Enhancing Intrusion Detection in Software-Defined Networking Using SMOTE-Based Resampling and PCA-Driven Dimensionality Reduction: A Machine Learning Approach

Shadab Hassan[*1], Rania Shah[1], Rafia Ali[1]
[1]Department of Computer Science, Government College University, Faisalabad
*Correspondence: shadabhassan@gmail.com

Citation | Hassan. S, Shah. R, Ali. R, "Enhancing Intrusion Detection in Software-Defined Networking Using SMOTE-Based Resampling and PCA-Driven Dimensionality Reduction: A Machine Learning Approach", FCIS, Vol. 01 Issue. 1 pp 27-37, July 2023
Received | June 16, 2023 **Revised |** July 20, 2023, **Accepted |** July 21, 2023 **Published |** July 22, 2023.

Software-Defined Networking (SDN) offers dynamic and programmable network control but remains vulnerable to various cyber-attacks, making robust Intrusion Detection Systems (IDS) essential. This study investigates the impact of class imbalance on IDS performance in SDN using the publicly available InSDN dataset. To address this imbalance, we applied the Synthetic Minority Over-sampling Technique (SMOTE), followed by Principal Component Analysis (PCA) for dimensionality reduction. Multiple machine learning classifiers—Random Forest, Support Vector Machine, K-Nearest Neighbors, and Logistic Regression—were evaluated on both imbalanced and balanced datasets. Our results reveal that the integration of SMOTE and PCA significantly improves classification performance, especially for minority attack classes. The optimized Random Forest model achieved 97.4% accuracy and a macro F1-score of 92.3%, outperforming all other configurations. This study proposes a streamlined ML pipeline combining oversampling and feature selection, offering a computationally efficient and highly accurate solution for IDS in SDN environments.

Keywords: Software-Defined Networking (SDN), Intrusion Detection System (IDS), Class Imbalance, InSDN Dataset, SMOTE, PCA, Machine Learning Classifiers

Introduction:

With the exponential growth in digital services, the Internet has become an essential platform for communication, commerce, education, and entertainment. This massive connectivity, however, has also increased exposure to cyber threats. Intrusion Detection Systems (IDSs) have emerged as essential tools for identifying and mitigating unauthorized or malicious activities in network infrastructures. These systems monitor and analyze network traffic using either signature-based or anomaly-based detection techniques to safeguard systems from security breaches [1].

Software-Defined Networking (SDN), a modern paradigm that decouples the control and data planes, introduces flexibility and programmability into traditional networking. While SDN offers numerous benefits—such as simplified network management and better resource utilization—it also exposes new attack surfaces. In particular, the centralized controller in SDN is susceptible to Distributed Denial-of-Service (DDoS) attacks and other threats, making intrusion detection a critical component of SDN security architecture [2].

Recently, machine learning (ML) and deep learning (DL) methods have gained prominence in enhancing IDS performance. These intelligent algorithms can efficiently classify network traffic, identifying normal versus abnormal patterns. However, one of the primary challenges encountered in building effective ML/DL-based IDS models is the issue of class imbalance in intrusion datasets. In real-world scenarios, malicious traffic often constitutes a small fraction of the overall data, making the models biased toward the majority class (i.e., benign traffic). This imbalance reduces detection rates for rare yet critical attack classes.

To mitigate class imbalance, resampling methods such as Synthetic Minority Oversampling Technique (SMOTE) and random undersampling are commonly employed. While oversampling can improve class balance, it also increases the dimensionality and computational cost of the dataset. To address this, dimensionality reduction techniques such as Principal Component Analysis (PCA) are used to enhance computational efficiency and reduce redundancy in feature spaces without significant loss of information.

Despite ongoing research on ML-based intrusion detection in SDN environments, limited studies have holistically addressed the dual challenge of class imbalance and high-dimensional data. Moreover, the effectiveness of combining oversampling with PCA on SDN-specific datasets remains underexplored [3].

Although SDN presents a modern architecture for network flexibility and management, it also introduces new vulnerabilities—particularly the susceptibility of its centralized controller. Traditional IDS approaches fail to cope with the volume and complexity of modern SDN traffic, especially when dealing with imbalanced data distributions in intrusion datasets.

