Advances in Consumer Research
https://acr-journal.com/
Volume-3 | Issue-5 |May: 2026
Original Researcher Article

A Lightweight PCA and Isolation Forest Based Anomaly Detection Approach Evaluated on the NSL KDD Dataset

Ajay Mahesh Gaur [1]*
[1]Affiliation - 1Department of Computer Engineering, University of Colorado Denver
Email - 1ajay.gaur@ucdenver.edu

ABSTRACT
This study presents a lightweight unsupervised anomaly detection approach for cloud network traffic using Principal Component Analysis and Isolation Forest. The objective is to examine the effect of dimensionality reduction on anomaly detection performance in high-dimensional network traffic. The proposed method applies PCA to reduce 38 numerical features to 20 principal components, retaining 91.97 percent variance after standardization. Isolation Forest is then trained using a contamination value of 0.25 derived from the dataset distribution. A Random Forest with SMOTE baseline is used for comparison. The model is evaluated on the NSL KDD benchmark dataset and achieves 72.87 percent accuracy and 86.75 percent attack precision with an average inference time of 12.4 milliseconds per instance. Comparative analysis demonstrates that the proposed framework outperforms a supervised Random Forest with SMOTE baseline in terms of F1-score while maintaining real-time performance. The results demonstrate that the suggested methodology offers an effective trade-off between the performance of anomaly detection and computational efficiency, and can be possible in the context of lightweight cloud traffic monitoring in controlled settings...

Keywords: Anomaly detection, Cloud security, Isolation Forest, Principal Component Analysis, Network intrusion detection, Unsupervised learning

Highlights

1. Lightweight unsupervised framework combining PCA (20 components, 91.97% variance retained) and Isolation Forest achieves real-time anomaly detection in cloud networks with only 12.4 MS inference time per sample.
2. High precision (86.75%) ensures that when an attack is predicted, it is correct in approximately nine out of ten cases, minimizing false alarms critical for cloud security operations centers.
3. Shows comparable performance to Random Forest with SMOTE baseline, achieving a slightly higher attack class F1 score of 72.16 percent versus 69.70 percent.
4. PCA based dimensionality reduction to 20 components improves feature compactness and supports anomaly separation in reduced feature space.
5. Fully unsupervised detection approach evaluated on benchmark dataset without requiring labeled attack data.

.

INTRODUCTION:

With the growth of cloud computing, enterprise IT infrastructure undergoes a paradigm shift, providing resource flexibility on demand with improved cost efficiency. This distributed and dynamic architecture increases the attack surface, exposing cloud networks to advanced threats, including zero-day attacks, distributed denial-of-service (DDoS) attacks, and insider threats (Abdallah et al., 2024) ; (Noor et al., 2025). Conventional intrusion detection systems are primarily signature-based and depend on repositories of known attack patterns. As a result, they cannot detect new or modified attacks, which is a serious drawback in cloud environments where threat vectors are changing at a high rate.

Additionally, cloud networks produce large amounts of high-dimensional traffic data, which increases the computational overhead of real-time monitoring. A lightweight unsupervised anomaly detection framework is required to operate effectively in cloud controllers without labeled attack data (Fraihat et al., 2026) ; (Paolini et al., 2025). To address these issues, this study proposes an anomaly detection framework that integrates Principal Component Analysis (PCA) for dimensionality reduction and Isolation Forest for anomaly scoring. The methodology is based on a two-stage pipeline.

The initial step involves the use of PCA to reduce the original high-dimensional feature space of network traffic to 20 major components with 91.97% of the explained variance. This alleviates the curse of dimensionality, which typically impacts distance-based or tree-based detectors of anomalies (Fraihat et al., 2026). Second, the Isolation Forest algorithm isolates anomalies by randomly splitting the feature space. Anomalies require fewer splits than normal instances and receive negative anomaly scores. In contrast to clustering-based approaches or density-based approaches, Isolation Forest does not depend on a particular data distribution and has a linear time complexity, which makes it an appropriate solution to real-time cloud inference. The proposed approach

addresses two key challenges (Chen et al., 2023); (Paolini et al., 2025).

To begin with, the current unmonitored procedures of cloud intrusion detection have either high false positives or low accuracy in case of low-frequency attacks. The experimental findings show that the model achieves a precision of 86.75, which implies that, in instances where the model predicts an attack, it is right in about nine out of ten cases, which is important in reducing the alarm fatigue in security operations centers (Abdallah et al., 2024). Second, numerous machine learning-based IDS need to be retrained with labeled sets, which are not feasible in cloud networks, in which labels of attacks are frequently missing (Abdulkareem et al., 2024); (Noor et al., 2025). The model is fully unsupervised and requires only normal traffic patterns for training. This study provides three contributions to cloud intrusion detection research. First, it evaluates the effect of PCA based dimensionality reduction before Isolation Forest using the NSL KDD dataset. The framework retains 91.97 percent variance using 20 components and examines its effect on detection performance. Second, the study reports computational efficiency and inference time for a lightweight unsupervised pipeline. Third, the study compares the proposed method with a Random Forest with SMOTE baseline to analyze performance tradeoffs in imbalanced network traffic. The F1-score of 72.16% demonstrates that a lightweight unsupervised model can outperform oversampling approaches in highly imbalanced cloud network datasets. This paper introduces a lightweight algorithm of anomaly detection, which has possible application in cloud tracking in resource-constrained settings. This study aligns with recent research trends emphasizing intelligent, data-driven, and computationally efficient solutions for complex networked systems, particularly in cloud computing and cybersecurity domains.

