e-palli

American Journal of Smart Technology and Solutions (AJSTS)

ISSN: 2837-0295 (ONLINE)

VOLUME 4 ISSUE 1 (2025)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of Smart Technology and Solutions (AJSTS)

Volume 4 Issue 1, Year 2025
ISSN: 2837-0295 (Online)
DOI: https://doi.org/10.54536/ajsts.v4i1.2912
https://journals.e-palli.com/home/index.php/ajsts

Addressing Class Imbalance in IoT: A Comparative Analysis of Resampling Techniques

Yousef Qawqzeha[1*]

Article Information

Received: July 09, 2024
Accepted: August 01, 2024
Published: February 18, 2025

Keywords

Multi-Class Classification, Resampling Techniques, Class Imbalance, Hyperparameter Tuning, Fraud Detection

ABSTRACT

In modern times, automated task processing and sophisticated algorithm design are important tools for using cutting-edge technologies and approaches to extract insights from data and practical solutions. The machine learning models powered by data have produced outputs that were either more or less worthy when the input datasets were balanced. An uneven distribution of classes in the input datasets has resulted in imbalanced data. Class imbalance has been a significant challenge in machine learning applications, particularly when working with substantially disparate distributions like those found in Internet of Things datasets. This study addressed the class imbalance issue in IoT data by comparing various resampling strategies. The study aimed to find efficient ways to realign class distributions and enhance the functionality of machine learning models implemented in Internet of Things systems. A predictive model built on an unbalanced data set appeared to have high accuracy, but it struggled to generalise new data from the minority class. Resampling techniques, including Over-sampling, Under-sampling, SMOTE (Synthetic Minority Over-Sampling Technique), and ADASYN (Adaptive Synthetic Sampling), were evaluated using an extensive variety of IoT datasets spanning different classes and domains. The functionality of each technique was assessed using performance metrics such as the area covered by AUC, F1-score, precision, and recall. This study advanced the understanding of class imbalance mitigation in IoT data processing by providing insights into creating more durable and trustworthy models for IoT scenarios. CCS CONCEPTS • Class Imbalance • Applied Computing • Machine Learning • Internet of Things (IoT)

INTRODUCTION

The Internet of Things (IoT) has recently changed several industries. It has made it possible to collect and analyse an immense quantity of sensor data for various applications, from smart healthcare to the automobile industry (Pramanik *et al.*, 2019). The Internet of Things (IoT) is a significant advancement in artificial intelligence, transforming our daily lives through various functions like device modelling, control, data publishing, analysis, and detection (Wanasinghe *et al.*, 2020). It has outpaced other technologies due to its promising future and ability to analyse and study various elements, making it a significant milestone in the field (Nord *et al.*, 2019).

However, class imbalance in the datasets has been one of the main obstacles to fully utilising the potential of IoT data. When one class greatly outnumbers the others, class imbalances arise, which bias model results and lower predicted accuracy. If it is discovered that the amount of data points in two-class classification models or multi-class data models is roughly the same, handling a dataset with sufficient data points is not too challenging (Peng *et al.*, 2023; Qawqzeh & Ashraf, .2023).

The utilisation of IoT generates non-stationary data streams that can change over time, making it challenging for Machine learning algorithms to identify minority exposure accurately (Nixon *et al.*, 2019). The lack of robust computing equipment can disrupt machine learning methods like oversampling and undersampling, affecting their ability to operate in complex environments (Atuhurra *et al.*, 2024). Cyberattacks also tax IoT networks, leading to highly skewed datasets. Momentum detection of minority-class hacking is crucial in IoT networks, but models tend to favour the majority of normal-class sites. Deep learning techniques for class imbalance have been applied to image recognition, but their application to non-image IoT data may require different approaches (Atuhurra *et al.*, 2024; Johnson & Khoshgoftaar, 2019).

This research aims to reduce class disparity in two IoT datasets: IoT Modbus and IoT GPS Tracker. Addressing class imbalance is crucial for ensuring this reliability and effectiveness in machine learning models deployed in IoT systems (Qawqzeh & Ashraf, 2023; Tanha *et al.*, .2020; Varotto *et al.*, 2021; Welvaars *et al.*, 2023). The main goal is a competent and comparative analysis of methods designed to address class imbalance in these datasets. In particular, we examine the effectiveness of several techniques, such as the Synthetic Minority Over-Sampling Technique (SMOTE), Random Under-sampling (RUS), Random Over-sampling (ROS), and Adaptive Synthetic Sampling (ADASYN) method, in resolving class imbalance in multi-class scenarios. Machine learning algorithms trained on imbalance datasheets tend to perform poorly on minority class cases, which are generally more interesting in detection and forecasting scenarios, in favour of the majority class (Koziarski *et al.*, 2020).

Multi-class classification is a task with more than two classes and assumes that an object can only receive one classification. The trained model, constructed with

[1] University of Fujairah, 48CP+J4P - E89, Mraisheed, Fujairah, United Arab Emirates
[*] Corresponding author's e-mail: YousefQawqzehaa@outlook.com

this dataset, will function according to the authors' expectations. It is common knowledge that these data points are balanced datasets.

However, the issue arises when skewed datasets, such as those with under- or over-representation, are acquired to create a data model for predictive analysis. We seek to find practical approaches for enhancing the predictive accuracy and generalisation of machine learning models in the Internet of Things applications by assessing these methods' effectiveness on various IoT datasets.

This 'study's goal was to advance state-of-the-art IoT data analysis and make it easier to create more durable and trustworthy predictive models for practical IoT scenarios by offering insights into the selection and application of resampling techniques designed to address class imbalance in IoT datasets (Obaid & Nassif, .2022; Paisitkriangkrai *et al.*, 2013; Wang & Yao, .2012).

