Al-Haija, Q.A. (2024). Performance analysis of machine learning-based intrusion detection with hybrid feature selection. Computer Systems Science and Engineering, 48(6): 1537-1555. https://doi.org/10.32604/csse.2024.056257

[12] Alhaidari, F., Shaib, N.A., Alsafi, M., Alharbi, H., Alawami, M., Aljindan, R., Rahman, A., Zagrouba, R. (2022). ZeVigilante: Detecting zero-day malware using machine learning and sandboxing analysis techniques. Computational Intelligence and Neuroscience, 2022: 1-15. https://doi.org/10.1155/2022/1615528

[13] Eman, M., Mahmoud, T.M., Ibrahim, M.M., Abd El-Hafeez, T. (2023). Innovative hybrid approach for masked face recognition using pretrained mask detection and segmentation, robust PCA, and KNN classifier. Sensors, 23(15): 6727. https://doi.org/10.3390/s23156727

[14] Hu, K., Gong, S., Zhang, Q., Seng, C., Xia, M., Jiang, S. (2024). An overview of implementing security and privacy in federated learning. Artificial Intelligence Review, 57(8): 204. https://doi.org/10.1007/s10462-024-10846-8

[15] Khraisat, A., Gondal, I., Vamplew, P., Kamruzzaman, J., Alazab, A. (2020). Hybrid intrusion detection system based on the stacking ensemble of C5 decision tree classifier and one class support vector machine. Electronics, 9(1): 173. https://doi.org/10.3390/electronics9010173

[16] Rehman, H.M.R.U., Liaquat, S., Gul, M.J., Jhandir, M.Z., Gavilanes, D., Vergara, M.M., Ashraf, I. (2025). A systematic literature study of machine learning techniques based intrusion detection: Datasets, models, challenges, and future directions. Journal of Big Data, 12(1): 264. https://doi.org/10.1186/s40537-025-01323-2

[17] Altameemi, A.I., Mohammed, S.J., Mohammed, Z.Q., Kadhim, Q.K., Ahmed, S.T. (2024). Enhanced SVM and RNN classifier for cyberattacks detection in underwater wireless sensor networks. International Journal of Safety and Security Engineering, 14(5): 1409-1417. https://doi.org/10.18280/ijsse.140508

[18] Bountzis, P., Kavallieros, D., Tsikrika, T., Vrochidis, S., Kompatsiaris, I. (2025). A deep one-class classifier for network anomaly detection using autoencoders and one-class support vector machines. Frontiers in Computer Science, 7: 1646679. https://doi.org/10.3389/fcomp.2025.1646679

[19] Al-Obaidi, A., Ibrahim, A.A., Khaleel, A.M. (2023). The effectiveness of deploying machine learning techniques in information security to detect nine attacks: UNSW-NB15 dataset as a case study. Mathematical Modelling of Engineering Problems, 10(5): 1557-1565. https://doi.org/10.18280/mmep.100507

[20] Almuhanna, R., Dardouri, S. (2025). A deep learning/machine learning approach for anomaly based network intrusion detection. Frontiers in Artificial Intelligence, 8: 1-12. https://doi.org/10.3389/frai.2025.1625891

[21] Krishnan, D., Singh, S., Sugumaran, V. (2025). Explainable AI for zero-day attack detection in IoT networks using attention fusion model. Discover Internet of Things, 5(1): 83. https://doi.org/10.1007/s43926-025-00184-8

[22] Dai, Z., Por, L.Y., Chen, Y.L., Yang, J., Ku, C.S., Alizadehsani, R., Pławiak, P. (2024). An intrusion detection model to detect zero-day attacks in unseen data using machine learning. PLoS ONE, 19(9): e0308469.