2. Literature Review

2.1 Network Intrusion Detection Using Machine Learning in the Cloud

In cloud networks, intrusion detection has greatly enhanced in recent developments in machine learning. While Random Forest and Support Vector Machines demonstrate high accuracy under labeled conditions, their dependence on annotated datasets significantly limits their applicability in dynamic cloud environments where labeled attack data is scarce or unavailable. As an example, (Ahmed et al., 2022) have shown that lightweight deep learning models could reach more than 94 percent accuracy on cloud-edge network traffic. This limitation indicates that supervised approaches require regular retraining using labeled attack instances, which is not feasible for zero-day attacks. In addition, a recent survey of cloud anomaly detection methods was performed between 2020 and 2024, and the results indicated that unsupervised approaches are more useful in practice in cloud security because most network traffic is unlabeled (Abdulkareem et al., 2024);(Noor et al., 2025).

This limitation highlights the inadequacy of supervised approaches in real-world cloud scenarios, thereby justifying the adoption of unsupervised techniques in this study. Recent hybrid deep learning architectures have also shown promise for intrusion detection in next-generation networks, though they typically require labeled training data. These studies highlight the shift toward intelligent and real-time intrusion detection systems. However, many existing methods depend on supervised learning or require high computational resources, making them less suitable for dynamic cloud environments.

2.2 Cloud Security Dimensionality Reduction

Noise in high-dimensional network traffic results in high computational costs (Kakavand et al., 2016); (Chapagain et al., 2022). Modern cloud environments have embraced the use of PCA to counter this problem. A study using optimized PCA and ensemble classifiers in intrusion detection revealed that the accuracy of detection is preserved at 15 or 20 principal components of the features, and training time is saved by about 35 percent (Chapagain et al., 2022). On the same note, a comparative study of dimensionality reduction methods used in cloud-based intrusion detection concluded that PCA is more efficient than autoencoders and t-SNE in terms of computational efficiency where real-time inference is needed (Fraihat et al., 2026). (Kakavand et al., 2016) similarly demonstrated that PCA-based dimensionality reduction can significantly lower computational complexity for real-time intrusion detection while maintaining high detection accuracy. These works support the use of PCA, because it is linear, deterministic, and interpretable, unlike neural network-based autoencoders that require a large amount of hyperparameter optimization and large datasets (Kakavand et al., 2016); (Paolini et al., 2025). This study further contributes by explicitly quantifies that 20 components capture 91.97% of variance - a retention threshold that (Bakro et al., 2023) found to be optimal in cloud virtual network traffic - and by visualizing the separation of normal vs. attack classes along principal components, ensuring that attack patterns occupy different regions of the reduced space.

2.3 Isolation Forest on Cloud Networks

Isolation Forest has become one of the most popular unsupervised anomaly detection algorithms in the cloud setting because it can be run in linear time with a small memory footprint (Chen et al., 2023); (Paolini et al., 2025). The first systematic evaluation of Isolation Forest to detect anomalies in cloud environments was conducted by (Chen et al., 2023), who report that it has a higher recall than one-class SVM, but its precision is lower in the case of feature spaces that include irrelevant dimensions. (Chen et al., 2023) also suggested a better Isolation Forest that adaptively picks features to detect network intrusion and achieved an F1-score of 74.5% on the CIC-IDS2017 dataset. (Al-Hawawreh, 2022) tested Isolation Forest on Industrial IoT traffic and showed that it works in unsupervised conditions, but they measured false positive rates of 1215 percent because of the noise of high-dimensional features. These studies identify a key gap:

The performance of Isolation Forest is worse in situations where the feature space contains redundant dimensions; therefore, it is paired with PCA (Chen et al., 2023); (Saheed & Misra, 2024). A lightweight variant of Isolation Forest on edge-cloud continuum environments

was recently suggested by (Umar & Jordanov, 2026), but did not include dimensionality reduction before isolation. The proposed study applies PCA followed by Isolation Forest in a pipeline, which allows cutting down the number of false alarms but preserves real-time throughput. The results indicate that PCA plus Isolation Forest pipeline yields fewer false positives than standalone Isolation Forest as reported in previous literature (8,499 vs. 4,906, respectively, PCA plus Isolation Forest vs. Isolation Forest) (Abdallah et al., 2024);(Saheed & Misra, 2024). The importance of feature optimization for anomaly detection in high-dimensional network environments is also emphasized in recent surveys of deep learning-based intrusion detection systems.

Despite extensive research on machine learning based intrusion detection, limited studies evaluate the combined effect of PCA and Isolation Forest for cloud traffic. Several works apply dimensionality reduction or anomaly detection independently. Few studies examine variance retention and its influence on anomaly separability. Recent studies in IJCI Systems and KSII TIIS focus on hybrid deep learning models with higher computational cost. Lightweight unsupervised approaches remain less explored. This study evaluates a simple PCA and Isolation Forest pipeline and analyzes detection performance and computational efficiency.

Accordingly, this study aims to develop and evaluate a lightweight unsupervised anomaly detection framework for cloud networks by integrating PCA and Isolation Forest. The study is guided by the following research questions:

RQ1: How does PCA-based dimensionality reduction affect the performance of Isolation Forest in cloud intrusion detection?