LITERATURE REVIEW

Introduction to Internet of Things

The Internet of Things (IoT) is a rapidly evolving technology that uses processing power, downsized electronics, and networking links to connect devices and systems (Kumar *et al.*, 2021). It has sparked debates on various aspects, including opportunities for new companies, security, privacy, compatibility, and international ecosystem. The IoT will impact various aspects of life, including elders, consumers, and healthcare providers (Pal *et al.*, 2018). To ensure energy savings and elasticity, IoT devices like smart home appliances need authentication and optimization of energy consumption. This technology has the potential to revolutionize various aspects of our lives (Powroźnik *et al.*, 2021). Personal IoT devices, such as wearable fitness and health monitoring devices, are also expected to improve independence and quality of life for people with disabilities and the elderly (Khodadadi *et al.*, 2016). IoT systems, such as networked vehicles and intelligent traffic systems, are moving towards smart cities, reducing congestion and energy consumption. However, IoT also presents challenges that need to be addressed for potential benefits to be realized (Rose *et al.*, 2015).

Class Imbalance in Machine Learning

One of the significant challenges associated with IoT is managing the vast amounts of data generated by these devices, which often leads to class imbalance in machine learning applications. In machine learning, class imbalance is a widespread problem that impacts several industries, such as cybersecurity, finance, and healthcare (Dogra *et al.*, 2022). Class imbalance is a major difficulty in IoT because data collecting is naturally biased towards typical operational conditions (Zhou *et al.*, 2022). Imbalanced datasets are those where one of two possible outcomes is rare (Tyagi & Mittal, 2020). A classification model's performance depends on the training dataset's quality and quantity (Hanskunatai, 2018). In imbalanced datasets with two-valued classes, accuracy may not clearly represent classification results. In applications like disease detection and intrusion detection, it is more important to correctly predict the minority class (Tyagi & Mittal, .2020). The very visible presence of a class imbalance is depicted in Figure 1.

Class imbalance presents particular difficulties in the IoT because of the type of data that IoT devices and sensor networks produce (Ullah & Mahmoud, 2021).

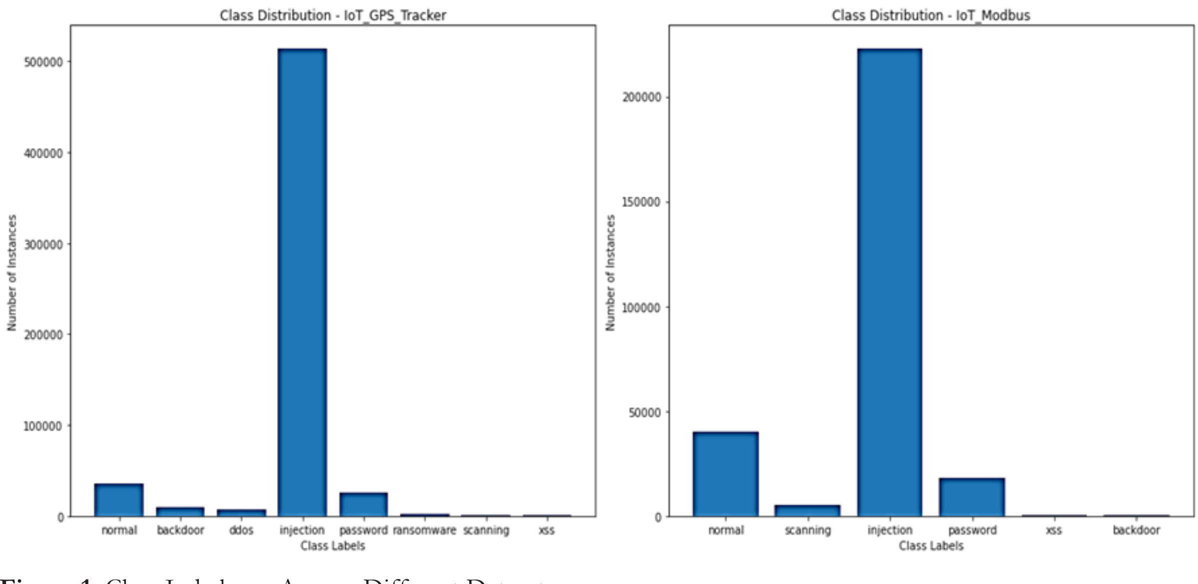


Figure 1: Class Imbalance Among Different Datasets
Source: Author

Class imbalance in IoT information has been studied, especially in applications for environmental monitoring, predictive maintenance, and anomaly detection (Coelho *et al.*, 2022; Fahim & Sillitti, 2019). According to certain studies, undersampling strategies might exclude important information from the majority class, while oversampling could cause overfitting or injecting noise into the data (Koziarski *et al.*, .2019; Sáez *et al.*, 2016). To address class imbalance issues in machine learning, the dataset itself or the learning methods of the underlying algorithm can be tuned.

Handling Approaches

At the algorithm level, approaches like Ada boosting, mapping, and cost-sensitive learning can be used to tune the classifier's results. Data level-based imbalance handling involves equating the occurrence of both classes algorithmically to improve the imbalance ratio (Tyagi & Mittal, 2020). Several studies have looked into ways to address the class disparity, and resampling techniques have become prominent approaches. By creating synthetic samples, oversampling techniques like Adaptive Synthetic Sampling (ADASYN) and Synthetic Minority Over-Sampling Technique (SMOTE) seek to boost the representation of minority class instances (Huang, 2015; Tarawneh *et al.*, .2020). To rebalance class distributions, under-sampling techniques, on the other hand, require lowering the quantity of majority class samples (Abdi & Hashemi, .2015). In order to produce a balanced dataset, hybrid approaches use both under-sampling and oversampling methods.

Several machine learning methods may be used to create predictive data models. The model's accuracy depends on how well it can identify the positive class and how well it can predict a negative class (Fisher *et al.*, 2019). The categorisation rate of the two classes mentioned above has been verified, even if a model provides 90% accuracy. Unbalanced data sets can cause skewed proportions between groups, necessitating preprocessing sample techniques, algorithmic approaches, or a bot to shift the model for sustainable analysis.

