

A COMPARATIVE STUDY OF SMOTE AND ADASYN FOR MULTICLASS CLASSIFICATION OF IOT ANOMALIES

*Fayez Alharbi**

Department of Information Technology, College of Computer and Information Sciences, Majmaah University. Al-Majmaah, 11952, Saudi Arabia

* Corresponding Author, e-mail: fs.alharbi@mu.edu.sa

Abstract: The advancement of IoT technologies requires stronger threat detection systems because cyber-physical risks have emerged. Traditional ML models demonstrate difficulty when working with datasets that have uneven class distributions specifically when detecting unusual security threats. To address this problem the research applies "IoT_Modbus" as its multiclass cybersecurity classification benchmark against which it evaluates advanced resampling approaches for better model results. Ensemble classifiers of Random Forest (RF), Gradient Boosting, XGBoost, LightGBM, and CatBoost together with Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN) enhanced the achievement of accuracy improvement. RF achieved the highest accuracy of 99.06% and 99.00% using SMOTE and ADASYN respectively. The XGBoost model demonstrated the highest accuracy rate at 86.21% with SMOTE and 86.63% with ADASYN among the various boosting-based models alongside CatBoost. The results achieved have shown an excellent promise of resampling techniques in addressing class imbalance coupled with predictive power. The analysis trends are better drawn with visual instruments like heatmaps and accuracy charts. The solution allows data security personnel to have an operational model in the identification of IoT threats while indicating that the right resampling strategies lead to better predictive results. This approach provides guidance on selecting imbalanced datasets while giving useful secus for the security protection of IoT systems in interconnected environments.

Key words: IoT; resampling techniques; class imbalance; OvO classification; accuracy optimization.

1. INTRODUCTION

The emerging modern information technology devices through IoT have altered several sectors malleability early include health care and industrial automation through enabling easily possible data connection. Enormous coupling of IoT devices has exposed intricate security weakness, making them easily prone to cyberattack threats against the

systems [1]. Interconnected devices should be afforded priority protection, since the breaches are severe operational disruptions coupled with data theft [2]. The security challenges associated with IoT are heavily contingent on the dataset imbalances that affect intrusion detection between different methods. Scavenging an anomaly in normal operational data from the malicious represents a greatly unbalanced condition since malfeasance is rarely generated within malefic activity data. The result of this limited, imbalanced condition across classes would severely influence ML models [29], as they could predict results toward the majority class and thus reduce their sensitivity toward rare attacks. IDS effectiveness depends on resolving this imbalance because this merits the ability to detect ordinary and unusual threats [2].

Different strategies aiming to address class imbalance have been studied for IoT intrusion detection systems in recent times [27, 28]. SMOTE proves effective in producing synthetic samples from minority classes which results in dataset balancing for better detection outcomes [3]. ADASYN continues the method of synthetic data creation especially for challenging examples to strengthen system reliability [4]. The application of these resampling methods together with the ensemble classifiers including RF, Gradient Boosting, XGBoost, LightGBM, and CatBoost shows great potential to enhance IDS performance in IoT environments as per research studies [2, 4]. We expand current progress in this study by conducting a systematic evaluation that measures SMOTE and ADASYN resampling methods and their impact on diverse ensemble classifiers. The proposed framework combines state-of-the-art ML algorithms with the solution of class imbalance problems [5] to develop a robust system for IoT intrusion detection which can achieve strong detection rates for all classes to secure IoT networks properly. To improve IoT intrusion detection systems, the primary goal of this study is to create a novel technique (Figure 1) that combines SMOTE and ADASYN with ensemble classifiers. Our work systematically evaluates multiple methods for class imbalance management and detection enhancement while proving their effectiveness for improving accuracy rates. The experimental results achieve their performance metrics by utilizing standard evaluation methods where accuracy and recall and precision along with F1-score metrics represent our original findings in the field. The research develops a total security framework for defeating multiple cyber dangers that affect IoT networks.