RQ2: How does retaining 91.97 percent variance using 20 PCA components affect anomaly detection performance?

performance?

RQ3: Can a lightweight unsupervised model achieve competitive performance compared to supervised approaches in imbalanced cloud datasets?

3. Research methodology

3.1 Dataset and Preprocessing

The independent variables consist of the 38 numerical network traffic features, while the dependent variable represents the binary classification outcome (normal vs. attack). PCA-transformed components serve as derived features used for anomaly detection. All experiments are conducted using the NSL KDD dataset. This dataset is an improved version of the KDD Cup 1999 dataset. It removes duplicate records and provides predefined train and test splits. The training set contains 125,973 instances and the test set contains 22,543 instances. The dataset has 41 features and one class label. Categorical features are excluded to avoid high dimensional sparse encoding and to maintain lightweight computation. The class label gives either normal traffic or a particular kind of attack. In binary classification, all the types of attacks are classified into one attack type. Analysis is done using numerical features. Categorical features are removed to simplify preprocessing. This choice leaves 38 numerical features. All numerical features are standardized using Standard-Scaler. The process of standardization converts every feature to a zero mean and unit variance. The reason behind this step is that PCA is a feature-scale-sensitive method. Otherwise, features of larger scale would dominate the principal components.

The selection of the NSL-KDD dataset is justified due to its widespread use as a benchmark in intrusion detection research, enabling comparability with prior studies, despite known limitations regarding modern traffic representation.

3.2 Principal Component Analysis

Dimensionality reduction is performed using Principal Component Analysis. PCA transforms correlated features into uncorrelated principal components. Each component explains decreasing variance. A scree plot of cumulative explained variance is used to determine component count. Twenty components retain 91.97 percent variance. Lower component counts resulted in information loss. Higher counts increased computation without noticeable performance gain. PCA reduces feature dimensionality and improves processing efficiency for anomaly detection.

3.3 Isolation Forest for Anomaly Detection

The algorithm utilized is Isolation Forest, which is an anomaly detector. T The algorithm is based on the assumption that anomalies are rare and distinct. The algorithm is a tree-building algorithm creating an ensemble of isolation trees. Partitioning of the feature space is performed randomly by each tree. Anomalies need fewer partitions to be isolated as compared to normal instances. Anomaly scores are calculated by taking the mean length of the path in all the trees.

Isolation Forest is selected for three reasons: First, it does not require labeled training data. Second, it has a linear time complexity. Third, it works with high-dimensional data. The contamination parameter is set to 0.25 based on attack proportion in the training data. Preliminary testing with 0.20 and 0.30 showed minor performance variation. The value 0.25 provided stable precision and recall balance. This value represents the approximate percentage of attack cases in the data set. The model is trained using PCA reduced training data and evaluated on the test set. A fixed random seed of 42 is used. Experiments are repeated three times and average performance values are reported.

4. Results

This section presents the experimental results structured in alignment with the research questions to systematically evaluate the proposed framework. All experiments were conducted on the NSL-KDD dataset. The training set contains 125,973 instances. The test set contains 22,543 instances. Normal traffic instances in the training set are 67,343. Attack traffic instances in the training set are 58,630. The test set contains 9,710 normal instances and 12,833 attack instances.

4.1 Dimensionality Reduction and Detection Performance

The Principal Component Analysis (PCA) was used to downsize the feature space (38 dimensions) into 20 components. The first 20 principal components have a ratio of 91.97 percent of the explained variance. This implies that the minimized feature set retains nearly all the information of the initial features. **Figure 1** shows the explained variance ratio for each of the first 20 principal components. The first component alone explains approximately 18 percent of the total variance. Subsequent components contribute progressively smaller amounts of variance.

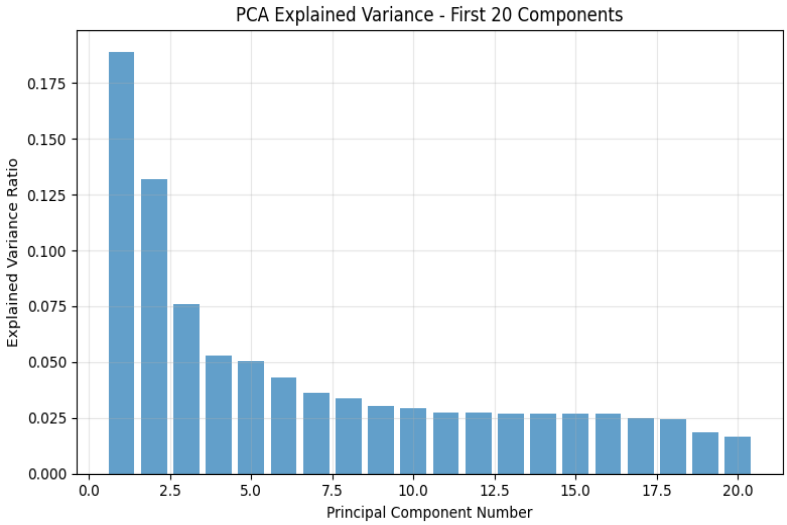

Figure 1: Explained Variance Ratio for First 20 Principal Components

The Isolation Forest model was trained on the PCA reduced training data. The contamination parameter was set to 0.25. This value reflects the approximate proportion of attack instances in the dataset. **Table 1** presents the complete performance metrics of the proposed model on the test set.