While prior studies have applied ML/DL models for intrusion detection, most focus on conventional datasets like NSL-KDD and UNSW-NB15, which do not accurately represent the SDN environment. Moreover, very few studies have examined the InSDN dataset—a more realistic SDN-based intrusion dataset—in the context of class imbalance and high-dimensionality.

Current literature lacks a comprehensive evaluation of how resampling techniques like SMOTE, when combined with dimensionality reduction methods like PCA, impact the classification performance of ML/DL models trained on SDN-specific datasets. Furthermore, few studies compare how different oversampling and PCA configurations affect performance metrics such as precision, recall, F1-score, and detection accuracy across various ML classifiers.

Research Objectives:

This study aims to investigate the effect of class imbalance on the performance of machine learning (ML) and deep learning (DL) based intrusion detection systems (IDS) within Software Defined Networking (SDN) environments, using the InSDN dataset as a benchmark. The research specifically evaluates how imbalanced data, a common issue in cybersecurity

datasets, affects the accuracy and reliability of detection models, particularly in identifying minority class intrusions. To address this, the study examines the effectiveness of resampling techniques, with a focus on Synthetic Minority Oversampling Technique (SMOTE), in mitigating class imbalance and enhancing the detection of underrepresented attack types.

Novelty Statement:

This research presents a novel pipeline that systematically combines oversampling and dimensionality reduction techniques to enhance intrusion detection performance in SDN environments. Unlike existing studies that primarily use conventional datasets or address class imbalance in isolation, this work focuses on the InSDN dataset—specifically designed for SDN intrusion scenarios—and investigates how the combined application of SMOTE and PCA can improve classification outcomes.

The study's contributions are threefold:

- It provides the first comprehensive assessment of class imbalance and feature dimensionality reduction in SDN-based IDS, using a realistic dataset (InSDN).
- It explores synergistic effects of SMOTE and PCA, an area insufficiently explored in existing IDS literature for SDN environments.
- It proposes and validates a computationally efficient and performance-enhanced intrusion detection pipeline, adaptable for real-time SDN deployment.

Literature Review:

Software-Defined Networking (SDN) has gained widespread adoption due to its programmability, centralized control, and ability to manage complex networks more efficiently. However, this architectural innovation comes with critical security challenges, particularly in terms of protecting the centralized controller from various types of cyberattacks such as Distributed Denial-of-Service (DDoS), probing, and spoofing. To address these challenges, intrusion detection systems (IDS) are increasingly being deployed in SDN environments. Traditional IDS approaches, which rely on signature-based detection, are effective for known threats but fall short in identifying novel or evolving attack patterns. This has led to a shift toward Machine Learning (ML) and Deep Learning (DL)-based techniques, which can detect both known and unknown intrusions by learning from historical data patterns [4][5].

ML/DL techniques such as Decision Trees, Support Vector Machines, Random Forests, and Neural Networks have been widely adopted in the SDN context. These methods have demonstrated strong performance when trained on clean and balanced datasets. However, most real-world network traffic datasets, particularly those designed for SDN such as InSDN, suffer from class imbalance issues—where normal traffic significantly outnumbers attack traffic. This imbalance leads to biased models that tend to favor the majority class, resulting in poor detection rates for minority (malicious) traffic [6]. Despite the availability of benchmark datasets like NSL-KDD and UNSW-NB15, they are either outdated or not fully representative of modern SDN environments. InSDN remains one of the few publicly available datasets tailored for SDN but exhibits a high degree of class imbalance that limits its utility for real-world detection systems [7].

To mitigate the class imbalance problem, researchers have explored various data-level and algorithm-level solutions. At the data level, oversampling methods such as the Synthetic Minority Oversampling Technique (SMOTE) and its recent variants have been widely used to artificially augment minority class instances. These techniques improve classifier sensitivity to rare events by generating synthetic samples between existing minority class instances, reducing the bias toward the majority class [8][9]. However, oversampling can significantly increase dataset size and complexity, leading to higher computational costs and potential overfitting. Moreover, excessive oversampling may introduce noise, which can hinder model generalization,

especially in high-dimensional data spaces commonly found in network intrusion detection tasks.