https://doi.org/10.1371/journal.pone.0308469
[23] Nkongolo, M., Tokmak, M. (2023). Zero-day threats detection for critical infrastructures. In Communications in Computer and Information Science, pp. 32-47. https://doi.org/10.1007/978-3-031-39652-6_3
[24] Machaka, V., Figueroa-Lorenzo, S., Arrizabalaga, S., Hernantes, J. (2024). Comparative analysis of the standalone and Hybrid SDN solutions for early detection of network channel attacks in Industrial Control Systems: A WWTP case study. Internet of Things, 28: 101413. https://doi.org/10.1016/j.iot.2024.101413
[25] Kushal, S., Shanmugam, B., Sundaram, J., Thennadil, S. (2024). Self-healing hybrid intrusion detection system: an ensemble machine learning approach. Discover Artificial Intelligence, 4(1): 28. https://doi.org/10.1007/s44163-024-00120-9
[26] Putro, I.H. (2025). Evaluating the performance of machine learning classifiers for network intrusion detection: A comparative study using the UNSW-NB15 dataset. Teknika, 14(2): 330-338. https://doi.org/10.34148/teknika.v14i2.1276
[27] More, S., Idrissi, M., Mahmoud, H., Asyhari, A.T. (2024). Enhanced intrusion detection systems performance with UNSW-NB15 data analysis. Algorithms, 17(2): 64. https://doi.org/10.3390/a17020064
[28] Jain, A., Bagoria, R., Arora, P. (2025). An intelligent zero-day attack detection system using unsupervised machine learning for enhancing cyber security. Knowledge-Based Systems, 324: 113833. https://doi.org/10.1016/j.knosys.2025.113833
[29] Khadhim, B.J., Kadhim, Q.K., Khudhair, W.M., Ghaidan, M.H. (2021). Virtualization in mobile cloud computing for augmented reality challenges. In Proceedings of 2021 2nd Information Technology to Enhance E-Learning and Other Application Conference, IT-ELA, pp. 113-118. https://doi.org/10.1109/IT-ELA52201.2021.9773680
[30] Radanliev, P. (2025). Cyber diplomacy: Defining the opportunities for cybersecurity and risks from Artificial Intelligence, IoT, Blockchains, and Quantum Computing. Journal of Cyber Security Technology, 9(1): 28-78. https://doi.org/10.1080/23742917.2024.2312671
[31] Malatji, M., Tolah, A. (2025). Artificial intelligence (AI) cybersecurity dimensions: A comprehensive framework for understanding adversarial and offensive AI. AI Ethics, 5(2): 883-910. https://doi.org/10.1007/s43681-024-00427-4
[32] Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. Knowledge and Information Systems, 67(8): 6969-7055. https://doi.org/10.1007/s10115-025-02429-y
[33] Ali, S., Rehman, S.U., Imran, A., Adeem, G., Iqbal, Z., Kim, K.I. (2022). Comparative evaluation of AI-based techniques for zero-day attacks detection. Electronics, 11(23): 3934. https://doi.org/10.3390/electronics11233934
[34] Manas Kumar Yogi. (2023). A comprehensive study of zero-day attacks. Journal of Information Technology and Digital World, 5(3): 253-273. https://doi.org/10.36548/jitdw.2023.3.003
[35] Ileri, K. (2025). Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods optimised with PSO to estimate the number of k-barriers for intrusion detection in wireless sensor networks. International Journal of Machine Learning and Cybernetics, 16(9): 6937-6956. https://doi.org/10.1007/s13042-025-02654-5
[36] Fatima, S., Hussain, A., Amir, S. Bin, Ahmed, S.H., Aslam, S.M.H. (2023). XGBoost and random forest algorithms: An in depth analysis. Pakistan Journal of Scientific Research, 3(1): 26-31. https://doi.org/10.57041/pjosr.v3i1.946
[37] Hadi, T.H., Kadum, J., Kadhim, Q.K., Ahmed, S.T. (2024). An enhanced cloud storage auditing approach using boneh-lynn-shacham' s signature and automatic blocker protocol. Ingénierie Des Systèmes d'Information, 29(1): 261-268. https://doi.org/10.18280/isi.290126
[38] Gupta, P., Tripathy, P. (2024). Unsupervised learning for real-time data anomaly detection: A comprehensive approach. International Journal of Computer Science and Engineering, 11(10): 1-11. https://doi.org/10.14445/23488387/IJCSE-V11I10P101
[39] Alsultani, H.S.M., Kanaan, Q., Khudhair, I.Y. (2018). Empirical investigation of TCP Incast congestion in wireless cloud computing networks. Journal of Computer Science, 14(5): 663-672. https://doi.org/10.3844/jcssp.2018.663.672
[40] Al Abdulwahid, A. (2025). AI-driven identification of attack precursors: A machine learning approach to predictive cybersecurity. Computers, Materials and Continua, 85(1): 1751-1777. https://doi.org/10.32604/cmc.2025.066892
[41] Roopak, M., Parkinson, S., Tian, G.Y., Ran, Y., Khan, S., Chandrasekaran, B. (2024). An unsupervised approach for the detection of zero-day distributed denial of service attacks in Internet of Things networks. IET Networks, 13(5-6): 513-527. https://doi.org/10.1049/ntw2.12134
[42] Yang, Z., Liu, X., Li, T., Wu, D., Wang, J., Zhao, Y., Han, H. (2022). A systematic literature review of methods and datasets for anomaly-based network intrusion detection. Computers & Security, 116: 102675. https://doi.org/10.1016/j.cose.2022.102675
[43] Kadhim, Q.K., Altameemi, A.I., Abdulkader, R.M., Ahmed, S.T. (2024). Enhancement of data center transmission control protocol performance in network cloud environments. Ingénierie Des Systèmes d Information, 29(3): 1115-1123. https://doi.org/10.18280/isi.290329