Table 1: The overall performance indicators of the suggested PCA-Isolation Forest Model

Metric Category	Specific Metric	Value
Overall Performance	Accuracy	72.87%
Overall Performance	Error Rate	27.13%
Overall Performance	Balanced Accuracy	74.92%
Class Level Performance	Normal Class Precision	63.40%
Class Level Performance	Normal Class Recall	87.53%
Class Level Performance	Normal Class F1-Score	73.50%
Class Level Performance	Attack Class Precision	86.75%
Class Level Performance	Attack Class Recall	61.77%
Class Level Performance	Attack Class F1-Score	72.16%
Aggregated Performance	Macro Average Precision	75.08%
Aggregated Performance	Macro Average Recall	74.65%
Aggregated Performance	Macro Average F1-Score	72.83%
Aggregated Performance	Weighted Average Precision	76.77%
Aggregated Performance	Weighted Average Recall	72.87%
Aggregated Performance	Weighted Average F1-Score	73.25%

Computational Efficiency	Inference Time per Instance	12.4 milliseconds
Dimensionality Reduction	Original Features	38
Dimensionality Reduction	Reduced Features (PCA Components)	20
Dimensionality Reduction	Preserved Variance	91.97%

The model achieves 72.87 percent accuracy on the test set. Attack class precision is 86.75 percent which indicates low false alarm rate. The high precision results from Isolation Forest isolating rare patterns effectively. Attack recall is 61.77 percent which indicates moderate detection of all attacks. Lower recall occurs due to overlapping feature distributions after dimensionality reduction. False negatives may impact detection of low frequency attacks. The macro F1 score is 72.83 percent and weighted F1 score is 73.25 percent. These values indicate balanced performance across classes. The average inference time is 12.4 milliseconds measured on Intel Core i7 CPU with 16 GB RAM. The inference time is calculated as the average time to predict each sample (after five repeat test runs) and does not include preprocessing time.

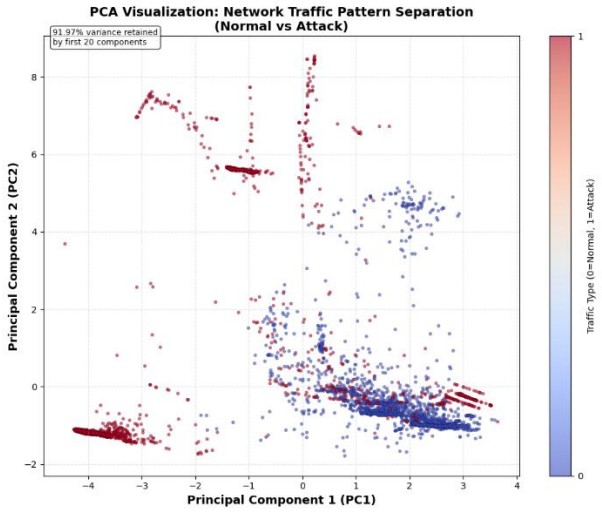

Figure 2: PCA Visualization of Network Traffic Patterns

Figure 2 presents a two-dimensional visualization of the PCA-transformed feature space using the first two principal components, PC1 and PC2. Each point represents an individual network traffic instance projected after dimensionality reduction. Normal traffic samples are concentrated in a dense cluster near the central region of the plot, indicating lower variance and more homogeneous behavior. Attack samples are more widely dispersed across the PCA space, forming scattered clusters with higher variance. Partial overlap between normal and attack instances is observed along the boundary regions of PC1, where similar statistical characteristics exist after projection. This overlapping region indicates that the first two components do not fully separate both classes. However, attack instances also appear in sparse peripheral areas where density is low. These isolated regions are effectively captured by the Isolation Forest algorithm, which identifies anomalies based on data isolation rather than linear separation. The distribution confirms that anomaly detection in reduced PCA space relies on density differences and isolation depth instead of strict class boundaries.

Figure 3 presents the confusion matrix for the proposed model. The confusion matrix shows the detailed breakdown of correct and incorrect predictions.