2. RELATED WORK

The research on IoT security now emphasizes the improvement of intrusion detection through the resolution of class imbalance issues [6]. Resampling methods SMOTE and ADASYN rose in popularity because they can create synthetic minority class samples for dataset balancing purposes which leads to model performance improvements [7]. The utilization of ensemble classifiers namely XGBoost and RF with these methods, leads to substantial increases in detection rates for unusual attacks according to research studies [8, 9]. A wide range of research has proved that SMOTE which works in multiclass classification and enhances IoT accuracy [8] while ADASYN demonstrates its effectiveness by focusing on difficult-to-teach samples to improve model stability [10, 11]. The research of [12] presented findings about combining LightGBM and CatBoost algorithms with resampling methods when dealing with high-

dimensional IoT data. Results of several investigations indicate the critical function of resampling strategies in the battle against class imbalance during the development of IoT security frameworks. The prospects of fusing resampling with feature selection and deep learning in addressing particular challenges faced in IoT datasets. The authors at [13] created a hybrid system combining SMOTE with convolutional neural networks (CNNs) to boost the detection of minority class attacks in the IoT networks [14]. The authors of [15] incorporated a feature selection technique with ADASYN to reduce the computational load, thus achieving their aim without compromise to detection performance as per [16]. In IoT security, resampling techniques have gained great popularity along with advanced algorithms for handling problems related to imbalanced data. In this study, with the added impetus of scholarly work of the utmost significance in this area, new objectives are established for an impartial evaluation of the SMOTE and ADASYN algorithms, used in conjunction with contemporary ensemble classifiers, meant to bolster the IoT anomaly detection systems.

3. RESEARCH METHODS

The research methodically addresses class imbalance in IoT cybersecurity dataset through the analysis of the dataset, namely "IoT_Modbus". This phase analyzes the first stage of data preparation, which most importantly contains preprocessing all raw data to make it suitable for all later analytical assessment. This phase also eliminates all non-hoarded timestamps in order to bring numerical values into a standard scale for models that learn by machine. Finally, the particular forms of attack shown by the target variable were transformed to numerical types through label encoding for the multiclass classification. The procedure made the dataset ready for analysis by establishing its consistent format. SMOTE and ADASYN operated as advanced resampling approaches that addressed the class imbalance problem. The minority class in SMOTE receives synthesized records through data point interpolation and ADASYN produces synthetic data elements targeting difficult learning components to boost generalization capabilities of the model [17]. Data balancing techniques proved effective by creating a more proper distribution of specimens suitable for training multivariate analytical models.

A group of five modern ensemble classifiers known as RF, Gradient Boosting, XGBoost, LightGBM, and CatBoost underwent assessment on resampled data. The selection of these prediction models occurred because they demonstrated success when used in complex classifications while maintaining their effectiveness with imbalanced datasets. Hyperparameter optimization through grid search with cross-validations allowed for finding the best performance while maintaining avoided overfitting scenarios [18]. An evaluation process employed accuracy scores together with confusion matrices and classification reports for a complete assessment of the model performance. Visual presentations were essential for understanding the analysis outcomes. The model comparison required Combined accuracy plots along with consolidated confusion matrix displays from each resampling strategy to track each class classification outcomes. All graphical outputs received high-resolution format preservation for true representation and result verification. The defined methodology stands as both research-intensive and

practical thereby providing essential details to enhance IoT security system performance. The research method is shown diagrammatically in Figure 1.