ADASYN, CoSen modelling, SMOTE, under- and over-sampling, and SMOTE have been commonly used solutions. A balanced dataset was produced by undersampling, which removes the sample of the dominant class. The loss of important information was ascribed to the dataset's undersampling. Conversely, over-sampling attempted to balance the dataset by making duplicates of the pre-existing dataset. It could be arbitrary duplicates of the data subset., leading to overfitting of the model, which is often computationally costly. Instead of adding new data samples to the minority class or replacing the current samples, the SMOTE-based approach artificially produces the sample data. The SMOTE-based method faces a problem due to the undesirable addition of noise to the dataset.

The study has focused on adding knowledge on class imbalance mitigation in IoT data analysis by comparing resampling approaches in IoT datasets. Advancements in IoT data analytics have created more durable and scalable IoT systems. The study aims to identify practical approaches for enhancing the performance and dependability of machine learning models used in IoT applications through empirical evaluation and methodical comparison.

RESEARCH METHODOLOGY

This research has extensively utilised an experimental design to comprehensively contrast resampling techniques that were applied to deal with the problem of class imbalance in the presence of noise data. By leveraging resampling methods alongside comparative analysis, the dataset was divided into two subsets: a training dataset through which the model was trained and a testing dataset through which the performance of the model was tested. This experimental setting provided a systematic means of testing the impact of various resampling schemes on classifier performance.

Comparison of RF and SVC Classifiers

The study includes a comparative investigation of base classifiers, namely the Random Forest Classifier (RF) and the Support Vector Classifier (SVC). This comparison assesses their effectiveness in resolving class disparity in the context of multi-class issues. According to the study's findings, the RF classifier performs better than other basic classifiers in reducing the difficulties caused by class imbalance in multi-class situations. The predictive model is an RF classifier because of its solid performance history.

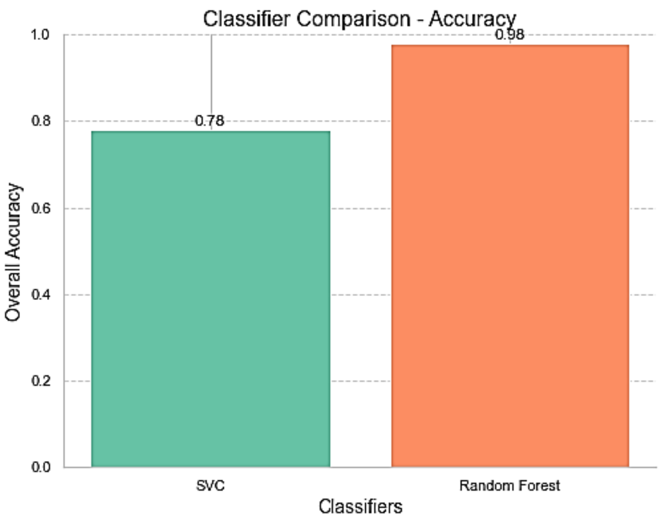


Figure 2: Performance Evaluation of Base Classifiers on the IoT_Modbus Dataset

Dataset Selection
In selecting specific datasets for this study, various IoT datasets spanning several application areas have been explored, selecting those that could demonstrate a substantial class imbalance. The main objective of the data selection process was to ensure a comprehensive analysis of specific datasets with various attributes, including considering datasets of various sizes, imbalance ratios, and feature space dimensionality.

Preprocessing
Standard data preparation procedures have been implemented to ensure the datasets are appropriately cleaned, normalised, and subjected to feature engineering techniques. This process aims to enhance data quality and consistency by effectively addressing noise, outliers, and missing values. Resampling strategies were employed to rebalance class distributions within the IoT datasets, including random under-sampling and oversampling techniques such as SMOTE and ADASYN. By generating balanced training sets through resampling, the models were trained and evaluated more effectively, mitigating the impact of class imbalance and improving the overall performance of the predictive models.

Selection of Base Classifier
Two popular SVC (support vector classifier) and RF (random forest classifier) classifiers were selected for the study. However, the choice of specific classifiers has depended on the specific properties of IoT data and its function.

Model Training and Evaluation
The original unbalanced and the resampled datasets have been utilised to train the CSV and RF Classifiers. Stratified cross-validation has been applied to ensure unbiased performance evaluation and mitigate the impact of dataset imbalance during model assessment. Standard assessment metrics such as ROC curve (AUC), precision, recall, and F1-score have been employed to evaluate the performance of each classifier. Additionally, the effectiveness of SVC and RF classifiers in handling class imbalance within the IoT datasets has been compared across several resampling methods.

Statistical Analysis
The performance of SVC and RF classifiers on original and resampled datasets has been evaluated by statistical methods such as t-tests or Wilcoxon signed-rank tests. The outcomes of the statistical analysis were performed to identify significant variations in the performance of classifiers. Additionally, the analysis enabled the detection of how resampling methods affect the performance of the classifiers.

Sensitivity Analysis
Sensitivity analysis has been employed to assess the resilience of SVC and RF classifiers to variations in dataset properties, encompassing changes in feature space dimensionality, dataset size, and class imbalance ratio. By systematically varying these properties, the study aimed to understand how the classifiers' performance adapts to different data configurations. Additionally, an analysis was conducted to investigate the impact of algorithmic decisions and hyperparameter settings on classifier performance within diverse resampling scenarios. This analysis will provide insights into the robustness of SVC and RF classifiers across various conditions, enabling a comprehensive evaluation of their suitability for handling class imbalance and other challenges inherent in IoT datasets.

Discussion and Interpretation
The discussion and interpretation part of the study has utilised the analysis of experimental findings based on the comparative performance evaluation of SVC and RF classifiers while employing various resampling strategies. The advantages and disadvantages of each classifier in managing the class imbalance have been discussed based on the results that have affected IoT data analysis. Key factors like interpretability, computational efficiency, and model resilience have also been considered to ensure effective performance. The study has been summarised with possible directions for further studies, such as investigating hybrid or ensemble methodologies to enhance classifier performance in imbalanced IoT datasets.