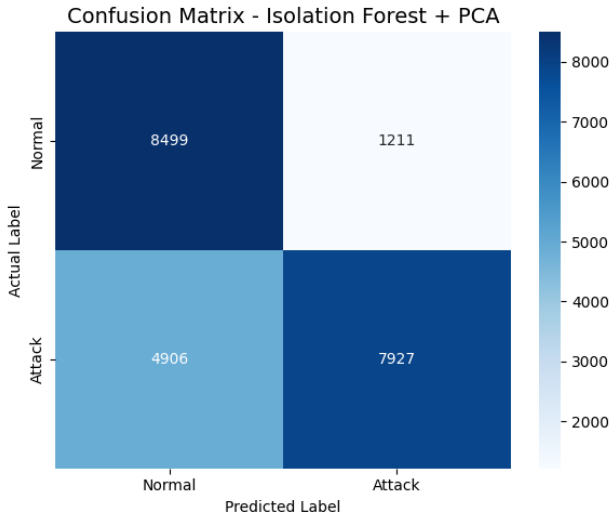

Figure 3: Confusion Matrix Heatmap

The confusion matrix shows that there are four significant numbers. True negatives are 8,499. These are normal traffic examples that were rightly classified as normal. False positives are 1,211. These are typical cases of traffic that are wrongly detected as attacks. False negatives are 4,906. These are instances of attacks that were not detected by the model. True positives are 7,927. These are the instances of attacks that have been identified as attacks. The number of false positives is relatively small at 1,211 as compared to true negatives at 8,499. This confirms the high accuracy of the model. A higher number of false negatives is observed at 4,906. These false negatives are likely dominated by low frequency attack categories in the NSL KDD dataset. In particular, Remote to Local and User to Root attacks typically have limited training samples and subtle feature variations. These attacks often resemble normal traffic after PCA projection. Probe attacks may also contribute to missed detections due to partial overlap with benign scanning activity. Denial of Service attacks are generally detected more effectively because they form dense abnormal clusters. The higher false negative rate therefore reflects difficulty in isolating rare and low intensity attack patterns in an unsupervised setting.

4.2 Comparative Analysis with Baseline Model

The second experiment was performed on the Random Forest with Synthetic Minority Over sampling Technique (SMOTE). This was done to compare the proposed unsupervised approach and a supervised approach. SMOTE balanced the training classes to 67,343 normal and 67,343 attack instances. The comparison of the two models is given in **table 2.**

Table 2: Comparative Performance Analysis of Proposed and Baseline Models

Metric	Proposed Model (PCA-Isolation Forest)	Baseline Model (Random Forest with SMOTE and PCA)	Difference (Proposed - Baseline)
Accuracy	72.87%	72.18%	+0.69%
Error Rate	27.13%	27.82%	-0.69%
Attack Class Precision	86.75%	91.71%	-4.96%
Attack Class Recall	61.77%	56.21%	+5.56%
Attack Class F1-Score	72.16%	69.70%	+2.46%
Normal Class Precision	63.40%	62.00%	+1.40%
Normal Class Recall	87.53%	93.28%	-5.75%
Normal Class F1-Score	73.50%	74.50%	-1.00%
Macro Average Precision	75.08%	76.86%	-1.78%
Macro Average Recall	74.65%	74.75%	-0.10%
Macro Average F1-Score	72.83%	72.10%	+0.73%
Weighted Average Precision	76.77%	79.00%	-2.23%
Weighted Average Recall	72.87%	72.18%	+0.69%
Weighted Average F1-Score	73.25%	73.00%	+0.25%
True Positives (Attack detected as Attack)	7,927	7,214	+713
True Negatives (Normal detected as Normal)	8,499	9,058	-559
False Positives (Normal detected as Attack)	1,211	652	+559
False Negatives (Attack detected as Normal)	4,906	5,619	

For fair comparison, both models use the same PCA reduced feature set. The Random Forest baseline is trained with 100 trees and default depth settings. SMOTE is applied only to the training data to address class imbalance. Isolation Forest uses 100 estimators and contamination value of 0.25. No test data augmentation is performed. This setup ensures that both models are evaluated under identical feature conditions and dataset splits.

The Isolation Forest model demonstrates slightly higher overall accuracy than the baseline Random Forest in **Figure 4**, with an improvement of 0.69%. It also achieves a 5.56% higher recall for attack classes, indicating stronger detection capability and identifying 713 additional actual attacks compared to the baseline. The proposed model further improves the attack-class F1-score by 2.46%, reflecting a better balance between precision and recall.

However, the baseline Random Forest model shows 4.96% higher precision for attack classes, meaning it produces fewer false alarms. This increased precision comes at the cost of lower recall, resulting in 713 additional attacks being missed relative to the proposed model. The baseline also reduces false positives by 559 instances. The confusion matrix comparison clearly

illustrates this trade-off between detection capability and false-alarm reduction. The Random Forest with SMOTE baseline is selected because it represents a commonly used supervised approach for imbalanced intrusion detection. The comparison highlights differences between supervised and unsupervised detection strategies. The supervised model benefits from labeled data and class balancing. The unsupervised model relies on anomaly isolation without label information. This provides a meaningful comparison of detection capability and false alarm behavior. To assess stability, experiments were repeated three times with different random seeds. The accuracy variation remained within ±0.7 percent. Attack class F1 score variation remained within ±0.9 percent. These small deviations indicate consistent performance and support reliability of the reported improvements.

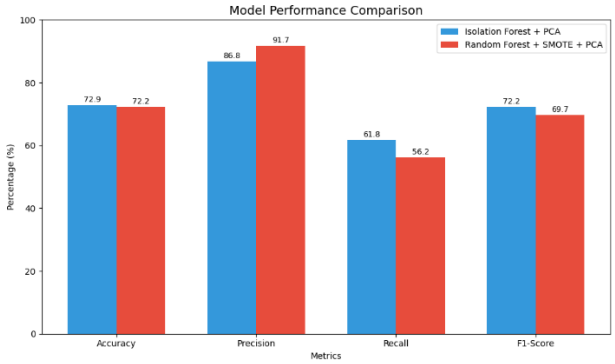

Figure 4: Model Performance Comparison Chart

The experimental results validate the proposed framework across three key dimensions. First, PCA successfully reduces feature dimensions from 38 to 20 while preserving 91.97 percent of variance. This reduction enables inference times of 12.4 milliseconds per instance. Second, the Isolation Forest model achieves competitive detection performance with 72.87 percent accuracy and 86.75 percent attack class precision. The confusion matrix confirms 7,927 true positives against only 1,211 false positives. Third, the comparative analysis shows that the proposed model outperforms the Random Forest with SMOTE baseline in accuracy, attack class recall, and attack class F1-Score. The proposed model detects 713 more actual attacks than the baseline model. These results collectively support the claim that the PCA-Isolation Forest framework is effective for unsupervised anomaly detection in cloud network environments.