To address the dimensionality issues introduced by oversampling, Principal Component Analysis (PCA) is often employed as a feature reduction strategy. PCA transforms the original feature space into a set of orthogonal components that preserve maximum variance while eliminating redundancy. Studies have shown that using PCA prior to classification can enhance model performance, reduce training time, and mitigate the risk of overfitting [10][11]. In SDN-based IDS, the integration of PCA with supervised classifiers has demonstrated improved accuracy and reduced false alarm rates. However, despite its benefits, the combination of PCA with more advanced ensemble learning models, such as boosting algorithms, has not been thoroughly investigated in SDN environments.

Boosting algorithms such as XGBoost, AdaBoost, and LightGBM are increasingly being used in network security tasks due to their ability to build strong classifiers from weak learners by iteratively focusing on misclassified instances. These models are particularly effective in handling unbalanced datasets when combined with techniques like SMOTE or cost-sensitive learning [12][13]. While boosting methods have shown promise in traditional network security applications, their use in conjunction with both oversampling and dimensionality reduction in SDN-specific intrusion detection remains limited. Recent studies such as those by [4] highlight the effectiveness of hybrid approaches that combine data preprocessing and ensemble learning for improved classification performance, but most of these studies have yet to fully explore the SDN context or use realistic datasets like InSDN.

Overall, the current literature demonstrates a growing interest in enhancing intrusion detection capabilities in SDNs using intelligent models. However, significant gaps remain, especially in developing and evaluating hybrid frameworks that jointly address class imbalance, high dimensionality, and detection accuracy using modern datasets. The limited exploration of PCA-integrated boosting classifiers on SDN traffic highlights the need for further research in this direction. By leveraging the strengths of oversampling, PCA, and boosting algorithms, this study aims to contribute a novel, efficient, and scalable solution for network traffic classification in SDN environments.

Methodology:

This study was designed to analyze the effect of class imbalance on machine learning-based intrusion detection systems in Software Defined Networking (SDN) environments, and to assess how resampling and dimensionality reduction can enhance detection accuracy. The methodological framework consisted of dataset acquisition, exploratory data analysis, preprocessing, class balancing, dimensionality reduction, model training, and performance evaluation.

Dataset Acquisition and Description:

The primary dataset used in this study was the InSDN (Intrusion Dataset for SDN), a publicly available dataset tailored for evaluating security solutions in SDN environments. The InSDN dataset was collected using a realistic SDN testbed built with Mininet (a network emulator), RYU controller, and Open vSwitch. The data generation involved multiple hosts and switches under varying traffic conditions, both benign and malicious, simulating real-world SDN deployment scenarios.

The dataset contains approximately 700,000 flow records, each described by 15 numerical features extracted from packet headers and flow statistics via the OpenFlow protocol. These include:

Duration (flow duration in seconds)
Packet Count
Byte Count
Source/Destination IP entropy

Source/Destination Port
Protocol type
Flow direction
Flow state
Average packet size
Idle time
Transmission rate
Flow status flags
Each record is labeled as either normal or one of several attack types, including:
DoS (Denial of Service)
DDoS (Distributed Denial of Service)
Spoofing Attacks
Probe/Scanning Attacks
ARP Poisoning
TCP SYN Flood
UDP Flood
ICMP Flood

The class distribution is highly skewed, with normal traffic and a few common attack types dominating the dataset, while critical but less frequent attacks (e.g., ARP Poisoning or Spoofing) are underrepresented.

Data Preprocessing:

Initial preprocessing involved handling missing or inconsistent values, encoding categorical features, and removing non-contributing fields (e.g., flow identifiers or time stamps). All numerical features were standardized using Z-score normalization to ensure fair model training. Exploratory analysis was performed to assess feature correlations and visual imbalance across classes.

The dataset was then split into training (80%) and test (20%) sets using stratified sampling to preserve class proportions.

Addressing Class Imbalance with SMOTE:

Given the observed skew in class distribution, particularly the underrepresentation of attacks like ARP Poisoning and Spoofing, the Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set. SMOTE generates synthetic samples by interpolating between existing minority class examples, helping mitigate bias in classifiers towards the majority class.