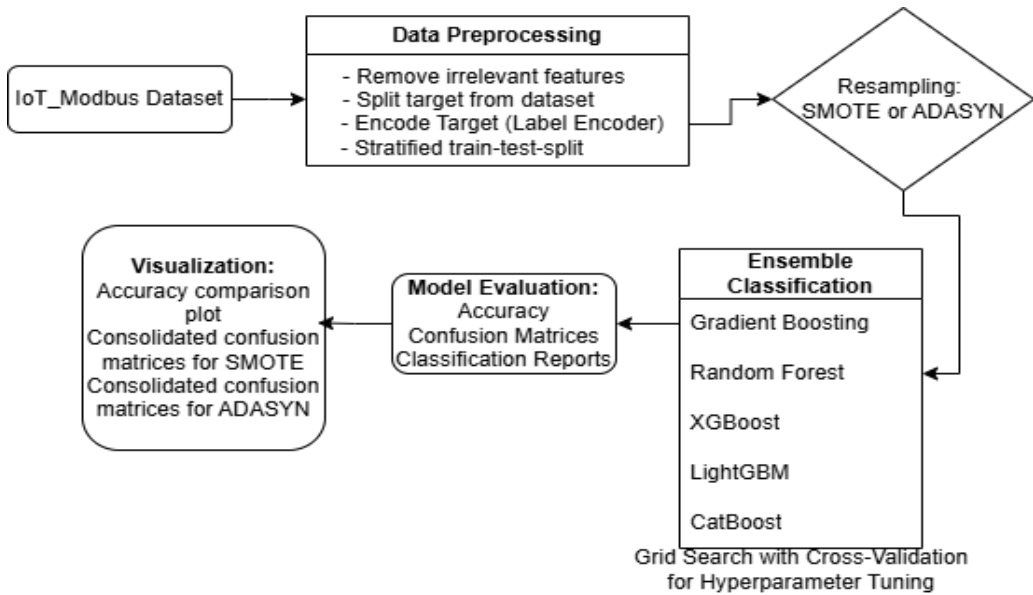


Figure 1. The implemented research methodology

4. RESULTS AND DISCUSSIONS

The assessment of model performance through resampling techniques involved multiple classifiers testing with SMOTE and ADASYN. The researchers used visual bar chart comparison and heatmaps to monitor accuracy changes of different models examined through the study. The stability and pattern changes that take place when classifiers are applied to unbalanced datasets are crucially illustrated by these visual displays. Both measurement stability and accuracy trends for each classifier across resampling techniques are displayed in this heatmap, and the bar chart shows direct evaluations of classifier performance between SMOTE and ADASYN. These analytical instruments enhance the perception of how data resampling approaches affect prediction accuracy by helping users select optimized models [19-20]. RF maintained its leading position in accuracy measurements regardless of utilizing either SMOTE or ADASYN resampling methods which indicates its strong capability to deal with unbalanced data. The boosting-based classifiers comprised Gradient Boosting and XGBoost and LightGBM and CatBoost while XGBoost displayed slightly better achievement than the others yet CatBoost achieved the lowest accuracy level. Both resampling methods approved similar levels of accuracy stability regardless of the testing technique being SMOTE or ADASYN. Among different boosting methods, RF proved itself as the most dependable classifier for this particular task.

4.1. Performance Metrics Evaluation

A full performance evaluation of the classifier required calculation of Accuracy and Precision together with Recall and F1-score metrics. Standard mathematical formulations from the literature defined the metrics that originated from the confusion matrix (Equations 1–4). Accuracy tracks how well the classifier operates in identifying correct responses yet Precision shows how well it identifies genuine outcomes among all predictions. The classifier's ability to find all essential positive cases is measured by Recall which operates under the name sensitivity and the F1-score calculates this balance using harmonic means between Precision and Recall values.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision+Recall} \tag{4}$$

While the comparable metrics from the ADASYN resampling method are shown in Table 2, Table 1 gives evaluation metrics for classifiers under SMOTE resampling. To comprehend the distinct advantages and disadvantages of each approach during assessment, the analysis compares the classifier performances of SMOTE and ADASYN. The outcome of classification depends on different resampling approaches according to evaluation metrics thus identifying models better suited to manage imbalanced datasets. The thorough evaluation process demonstrates both the robust capabilities of particular classifiers as well as it provides real-world knowledge for picking and enhancing models that work with imbalanced datasets.

Table 1. Utilising SMOTE Resampling to Measure Classifier Performance

Classifier	Accuracy (%)	Precision (Avg) (%)	Recall (Avg) (%)	F1-score (Avg) (%)
RF	99	70	76	72
Gradient Boosting	82	36	46	30
XGBoost	86	41	60	38
LightGBM	88	43	43	41
CatBoost	83	36	30	30

Table 2. Classifier Performance Metrics Using ADASYN Resampling

Classifier	Accuracy (%)	Precision (Avg) (%)	Recall (Avg) (%)	F1-score (Avg) (%)
RF	99	69	75	71
Gradient Boosting	82	36	44	30
XGBoost	86	41	57	39
LightGBM	88	42	42	40
CatBoost	83	36	40	31

4.2. Key Findings and Insights

RF demonstrates superiority over alternative models by delivering maximum accuracy results through the use of SMOTE and ADASYN indicating its powerful ability to handle class unbalance issues [21-22]. XGBoost demonstrates stable performance among boosting-based classifiers through its regular results when using both SMOTE and ADASYN resampling techniques. The lower accuracy performance of CatBoost shows that this model may not work efficiently for this classification project with this given dataset. SMOTE and ADASYN demonstrate equivalent effects on classifier behaviors which emerge from their similar performance patterns. The results demonstrate that selecting proper combination of model and resampling technique provides optimal outcomes for working with imbalanced IoT cybersecurity datasets. According to the study ensemble approaches such as RF and XGBoost show better resistance against data imbalance than single classifier systems making them strong candidates for real IoT security work. Researches could develop hybrid detection strategies that link ensemble with resampling techniques to achieve superior detection precision. The research data established the requirement for model adaptation that addresses distinctive characteristics of imbalanced cybersecurity datasets.