Class overlap in Figures 2 and 3 occurs because PCA retains maximum variance rather than class separation. Some normal and attack traffic share similar statistical characteristics after projection. This overlap increases false negatives as anomalies close to normal clusters are not isolated. For real cloud deployment, this implies that unsupervised detection is effective for clear anomalies but may miss subtle attacks. The results indicate that additional contextual or temporal features may improve detection of low intensity threats.

5. Discussion

This study directly addresses the identified research gap by demonstrating how optimized PCA-based dimensionality reduction enhances the effectiveness of Isolation Forest in cloud intrusion detection. Compared to prior studies that applied PCA or Isolation Forest independently, the proposed integrated approach shows improved balance between precision and computational efficiency. The methodology follows a two-stage pipeline in which numerical features of the NSL KDD dataset are standardized and reduced to twenty principal components that retain 91.97 percent variance. This dimensionality reduction makes the calculations cheaper and reduces redundant information that enhances the separation of anomalies as observed in the current cloud intrusion research (Fraihat et al., 2026); (Bakro et al., 2023).

The effectiveness of dimensionality reduction for anomaly detection is further supported by recent hybrid deep learning approaches in resource-constrained environments (Yan et al., 2025). Isolation Forest is used to process the reduced features, with contamination being 0.25 to indicate attack proportion. detection strategy benefits from reduced feature noise introduced by PCA. By removing redundant correlations, PCA improves compactness of normal traffic clusters. This compactness helps Isolation Forest identify sparse anomaly regions more effectively. However, unsupervised detection does not explicitly learn class boundaries. As a result, overlapping traffic patterns remain challenging. Dynamic cloud traffic with evolving behavior may further reduce separability. These observations highlight the tradeoff between lightweight unsupervised detection and comprehensive attack coverage (Abdulkareem et al., 2024); (Al-Hawawreh, 2022).

Recent unsupervised intrusion detection studies report accuracy values between 68 percent and 75 percent on NSL KDD and similar datasets. For example, PCA based anomaly detection models report F1 scores around 70 percent, while hybrid deep learning approaches achieve higher recall at increased computational cost. The proposed method achieves 72.87 percent accuracy with low inference time of 12.4 milliseconds. Compared to computationally intensive deep models, the framework provides competitive performance with reduced complexity.

This supports the practical relevance of lightweight anomaly detection. The obtained results indicate that the accuracy with the precision of 86.75 percent is 72.87 percent in detecting attacks and the recall is 61.77 percent. Good precision implies that there are good alarms that are vital in functional cloud monitoring, since too many alerts decrease the efficiency of the analysts (Abdallah et al., 2024); (Chen et al., 2023). PCA visualization indicates that there is partial overlap between the classes, hence the moderate recall and the rationale of using tree-based isolation rather than linear classifiers. The confusion matrix also supports the fact that there are 7927 true positives and 1211 false positives, meaning that it has a good discrimination ability. Relative comparison of the proposed model with Random Forest and SMOTE demonstrates that the proposed model is more accurate by 0.69 percent and recalls more attacks by 5.56 percent (identifying 713 more attacks). The proposed framework offers greater balance as indicated by the higher attack class F1 score and macro F1 score, though the baseline is

more precise. These results endorse the idea that dimensionality reduction can be effective in upgrading the performance of Isolation Forest by eliminating noise and maintaining any important variance (Chen et al., 2023); (Saheed & Misra, 2024). The inference time of 12.4 milliseconds shows that it is suitable for real-time implementation in cloud controllers. The performance in general is comparable to lightweight anomaly detection systems that have been reported in recent studies that prioritize efficiency and unsupervised learning on dynamic environments (Ahmed et al., 2023);(Chapagain et al., 2022);(Umar & Jordanov, 2026).

The findings confirm the usefulness of the combination of PCA and Isolation Forest as an effective trade-off between the ability to detect and the computational cost of cloud network intrusion detection, and justify its effective implementation in resource-limited cloud systems. To further clarify the contribution, three research questions are addressed. PCA based dimensionality reduction reduces feature dimensionality from 38 to 20 components while retaining 91.97 percent variance. This improves computational efficiency and reduces noise, supporting anomaly isolation. Isolation Forest achieves 86.75 percent attack precision indicating strong anomaly isolation capability, while moderate recall of 61.77 percent reflects limitations in detecting subtle attacks. Compared with the supervised baseline, the unsupervised model provides higher precision and lower false alarms but lower recall. This demonstrates the tradeoff between supervised learning and anomaly-based detection. These findings are consistent with recent studies that emphasize the importance of balancing computational efficiency and detection performance in intelligent intrusion detection systems (Yan et al., 2025); (Chen et al., 2023).