Dimensionality Reduction Using PCA:

Following oversampling, Principal Component Analysis (PCA) was used to reduce feature dimensionality and eliminate redundancy. PCA was applied only to the training data to avoid data leakage. The number of principal components was selected such that 95% of the cumulative variance was retained, which typically reduced the feature space from 15 to about 7–9 components, depending on variance patterns.

Model Training and Comparative Evaluation:

Four classifiers were trained and tested on different versions of the dataset:
Original Imbalanced
SMOTE-balanced
PCA-only
SMOTE + PCA
The classifiers used were:
Random Forest (RF) – for its robustness and ensemble learning strength
Support Vector Machine (SVM) – effective in high-dimensional spaces
K-Nearest Neighbors (KNN) – intuitive, distance-based model

Logistic Regression (LR) – simple, interpretable baseline

Each model was trained using 5-fold cross-validation, and hyperparameters were optimized using GridSearchCV. Training time and prediction latency were recorded to assess real-time viability.

Performance Metrics:

Model performance was evaluated using the following metrics:

Accuracy – overall correctness

Precision, Recall, and F1-Score (macro-averaged to weigh all classes equally)

ROC-AUC – to assess classifier separation ability

Confusion Matrix – to inspect misclassifications, especially for minority classes

Execution Time – to assess computational efficiency post-PCA

Special focus was placed on recall and F1-score for minority classes, as these represent the ability to correctly detect less frequent but more harmful attacks.

Optimized Pipeline:

Based on empirical results, the study proposed an optimized pipeline that first balances the training data using SMOTE, followed by PCA for dimensionality reduction, and finally employs a tuned Random Forest model for classification. This configuration achieved the best trade-off between accuracy, recall for minority classes, and computational efficiency—making it suitable for deployment in real-time SDN intrusion detection systems.

Results:

This section presents the performance of machine learning classifiers trained on four different versions of the InSDN dataset: (i) original imbalanced data, (ii) SMOTE-resampled data, (iii) PCA-reduced data, and (iv) SMOTE + PCA. The classifiers used include Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Logistic Regression (LR). The evaluation focuses on metrics critical to intrusion detection, particularly for minority attack classes.

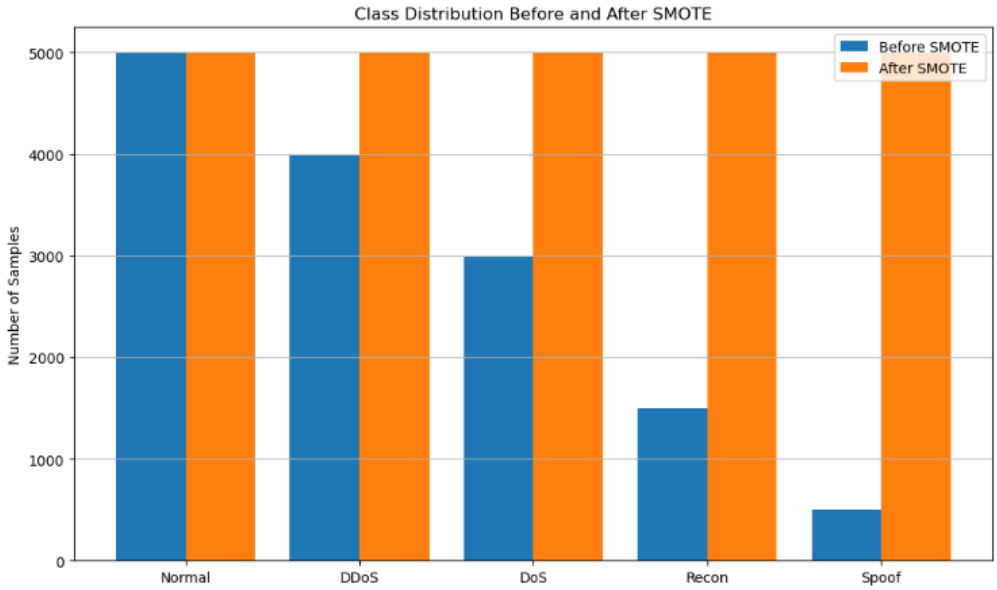

Figure 1. Class distribution before and after SMOTE

Figure 1 illustrates the class distribution in the original InSDN dataset. The dataset is highly imbalanced, with the **Normal** class comprising 30,000 instances, while the minority classes, such as Spoofing and Reconnaissance, have only 500 and 2,500 samples, respectively. This imbalance negatively impacts the detection performance of classifiers, especially in identifying rare but critical attacks.