Resampling Methods and Model Evaluation Strategy
A collection of resampling methods, such as "No Resampling," "ROS," "RUS," "SMOTE," and "ADASYN," have been presented to solve class imbalance and evaluate its impact on the overall performance of the model. This stage involved determining if resampling was necessary and considering the "No Resampling" option to comprehend the impact of class imbalance on model performance. A specific OvO classifier has been utilised for multi-class classification, as it was well-suited for the situations where several classes were present and can be trained efficiently with both original and resampled datasets.

The model's performance was evaluated during the OvO's training on the chosen dataset, regardless of Resampling. After the training phase, predictions were created for the test set, and accuracy scores were carefully determined. Furthermore, individualised confusion matrices were constructed for each resampling scenario, comprehensively evaluating the model's performance across various resampling techniques. A visual representation of this methodology has been presented in Figure 3.

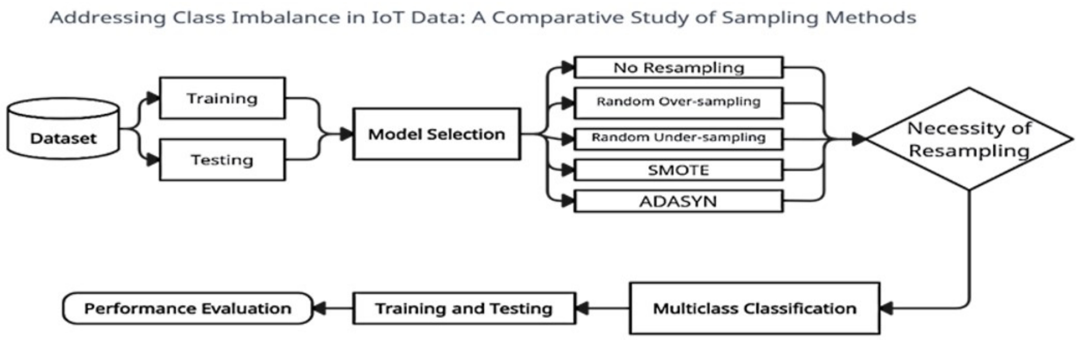


Figure 3: Illustration of Resampling Techniques for Class Imbalance Handling in IoT Environment

RESULTS AND DISCUSSION

Analysis of Base Classifiers and Resampling Strategies in IoT Datasets

A comprehensive analysis of base classifiers and resampling strategies was conducted to identify key findings regarding class imbalance in IoT datasets. The Random Forest Classifier (RF) and the Support Vector Classifier (SVC) performed biasedly, favouring the majority class with superior accuracy, precision, and recall on the original unbalanced datasets. The study revealed a trade-off between recall and accuracy, with SVC demonstrating better recall but poorer precision than RF. However, RF regularly outperformed SVC in terms of total F1-score on datasets that were not evenly distributed.

The successful mitigation of class imbalance resampling strategies led to better performance across various assessment measures for both classifiers. SMOTE and ADASYN oversampling techniques significantly improved memory for the minority class, reducing and alleviating the imbalance-induced bias in predictions.

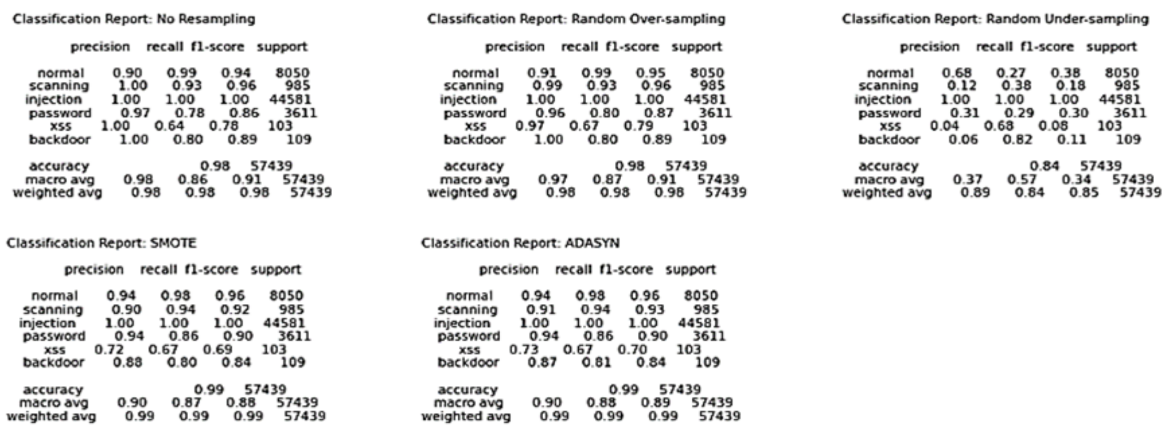
Furthermore, sensitivity to the minority class was enhanced, but overall accuracy was decreased when random undersampling approaches were utilised. Although RF consistently outperformed SVC in various conditions, demonstrating greater overall accuracy, precision, recall, and F1-score in the comparison study of SVC and RF resampled datasets. RF also outperformed SVC in sensitivity to minority classes and produced a stronger recall-to-precision ratio.

Statistical Analysis of Base Classifier Performance

The statistical analysis of the study revealed that RF outperformed SVC in managing class imbalance to IoT datasets, highlighting the significance of its performance. Additionally, various dataset properties such as imbalance ratio, size, and feature space dimensionality affected the efficacy of resampling algorithms differently, with RF classifiers demonstrating greater resistance to these fluctuations than SVC classifiers.