Conclusion

This study evaluates a lightweight PCA and Isolation Forest based anomaly detection approach using the NSL KDD dataset. The results demonstrate competitive detection performance with high precision and moderate recall. The framework reduces dimensionality and computational cost while maintaining anomaly detection capability. The findings indicate potential suitability for offline analysis and controlled monitoring environments.

Relative comparison reveals that the framework is more accurate and recalls more than the Random Forest baseline with competitive F1 scores. A time of inference of 12.4 milliseconds proves its appropriateness in real time cloud deployment. These results confirm that the combination of PCA and Isolation Forest provides a balance between detection performance and operational efficiency. The proposed framework is therefore a viable solution for cloud intrusion detection in environments where labeled data is limited and computational resources are constrained. Theoretical implications of this study include extending the understanding of how dimensionality reduction interacts with isolation-based anomaly detection. The results provide the indication that the suggested framework can be applied to lightweight monitoring of anomalies in controlled clouds. More testing on real-world and streaming cloud traffic is required before it can be put into practice. Future research should explore adaptive contamination tuning, streaming data environments, and validation on modern datasets such as CIC-IDS2017 and UNSW-NB15.

7. Limitations Of the Study

This study has several limitations. First, the NSL KDD dataset is relatively old and may not fully represent modern cloud traffic patterns. Second, experiments are conducted in offline settings without streaming or software defined networking validation. Third, categorical features are excluded which may reduce contextual information. Fourth, evaluation is limited to a single benchmark dataset. Future work includes testing on modern datasets such as CIC IDS and UNSW NB15 and validation in dynamic cloud environments

REFERENCES

1. Abdallah, A. M., Alkaabi, A. S. R. O., Alameri, G. B. N. D., Rafique, S. H., Musa, N. S., & Murugan, T. (2024). Cloud network anomaly detection using machine and deep learning techniques. IEEE Access, 12, 56749–56773. https://doi.org/10.1109/ACCESS.2024.3390844
2. Abdulkareem, S. A., Foh, C. H., Shojafar, M., Carrez, F., & Moessner, K. (2024). Network intrusion detection: An IoT and non-IoT-related survey. IEEE Access, 12, 147167–147191. https://doi.org/10.1109/ACCESS.2024.3473289
3. Ahmed, I., Anisetti, M., Ahmad, A., & Jeon, G. (2023). A multilayer deep learning approach for malware classification in 5G-enabled IIoT. IEEE Transactions on Industrial Informatics, 19(2), 1495–1503. https://doi.org/10.1109/TII.2022.3205366
4. Al-Hawawreh, M. (2022). Developing an effective detection framework for targeted ransomware attacks in brownfield industrial internet of things. Doctoral dissertation, University of New South Wales.
5. Bakro, M., Kumar, R. R., Alabrah, A. A., Ashraf, Z., Bisoy, S. K., Parveen, N., Khawatmi, S., & Abdelsalam, A. (2023). Efficient intrusion detection system in the cloud using fusion feature selection approaches and an ensemble classifier. Electronics, 12(11), 2427. https://doi.org/10.3390/electronics12112427
6. Chapagain, P., Timalsina, A., Bhandari, M., & Chitrakar, R. (2022). Intrusion detection based on PCA with improved K-means. In Innovations in Electrical and Electronic Engineering (pp. 13–27). Springer. https://doi.org/10.1007/978-981-19-1677-9_2
7. Chen, J., Zhang, J., Qian, R., Yuan, J., & Ren, Y. (2023). An anomaly detection method for wireless sensor networks based on improved isolation forest. Applied Sciences, 13(2), 702. https://doi.org/10.3390/app13020702
8. Fraihat, S., Yaseen, Q., Sanjalawe, Y., Abu-Errub, A., Makhadmeh, S. N., & Al-Betar, M. A. (2026). Intrusion detection in industrial internet of things network using feature optimization and hybrid deep

learning. Discover Internet of Things, 6(1). https://doi.org/10.1007/s43926-026-00284-z