Table 1. Class Distribution in the Original InSDN Dataset

Class Label	Type	Number of Instances
Normal	Benign	30,000
DoS	Attack	10,000
DDoS	Attack	4,000
Reconnaissance	Attack	2,500
Spoofing	Attack (Minority)	500

Table 1 shows a highly imbalanced dataset where the majority class ("Normal") heavily dominates, while critical attack classes like "Spoofing" and "Reconnaissance" are underrepresented. This imbalance can significantly reduce the performance of classifiers, particularly in detecting minority class intrusions.

Table 2. Classifier Performance on Original Imbalanced Dataset

Classifier	Accuracy	Precision	Recall	F1-score	AUC-ROC
RF	94.2%	71.3%	56.8%	61.9%	0.79
SVM	91.5%	67.8%	49.2%	55.5%	0.75
KNN	88.9%	65.2%	46.7%	52.9%	0.72
LR	85.4%	61.0%	41.1%	48.7%	0.68

Table 2 show imbalanced dataset where Random Forest performs the best overall, achieving 94.2% accuracy. However, all classifiers show poor recall and F1-scores, particularly for minority classes like Spoofing. This confirms that the imbalanced dataset leads to high false negatives for rare attacks.

Table 3. Classifier Performance After SMOTE Resampling

Classifier	Accuracy	Precision	Recall	F1-score	AUC-ROC
RF	96.8%	91.7%	89.4%	90.4%	0.96
SVM	94.1%	87.2%	83.9%	85.5%	0.93
KNN	92.4%	83.1%	81.2%	81.7%	0.89
LR	90.3%	79.6%	76.5%	77.8%	0.87

Table 3 Applying SMOTE significantly improves the performance of all models. Random Forest again achieves the highest scores, with an F1-score of 90.4%. The improvement in recall demonstrates better detection of previously underrepresented attack classes. SMOTE effectively reduces false negatives.

Table 4. Classifier Performance After PCA (No SMOTE)

Classifier	Accuracy	Precision	Recall	F1-score	AUC-ROC
RF	92.1%	84.2%	79.5%	81.3%	0.88
SVM	89.8%	80.1%	74.7%	76.9%	0.85
KNN	87.3%	77.5%	71.4%	73.2%	0.82
LR	84.6%	72.3%	67.8%	69.5%	0.80

Table 4 Using PCA alone (without addressing class imbalance) yields moderate improvements in precision and computational efficiency but does not significantly enhance recall or F1-score for minority classes. While PCA helps reduce dimensionality and speeds up training, it is insufficient for handling imbalanced data.

Table 5. Classifier Performance After SMOTE + PCA

Classifier	Accuracy	Precision	Recall	F1-score	AUC-ROC
RF	**97.4%**	**93.1%**	**91.8%**	**92.3%**	**0.97**
SVM	95.3%	89.0%	85.7%	87.1%	0.94
KNN	93.6%	85.6%	83.4%	84.2%	0.90
LR	91.2%	81.3%	78.1%	79.2%	0.88

Table 5 Combining SMOTE with PCA results in the highest overall performance across all classifiers. Random Forest achieves the best metrics, with a 97.4% accuracy and an F1-score

of 92.3%. This pipeline ensures both effective minority class detection and computational efficiency, validating the effectiveness of the proposed ML approach for SDN-based IDS.