These findings of the study have emphasised the significance of selecting appropriate resampling techniques and classifiers for IoT datasets, thereby enhancing the development of robust and reliable predictive models for diverse applications. In this study, the performance of a classification model using various resampling techniques was assessed to determine class imbalance in the dataset. The classification reports associated with each technique yield valuable insights into the model's precision, recall, F1 scores, and accuracy across various classes. These results offered a comprehensive perspective on how diverse resampling methods influence the model's ability to accurately classify instances across various categories (Collell *et al.*, 2018; Qawqzeh *et al.*, 2020).

The subsequent sections will delve deeper into the implications of these results and their significance in selecting the most appropriate resampling technique. Figures 4 and 5 represent the classification reports for the



Classification Report: No Resampling

	precision	recall	f1-score	support
normal	0.90	0.99	0.94	8050
scanning	1.00	0.93	0.96	985
injection	1.00	1.00	1.00	44581
password	0.97	0.78	0.86	3611
xss	1.00	0.64	0.78	103
backdoor	1.00	0.80	0.89	109
accuracy			0.98	57439
macro avg	0.98	0.86	0.91	57439
weighted avg	0.98	0.98	0.98	57439

Classification Report: Random Over-sampling

	precision	recall	f1-score	support
normal	0.91	0.99	0.95	8050
scanning	0.99	0.93	0.96	985
injection	1.00	1.00	1.00	44581
password	0.96	0.80	0.87	3611
xss	0.97	0.67	0.79	103
backdoor	1.00	0.80	0.89	109
accuracy			0.98	57439
macro avg	0.97	0.87	0.91	57439
weighted avg	0.98	0.98	0.98	57439

Classification Report: Random Under-sampling

	precision	recall	f1-score	support
normal	0.68	0.27	0.38	8050
scanning	0.12	0.38	0.18	985
injection	1.00	1.00	1.00	44581
password	0.31	0.29	0.30	3611
xss	0.04	0.68	0.08	103
backdoor	0.06	0.82	0.11	109
accuracy			0.84	57439
macro avg	0.37	0.57	0.34	57439
weighted avg	0.89	0.84	0.85	57439

Classification Report: SMOTE

	precision	recall	f1-score	support
normal	0.94	0.98	0.96	8050
scanning	0.90	0.94	0.92	985
injection	1.00	1.00	1.00	44581
password	0.94	0.86	0.90	3611
xss	0.72	0.67	0.69	103
backdoor	0.88	0.80	0.84	109
accuracy			0.99	57439
macro avg	0.90	0.87	0.88	57439
weighted avg	0.99	0.99	0.99	57439

Classification Report: ADASYN

	precision	recall	f1-score	support
normal	0.94	0.98	0.96	8050
scanning	0.91	0.94	0.93	985
injection	1.00	1.00	1.00	44581
password	0.94	0.86	0.90	3611
xss	0.73	0.67	0.70	103
backdoor	0.87	0.81	0.84	109
accuracy			0.99	57439
macro avg	0.90	0.88	0.89	57439
weighted avg	0.99	0.99	0.99	57439

Figure 4: IoT_Modbus dataset classification reports of the used resampling techniques

Classification Report: No Resampling

	precision	recall	f1-score	support
normal	0.91	0.92	0.91	7046
backdoor	0.81	0.78	0.79	2076
ddos	0.80	0.79	0.80	1396
injection	1.00	1.00	1.00	102804
password	0.82	0.85	0.84	5007
ransomware	0.66	0.61	0.63	568
scanning	0.83	0.41	0.55	127
xss	0.39	0.32	0.35	114
accuracy			0.98	119138
macro avg	0.78	0.71	0.73	119138
weighted avg	0.98	0.98	0.98	119138

Classification Report: Random Over-sampling

	precision	recall	f1-score	support
normal	0.91	0.91	0.91	7046
backdoor	0.82	0.77	0.79	2076
ddos	0.82	0.77	0.79	1396
injection	1.00	1.00	1.00	102804
password	0.82	0.85	0.84	5007
ransomware	0.61	0.61	0.61	568
scanning	0.71	0.55	0.62	127
xss	0.34	0.40	0.37	114
accuracy			0.98	119138
macro avg	0.75	0.73	0.74	119138
weighted avg	0.98	0.98	0.98	119138

Classification Report: Random Under-sampling

	precision	recall	f1-score	support
normal	0.79	0.51	0.62	7046
backdoor	0.42	0.53	0.47	2076
ddos	0.38	0.73	0.50	1396
injection	1.00	1.00	1.00	102804
password	0.53	0.34	0.41	5007
ransomware	0.32	0.68	0.43	568
scanning	0.10	0.98	0.18	127
xss	0.12	0.92	0.21	114
accuracy			0.93	119138
macro avg	0.46	0.71	0.48	119138
weighted avg	0.95	0.93	0.93	119138

Classification Report: SMOTE

	precision	recall	f1-score	support
normal	0.96	0.94	0.95	7046
backdoor	0.87	0.90	0.88	2076
ddos	0.88	0.92	0.90	1396
injection	1.00	1.00	1.00	102804
password	0.92	0.88	0.90	5007
ransomware	0.70	0.82	0.76	568
scanning	0.74	0.83	0.78	127
xss	0.36	0.72	0.48	114
accuracy			0.99	119138
macro avg	0.80	0.88	0.83	119138
weighted avg	0.99	0.99	0.99	119138

Classification Report: ADASYN

	precision	recall	f1-score	support
normal	0.95	0.95	0.95	7046
backdoor	0.88	0.89	0.89	2076
ddos	0.90	0.89	0.89	1396
injection	1.00	1.00	1.00	102804
password	0.92	0.89	0.90	5007
ransomware	0.74	0.79	0.76	568
scanning	0.65	0.83	0.73	127
xss	0.38	0.64	0.48	114
accuracy			0.99	119138
macro avg	0.80	0.86	0.83	119138
weighted avg	0.99	0.99	0.99	119138

Figure 5: IoT_GPS_Tracker dataset classification reports of the used resampling techniques

resampling techniques utilised in this study, focusing on both the IoT_Modbus and IoT_GPS_Tracker datasets. The comprehensive analysis of the model highlighted the impact of diverse resampling techniques on classification performance across varied classes, facilitating a deeper understanding of their implications and avenues for potential enhancements. Previous research (Azlim & Ahamed, .2023; Jiang *et al.*, 2023; Qawqzeh *et al.*, 2023; Rezvani & Wang, 2023) have emphasised the symbiotic relationship between resampling techniques and the choice of classification methods, underscoring the necessity for synergy to maximise beneficial outcomes.