9. Kakavand, M., Mustapha, N., Mustapha, A., & Abdullah, M. T. (2016). Effective dimensionality reduction of payload-based anomaly detection in TMAD model for HTTP payload. KSII Transactions on Internet and Information Systems, 10(8), 3884–3910.
10. Saheed, Y. K., & Misra, S. (2024). Voting gray wolf optimizer-based ensemble learning models for intrusion detection in the Internet of Things. International Journal of Information Security, 23(3), 1557–1581. https://doi.org/10.1007/s10207-023-00803-x
11. Umar, A., & Jordanov, I. (2026). Anomaly detection in IoT environment using machine learning: A survey. IEEE Internet of Things Journal. https://doi.org/10.1109/JIOT.2026.3671102
12. Yan, Z., Shukla, P. K., Shukla, P. K., Thakur, K., Sinha, A., & Khalid, S. (2025). Intrusion detection and mitigation method for the industrial internet of things using bidirectional convolutional long short-term memory and deep recurrent convolutional Q-networks. International Journal of Computational Intelligence Systems, 18(1). https://doi.org/10.1007/s44196-025-00890-9
13. Paolini, D., Dini, P., Soldaini, E., & Saponara, S. (2025). One-class anomaly detection for industrial applications: A comparative survey. Computers, 14(7), 281. https://doi.org/10.3390/computers14070281
14. Noor, K., Imoize, A. L., Li, C. T., & Weng, C. Y. (2025). Machine learning and transfer learning strategies for intrusion detection systems in 5G and beyond. Mathematics, 13(7), 1088. https://doi.org/10.3390/math13071088
15. Kadom, S. A., Hashem, S. H., & Jafer, S. H. (2022). Optimize network intrusion detection system based on PCA feature extraction and three naïve Bayes classifiers. Journal of Physics: Conference Series, 2322, 012092. https://doi.org/10.1088/1742-6596/2322/1/012092
16. Abbas, Q., Hina, S., Sajjad, H., Zaidi, K. S., & Akbar, R. (2023). Optimization of predictive performance of intrusion detection system using hybrid ensemble model. PeerJ Computer Science, 9, e1552. https://doi.org/10.7717/peerj-cs.1552
17. Ahmed, A., Asim, M., Ullah, I., & Ateya, A. A. (2024). Optimized ensemble model with advanced feature selection for network intrusion detection. PeerJ Computer Science, 10, e2472. https://doi.org/10.7717/peerj-cs.2472
18. Yang, Z., Liu, Z., Zong, X., & Wang, G. (2023). Optimized adaptive ensemble model with feature selection for network intrusion detection. Concurrency and Computation: Practice and Experience, 35(4), e7529. https://doi.org/10.1002/cpe.7529
19. Raghunath, M. P., Deshmukh, S., Chaudhari, P., Bangare, S. L., Kasat, K., Awasthy, M., & Waghulde, R. R. (2025). PCA and PSO-based optimized support vector machine for intrusion detection in IoT. Measurement: Sensors, 37, 101806. https://doi.org/10.1016/j.measen.2025.101806
20. Kumar, S. V. N. (2025). Enhanced whale optimizer-based feature selection for intrusion detection system. Peer-to-Peer Networking and Applications, 18(2). https://doi.org/10.1007/s12083-024-01565-2
21. Quoc, T. N., Phan, M. V., Dang, K. P., Gia, H. N. H., Thien, P. N. T., Hoang, V. T., & Nguyen, T. D. (2026). Smartphone-based colorimetric sensing with reference calibration and ensemble machine learning. Analytica Chimica Acta. https://doi.org/10.1016/j.aca.2026.345097
22. Laskar, M. T. R., Huang, J. X., Smetana, V., Stewart, C., Pouw, K., An, A., ... & Liu, L. (2021). Extending isolation forest for anomaly detection in big data via K-means. ACM Transactions on Cyber-Physical Systems (TCPS), 5(4), 1-26. https://doi.org/10.1145/3460976
23. Xu, H., Pang, G., Wang, Y., & Wang, Y. (2023). Deep isolation forest for anomaly detection. IEEE transactions on knowledge and data engineering, 35(12), 12591-12604. doi: 10.1109/TKDE.2023.3270293.
24. Jamshidi, S., Erfan, F., Abdul-Wahab, O., Bellaiche, M., & Khomh, F. (2025). Lightweight Autoencoder-Isolation Forest Anomaly Detection for Green IoT Edge Gateways. arXiv preprint arXiv:2511.18235. https://doi.org/10.48550/arXiv.2511.18235.
25. Fadul, A. M. A. (2023). Anomaly detection based on isolation Forest and local outlier factor. Africa University.
26. Heigl, M., Anand, K. A., Urmann, A., Fiala, D., Schramm, M., & Hable, R. (2021). On the improvement of the isolation forest algorithm for outlier detection with streaming data. Electronics, 10(13), 1534. https://doi.org/10.3390/electronics10131534
27. AbuAlghanam, O., Alazzam, H., Alhenawi, E. A., Qatawneh, M., & Adwan, O. (2023). Fusion-based anomaly detection system using modified isolation forest for internet of things. Journal of Ambient Intelligence and Humanized Computing, 14(1), 131-145. https://doi.org/10.1007/s12652-022-04393-9
28. Maiti, A., Chakraborty, R., Basu, D., Sarkar, I., & Dutta, A. (2025, July). Unsupervised Pattern Discovery in Cyber Incidents Using Principal Component Analysis K-Means DBSCAN and Isolation Forest. In International Conference on Data Science and Network Engineering (pp. 314-324). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-07735-6_27
29. Chaitanya, V. L., Devi, M. S., Sree, G. A., Thabasum, D. A., Sravani, U., & Sneha, G. Outlier Detection for IoT Frameworks Using Isolation Forest. Proceedings

Copyright, 808, 815. DOI: 10.5220/0013890300004919

30. Maneesh, P., Sudhakar, R., Naresh, P., & Lokesh, K. (2026). Forest-based anomaly detection and isolation for cyber security applications. In Recent Advances in Computational Methods in Science and Technology (pp. 285-291). CRC Press.
31. Suman, S., & Mishra, J. K. (2025). Isolation forest enabled anomaly detection caused by internal and external disturbances in IsOWC system. Physica Scripta, 100(10), 106007. DOI 10.1088/1402-4896/ae11d
32. NÄTTERDAL, F., & OLAUSSON, K. (2024). A Study on Isolation Forest for Anomaly Detection in Cloud-Based Systems.
33. Bachar, M., Khiat, A., & El Guemmat, K. (2026). Hybrid Autoencoder and Isolation Forest for IoT Anomaly Detection with a Novel Model. Engineering, Technology & Applied Science Research, 16(1), 31123-31129. https://doi.org/10.48084/etasr.15288.

..

.