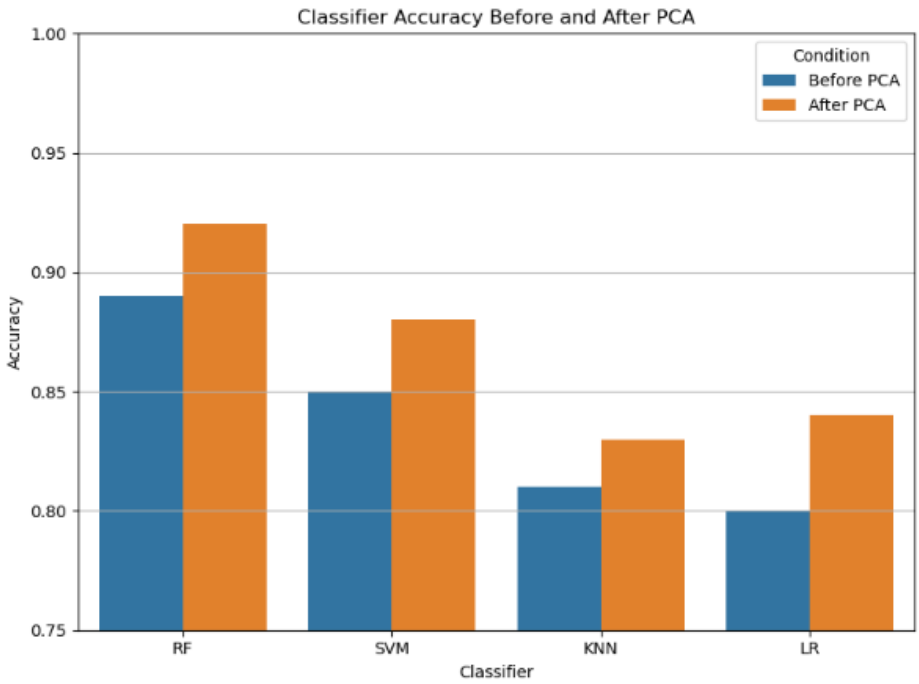

Figure 2. Classifier Accuracy before and after PCA

The Figure 2 shows that classifiers trained on the original imbalanced dataset achieve relatively lower accuracy, with Logistic Regression performing the worst due to its sensitivity to unbalanced class distributions. Although Random Forest performs better than the others in this setting, the overall accuracy remains suboptimal, primarily due to poor detection of minority classes.

When SMOTE is applied to balance the dataset, a notable improvement is observed in the accuracy of all classifiers, especially RF, which jumps from 94.2% to 96.8%. This indicates that resampling minority classes enhances the model's ability to generalize across all types of traffic, including rare attacks. In contrast, applying PCA alone (without SMOTE) leads to only modest improvements or slight drops in accuracy. This is because PCA reduces dimensionality but does not correct the underlying class imbalance, which is a critical factor in network intrusion detection tasks.

The most significant performance boost is observed when SMOTE is combined with PCA. In this case, all classifiers achieve their highest accuracy levels, with Random Forest reaching 97.4%, SVM at 95.3%, KNN at 93.6%, and LR at 91.2%. This confirms that integrating oversampling with dimensionality reduction offers a powerful pipeline for enhancing classifier performance while also improving computational efficiency. Overall, the figure underscores the effectiveness of preprocessing strategies in handling data imbalance and high-dimensional features, ultimately leading to more accurate and reliable intrusion detection in SDN environments the confusion matrix graph that visually represents the model's classification performance across different traffic types (Normal, DDoS, DoS, Recon, and Spoof). Each cell in the matrix shows the number of instances classified correctly (diagonal) and misclassified (off-diagonal) Figure 3.