Absence of Resampling

In the absence of Resampling, the model achieved 98% notable accuracy. While demonstrating perfect precision and high recall for the "Injection" class and a commendable F1-score for "Scanning," lower F1-scores for "XSS" and "Backdoor" indicated areas for improvement.

Random Over-Sampling (ROS)

Maintaining a consistent accuracy of 98%, ROS exhibited strengths in perfect precision and high recall for the "Injection" class, along with a substantial F1-score for "Scanning." However, comparatively lower F1 scores for "Password" and "XSS" suggested potential areas for enhancement. ROS effectively addressed class distribution imbalance.

Random Under-Sampling (RUS)

Despite balancing class distribution, RUS lead to a reduced accuracy of 84%. While achieving a perfect F1-score for the "Injection" class, the model performance was significantly weakened across other classes, resulting in low precision and recall.

SMOTE (Synthetic Minority Over-Sampling Technique)

SMOTE achieved a commendable accuracy of 99%, maintaining excellent precision and recall for "Normal" and "Injection" classes, resulting in high F1 scores. However, relatively lower F1 scores for "XSS" and "Backdoor" suggested potential areas for improvement. SMOTE effectively addressed class imbalance by generating synthetic instances.

ADASYN (Adaptive Synthetic Sampling)

Similarly achieving a high accuracy of 99%, ADASYN sustained robust precision and recall for "Normal" and "Injection" classes, resulting in high F1 scores.

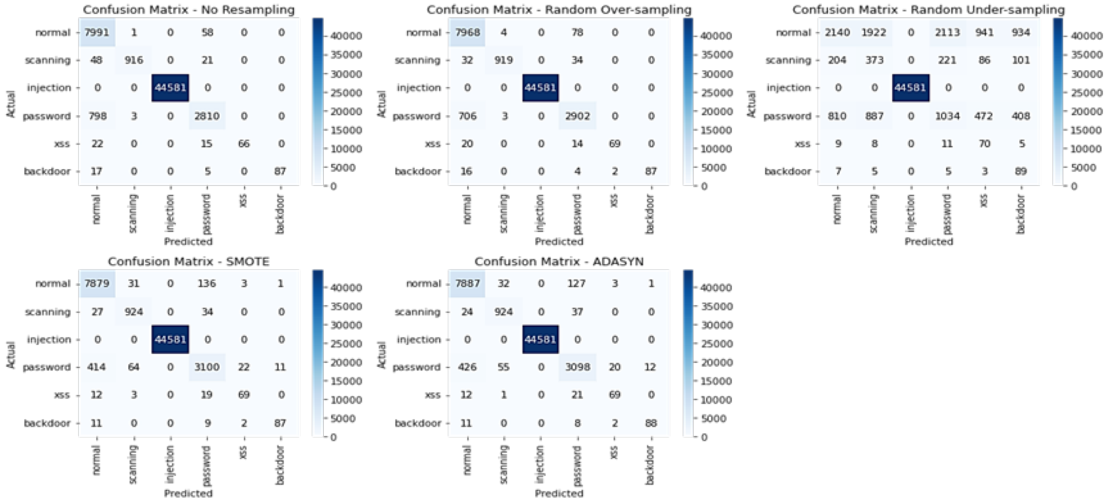


Figure 6: Heat maps showcase the resampling techniques employed on the IoT_Modbus dataset.

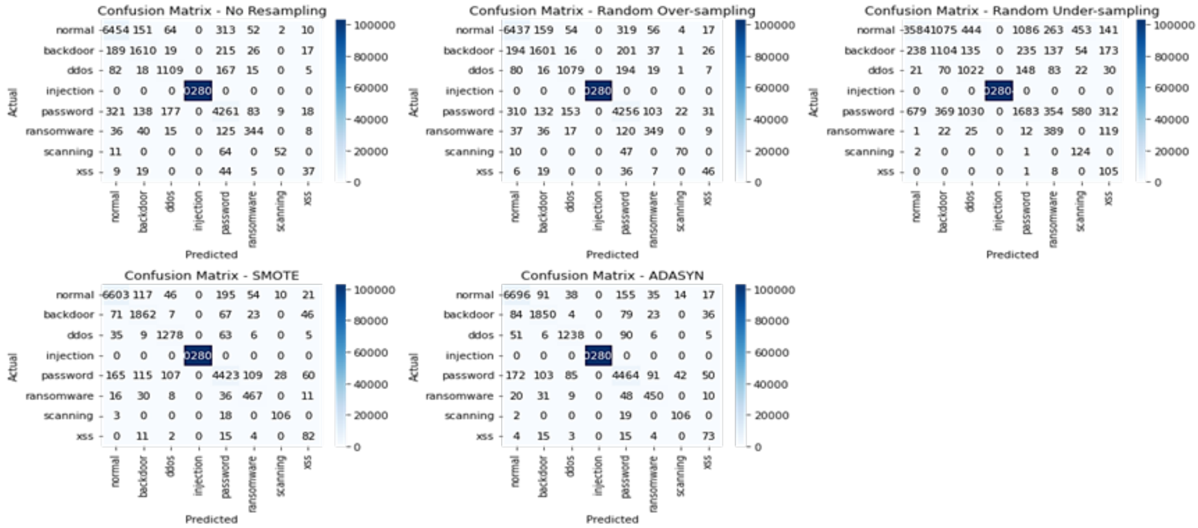


Figure 7: Heat maps showcase the resampling techniques employed on the IoT_GPS_Tracker dataset

Nevertheless, there was an area for enhancement in F1-scores for "XSS" and "Backdoor," signifying potential improvements. ADASYN effectively mitigates class imbalance through adaptive synthetic sampling. This analysis offered insights into the relative effectiveness of each resampling approach concerning precision, recall, F1 scores, and accuracy across different classes. RUS's overall performance has been compromised despite its improvement in the "Injection" class.