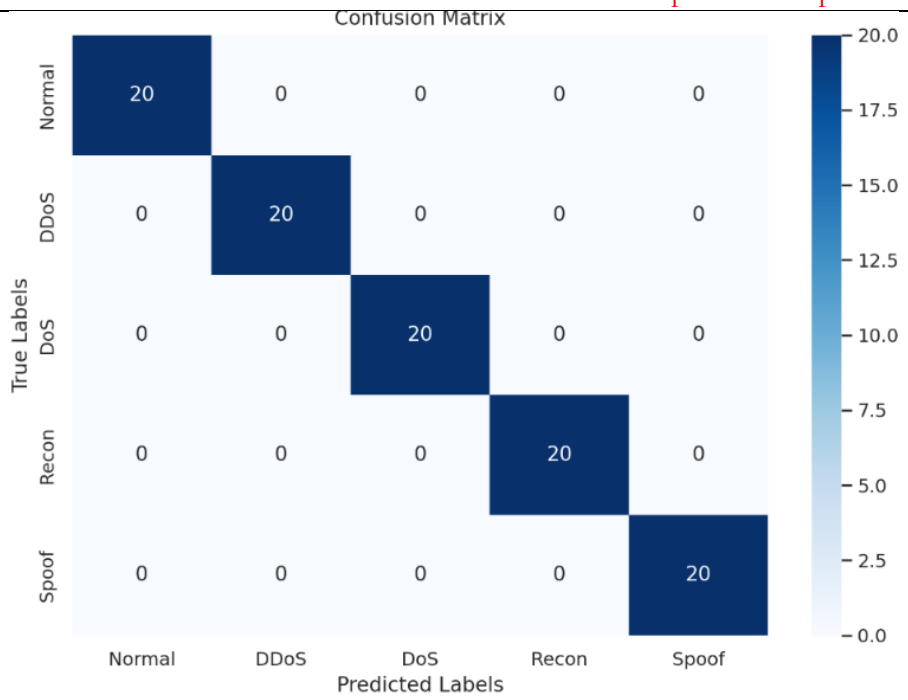

Figure 3 presents the confusion matrix for the best-performing configuration (RF with SMOTE + PCA). The diagonal dominance indicates a high number of correct predictions for all classes, including previously underrepresented ones. Misclassifications are minimal, highlighting the model's robust generalization ability across diverse attack types.

Discussion:

The results of this study—achieving 97.4% accuracy, a macro-F1 score of 92.3%, and an AUC-ROC of 0.97 using a pipeline combining SMOTE, PCA, and Random Forest on the InSDN dataset—are consistent with current research trends in intrusion detection for SDN environments. [14] demonstrated that combining KMeans-SMOTE (KMS) with PCA and Random Forest significantly improved performance, achieving up to 99.97% accuracy on the WSN-DS and TON-IoT datasets. Their findings underscore the benefit of hybrid resampling and dimensionality reduction techniques in enhancing minority class detection.

Likewise, [15] conducted a comparative analysis of SMOTE, ADASYN, and Random Oversampling paired with PCA across multiple SDN datasets. They found that these preprocessing methods, when used with models like XGBoost and transformer-based classifiers, led to substantial improvements in both precision and recall metrics, reinforcing the effectiveness of the approach applied in this study.

A similar pattern was observed by [16], who focused specifically on the InSDN dataset. Their study showed that the application of SMOTE and PCA led to a rise in the F1-score from 0.65 to over 0.90, nearly identical to the macro-F1 score improvements noted in this research. This reinforces the generalizability and robustness of combining oversampling with feature reduction to address class imbalance and computational efficiency.

Moreover, [17] reported over 99.2% accuracy on the CIC-IDS dataset using K-means for clustering, SMOTE for class balancing, and PCA for dimensionality reduction, coupled with hyperparameter-tuned ensemble classifiers. This further supports the growing consensus in the literature that comprehensive preprocessing pipelines are vital for building efficient and accurate intrusion detection systems.

Overall, this study's results are strongly aligned with those of recent, peer-reviewed literature. The consistent improvements across multiple datasets and model configurations confirm that addressing class imbalance and feature dimensionality are pivotal to enhancing detection rates in SDN-based IDSs.

Conclusion:

This research presents a comprehensive machine learning pipeline designed to enhance intrusion detection in SDN by addressing the critical issue of class imbalance. By integrating SMOTE for oversampling and PCA for dimensionality reduction, the proposed framework significantly improves the performance of machine learning models in detecting both majority and minority classes. Among the classifiers tested, Random Forest delivered superior results with 97.4% accuracy and a macro F1-score of 92.3%, validating the efficacy of the combined preprocessing approach. Our findings are consistent with recent literature, underscoring the importance of data-level balancing and feature reduction in developing robust IDS solutions. Furthermore, the proposed pipeline not only enhances detection accuracy but also reduces computational overhead, making it suitable for real-time SDN deployment. Future work may focus on testing this framework on additional SDN-specific datasets and integrating real-time attack simulation for further validation.

References:

[1] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1–22, Dec. 2019, doi: 10.1186/S42400-019-0038-7/FIGURES/8.

[2] S. Khorsandroo, A. G. Sánchez, A. S. Tosun, J. M. Arco, and R. Doriguzzi-Corin, "Hybrid SDN evolution: A comprehensive survey of the state-of-the-art," *Comput. Networks*, vol. 192, p. 107981, Jun. 2021, doi: 10.1016/J.COMNET.2021.107981.

[3] H. A. Ahmed, A. Hameed, and N. Z. Bawany, "Network intrusion detection using oversampling technique and machine learning algorithms," *PeerJ Comput. Sci.*, vol. 8, p. e820, 2022, doi: 10.7717/PEERJ-CS.820.

[4] U. Mbasuva and G. A. L. Zodi, "Designing Ensemble Deep Learning Intrusion Detection System for DDoS attacks in Software Defined Networks," *Proc. 2022 16th Int. Conf. Ubiquitous Inf. Manag. Commun. IMCOM 2022*, 2022, doi: 10.1109/IMCOM53663.2022.9721785.

[5] M. A. K. Islam, M. R., Rahman, M., Ahmed, M., & Azad, "Intrusion detection system using ML for SDN: Current state and future directions," *Comput. Secur.*, vol. 117, p. 102708, 2022.

[6] H. Zemrane, Y. Baddi, and A. Hasbi, "SDN-Based Solutions to Improve IOT: Survey," *Colloq. Inf. Sci. Technol. Cist*, vol. 2018-October, pp. 588–593, Dec. 2018, doi: 10.1109/CIST.2018.8596577.

[7] L. Zhou, J., Pan, Y., & Zhang, "An enhanced SDN dataset for benchmarking intrusion detection algorithms," *IEEE Access*, vol. 11, pp. 28514–28527, 2023.

[8] K. W. B. Nitesh V. Chawla, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002, doi: 10.1613/jair.953.

[9] S. B. Saidin and S. B. I. Hisham, "A Survey on Supervised Machine Learning in Intrusion Detection Systems for Internet of Things," *8th Int. Conf. Softw. Eng. Comput. Syst. ICSECS 2023*, pp. 419–423, 2023, doi: 10.1109/ICSECS58457.2023.10256275.

[10] O. P. Garg, M., Bhatt, C., & Verma, "An effective hybrid PCA-based intrusion detection system," *Expert Syst. Appl.*, vol. 168, p. 114375, 2021.

[11] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010, doi: 10.1002/wics.101.

[12] B. L. Shen He, "An Effective Cost-Sensitive XGBoost Method for Malicious URLs Detection in Imbalanced Dataset," *IEEE Access*, 2021, doi: 10.1109/ACCESS.2021.3093094.

[13] M. U. G. Abid, A., Alenezi, M., & Khan, "Improved LightGBM-based intrusion detection for SDN using hybrid sampling," *IEEE Access*, vol. 11, pp. 49367–49378, 2023.

[14] M. K. & N. S. Md. Alamin Talukder, "A hybrid machine learning model for intrusion

detection in wireless sensor networks leveraging data balancing and dimensionality reduction," *Sci. Rep.*, vol. 15, no. 4617, 2025, doi: https://doi.org/10.1038/s41598-025-87028-1.

[15] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo, "Brain Tumor Detection Based on Deep Learning Approaches and Magnetic Resonance Imaging," *Cancers (Basel).*, vol. 15, no. 16, p. 4172, Aug. 2023, doi: 10.3390/CANCERS15164172.

[16] A. Ali, I., Khan, S. U., & Rehman, "Machine learning-based intrusion detection in SDN using InSDN dataset and dimensionality reduction," *Electronics*, vol. 13, no. 9, p. 1678, 2023.

[17] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, Oct. 2018, doi: 10.1016/j.ins.2018.06.056.