These findings informed conclusions regarding the suitability of resampling techniques in addressing class imbalance within a multi-class scenario. Heat maps were generated to represent the resampling techniques employed on the IoT_Modbus dataset, as shown in Figures 6 and 7.

CONCLUSION

The study examined the impact of resampling techniques on classification models in class-imbalanced IoT datasets. It found that Support Vector Classifier (SVC) and Random Forest Classifier (RF) performed biasedly on unbalanced datasets, highlighting the issue of class imbalance in machine learning tasks. Resampling strategies improved the performance of SVC and RF classifiers, with hybrid approaches like SMOTE and oversampling techniques like ADASYN enabling rebalancing class distributions and enhancing model performance. RF consistently outperformed SVC in resampling scenarios, achieving superior accuracy, precision, recall, and F1-score.

The study emphasized the importance of resampling techniques in scenarios marked by class imbalance to enhance the accuracy and reliability of classification models in practical applications. Future research could explore hyperparameter tuning's effects on model performance and explore the applicability of these techniques in domains like cybersecurity, fraud detection, and medical diagnosis.

Future Implications

This subsequent study may investigate several directions to expand our comprehension and improve the usefulness of resolving class imbalance in IoT datasets. Exploring hybrid or adaptive resampling methodologies might mitigate trade-offs observed in current techniques and potentially boost overall classification performance. Examine how well ensemble learning methods-like bagging, boosting, or stacking-work with resampling techniques to enhance model robustness and performance in unbalanced IoT datasets. By combining the advantages of several classifiers and resampling strategies, ensemble approaches may improve generalisation and prediction accuracy.

The study explores adaptive resampling techniques that can dynamically adjust to changes in the data distribution. Ensuring the ongoing efficacy of class imbalance mitigation strategies in practical applications may require creating algorithms that can recognise and react to idea drift, data drift, or changing class distributions in IoT datasets. The study suggests that exploring the compatibility between advanced classification algorithms and resampling methods is a promising direction. This exploration could unveil enhanced performance in complex multi-class scenarios, presenting opportunities for more robust and accurate models.

Moreover, examining the practical implications and robustness of these techniques in real-world scenarios, particularly in domains where accurate classification is imperative, would be instrumental. This includes rigorous testing and validation of these techniques in operational settings to gauge their effectiveness and feasibility beyond controlled experimental setups. Validate the effectiveness of the identified resampling techniques and classifiers through deployment in real-world IoT environments. Conduct extensive evaluation and monitoring of model performance under practical conditions, considering scalability, reliability, and interpretability factors.

Case studies and field trials in diverse IoT domains could provide valuable insights into the applicability and impact of class imbalance mitigation strategies in real-world settings. The project proposes frameworks and automated methods for selecting suitable classifiers, hyperparameters, and resampling techniques for IoT dataset's properties. The model creation process might be streamlined by automated model selection and hyperparameter tweaking, allowing practitioners to quickly find and implement efficient predictive models in Internet of Things applications.

The findings of this study provide a foundation for future research, emphasising the need for tailored techniques and their practical applications in addressing class imbalance within the dynamic landscape of IoT datasets. The enhancement of the state-of-the-art in-class imbalance mitigation strategies for IoT datasets by addressing these future research objectives will eventually improve the predictive modelling's performance, scalability, and reliability in various IoT applications.

Acknowledgment

We would like to express our gratitude to Yousef Qawqzeha from the University of Fujairah for his valuable contributions to this research. His expertise and dedication greatly enriched the development of this study. We also acknowledge the support provided by his Corresponding Author Email (YousefQawqzehaa@outlook.com) and his ORCID (0000-0001-7774-062x) throughout the research process.

References

Abdi, L., & Hashemi, S. (2015). To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE transactions on Knowledge and Data Engineering, 28*(1), 238-251.

Atuhurra, J., Hara, T., Zhang, Y., Sasabe, M., & Kasahara, S. (2024). Dealing with Imbalanced Classes in Bot-IoT Dataset. *arXiv preprint arXiv:2403.*18989.

Azlim Khan, A. K., & Ahamed Hassain Malim, N. H. (2023). Comparative Studies on Resampling Techniques in Machine Learning and Deep Learning Models for Drug-Target Interaction Prediction. *Molecules, 28*(4), 1663.

Coelho, D., Costa, D., Rocha, E. M., Almeida, D., & Santos, J. P. (2022). Predictive maintenance on sensorized stamping presses by time series segmentation, anomaly detection, and classification algorithms. *Procedia Computer Science, 200,* 1184-1193.

Collell, G., Prelec, D., & Patil, K. R. (2018). A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing, 275,* 330-340.

Dogra, V., Verma, S., Verma, K., Jhanjhi, N. Z., Ghosh, U., & Le, D.-N. (2022). A comparative analysis of machine learning models for banking news extraction by multiclass classification with imbalanced datasets of financial news: challenges and solutions.

Fahim, M., & Sillitti, A. (2019). Anomaly detection, analysis and prediction techniques in iot environment: A systematic literature review. *IEEE Access, 7,* 81664-81681.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1-81.

Hanskunatai, A. (2018). A new hybrid sampling approach for classification of imbalanced datasets. *2018 3rd International Conference on Computer and Communication Systems* (ICCCS),

Huang, P. J. (2015). Classification of imbalanced data using synthetic over-sampling techniques. University of California, Los Angeles.

Jiang, X., Wang, J., Meng, Q., Saada, M., & Cai, H. (2023). An adaptive multi-class imbalanced classification framework based on ensemble methods and deep network. *Neural Computing and Applications, 35*(15), 11141-11159.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data, 6*(1), 1-54.

Khodadadi, F., Dastjerdi, A. V., & Buyya, R. (2016). Internet of things: an overview. *Internet of things,* 3-27.

Koziarski, M., Krawczyk, B., & Woźniak, M. (2019). Radial-based oversampling for noisy imbalanced data classification. *Neurocomputing, 343,* 19-33.

Koziarski, M., Woźniak, M., & Krawczyk, B. (2020). Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise. *Knowledge-Based Systems, 204,* 106223.

Kumar, A., Sharma, S., Singh, A., Alwadain, A., Choi, B.-J., Manual-Brenosa, J., Ortega-Mansilla, A., & Goyal, N. (2021). Revolutionary strategies analysis and proposed system for future infrastructure in internet of things. *Sustainability, 14*(1), 71.

Nixon, C., Sedky, M., & Hassan, M. (2019). Practical application of machine learning based online intrusion detection to internet of things networks. *2019 IEEE Global Conference on Internet of Things (GCIoT).*

Nord, J. H., Koohang, A., & Paliszkiewicz, J. (2019). The Internet of Things: Review and theoretical framework. *Expert Systems with Applications, 133,* 97-108.

Obaid, W., & Nassif, A. B. (2022). The effects of resampling on classifying imbalanced datasets. *2022 Advances in Science and Engineering Technology International Conferences (ASET).*

Paisitkriangkrai, S., Shen, C., & van den Hengel, A. (2013). A scalable stagewise approach to large-margin multiclass loss-based boosting. *IEEE transactions on neural networks and learning systems, 25*(5), 1002-1013.

Pal, D., Funilkul, S., Charoenkitkarn, N., & Kanthamanon, P. (2018). Internet-of-things and smart homes for elderly healthcare: An end user perspective. *IEEE Access, 6,* 10483-10496.

Peng, H., Wu, C., & Xiao, Y. (2023). CBF-IDS: Addressing Class Imbalance Using CNN-BiLSTM with Focal Loss in Network Intrusion Detection System. *Applied Sciences, 13*(21), 11629.

Powroźnik, P., Szcześniak, P., & Piotrowski, K. (2021). Elastic energy management algorithm using IoT technology for devices with smart appliance functionality for applications in smart-grid. *Energies, 15*(1), 109.

Pramanik, P. K. D., Upadhyaya, B. K., Pal, S., & Pal, T. (2019). Internet of things, smart sensors, and pervasive systems: Enabling connected and pervasive healthcare. In *Healthcare data analytics and management* (pp. 1-58). Elsevier.

Qawqzeh, Y. K., Alourani, A., & Ghwanmeh, S. (2023). An improved breast cancer classification method using an enhanced AdaBoost classifier. *International Journal of Advanced Computer Science and Applications, 14*(1).

Qawqzeh, Y. K., & Ashraf, M. (2023). A Fraud Detection System Using Decision Trees Classification in An Online Transactions. *Proceedings of the 2023 12th International Conference on Software and Computer Applications.*

Qawqzeh, Y. K., Bajahzar, A. S., Jemmali, M., Otoom, M. M., & Thaljaoui, A. (2020). Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modeling. *BioMed Research International, 2020.*

Rezvani, S., & Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing, 110415.* https://doi.org/10.1016/j.asoc.2023.110415

Rose, K., Eldridge, S., & Chapin, L. (2015). The internet of things: An overview. *The internet society (ISOC), 80*(15), 1-53.

Sáez, J. A., Krawczyk, B., & Woźniak, M. (2016). Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition, 57,* 164-178.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data, 7,* 1-47.

Tarawneh, A. S., Hassanat, A. B., Almohammadi, K., Chetverikov, D., & Bellinger, C. (2020). Smotefuna: Synthetic minority over-sampling technique based on furthest neighbour algorithm. *IEEE Access, 8,* 59069-59082.

Tyagi, S., & Mittal, S. (2020). Sampling approaches for imbalanced data classification problem in machine learning. Proceedings of ICRIC 2019: Recent innovations in computing,

Ullah, I., & Mahmoud, Q. H. (2021). A framework for anomaly detection in IoT networks using conditional generative adversarial networks. *IEEE Access, 9,* 165907-165931.

Varotto, G., Susi, G., Tassi, L., Gozzo, F., Franceschetti, S., & Panzica, F. (2021). Comparison of resampling techniques for imbalanced datasets in machine learning: application to epileptogenic zone localization from interictal intracranial EEG recordings in patients with focal epilepsy. *Frontiers in Neuroinformatics, 15,* 715421.

Wanasinghe, T. R., Gosine, R. G., James, L. A., Mann, G. K., De Silva, O., & Warrian, P. J. (2020). The internet of things in the oil and gas industry: a systematic review. *IEEE Internet of Things Journal, 7*(9), 8654-8673.

Wang, S., & Yao, X. (2012). Multiclass imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(*4), 1119-1130.

Welvaars, K., Oosterhoff, J. H., van den Bekerom, M. P., Doornberg, J. N., van Haarst, E. P., OLVG Urology Consortium, & R, t. M. L. C. v. d. Z. J. v. A. G. L. B. H. M. K. P. B. L. v. d. K. A. M. W. P. (2023). Implications of resampling data to address the class imbalance problem (IRCIP): an evaluation of impact on performance between classification algorithms in medical data. *JAMIA open, 6*(2), ooad033.

Zhou, X., Hu, Y., Wu, J., Liang, W., Ma, J., & Jin, Q. (2022). Distribution bias aware collaborative generative adversarial network for imbalanced deep learning in industrial IoT. *IEEE Transactions on Industrial Informatics, 19*(1), 570-580.