4.3. Future Directions and Considerations

Advancement in research should examine other resampling techniques including Random Oversampling, NearMiss and Tomek Links to achieve superior performance on imbalanced datasets. Using advanced ensemble learning techniques, as well as optimized hyperparameter tuning, with refined feature engineering would boost classifier robustness and generalization [23-24]. The model performance using SMOTE and ADASYN appears in heatmap confusion matrices of Figures 2 and 3 and Figure 4 displays performance variations across resampling techniques.

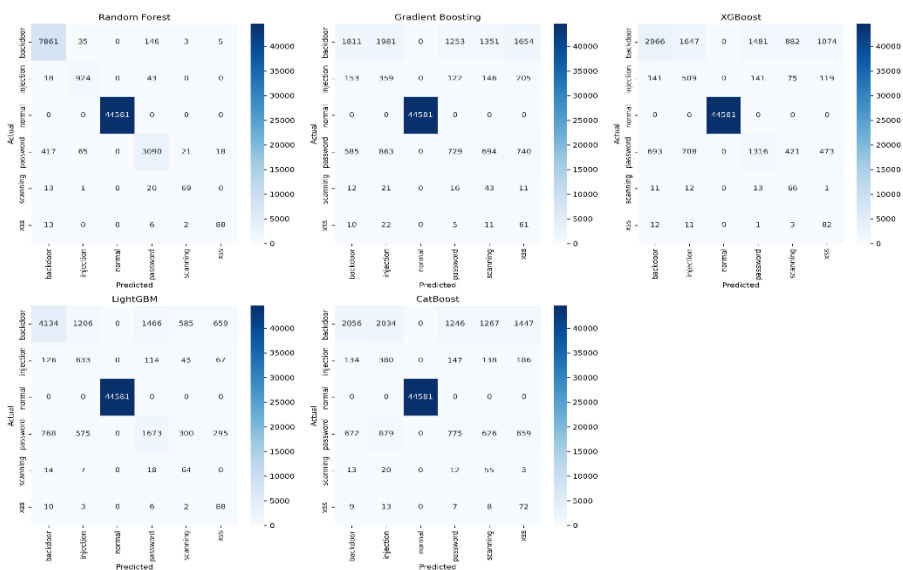


Figure 2. Models Performance Using SMOTE Resampling

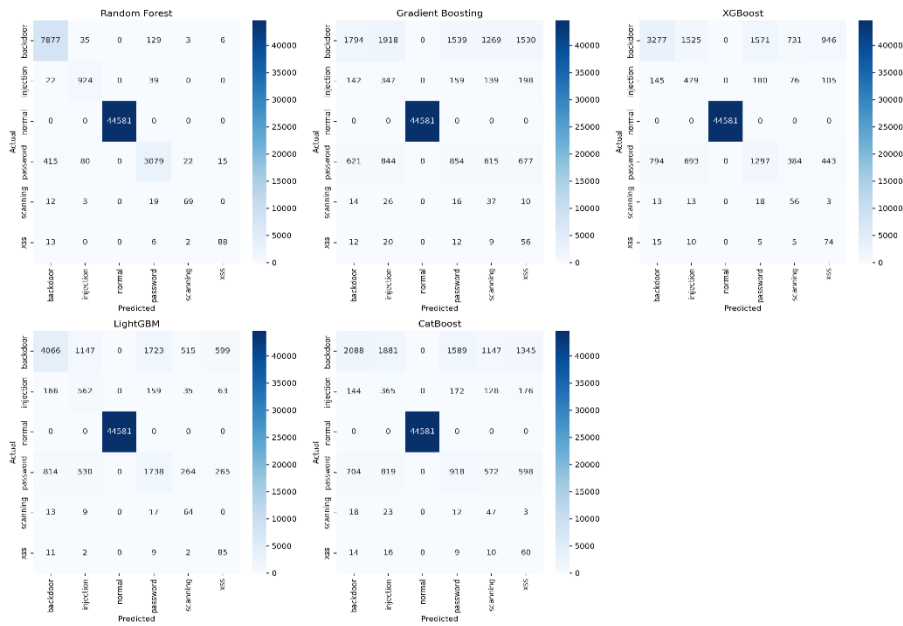


Figure 3. Models' Performance Using ADASYN Resampling

Future research should examine how these methods are applied in operational IoT contexts to evaluate performance scalability and operational efficiency in real time [25]. The evaluation of mixed techniques which integrate deep learning methods together with sampling approaches should be researched to deal with complex and changing security threats. Security frameworks in dynamic interactive environments will achieve additional strengthening through these proposed steps.

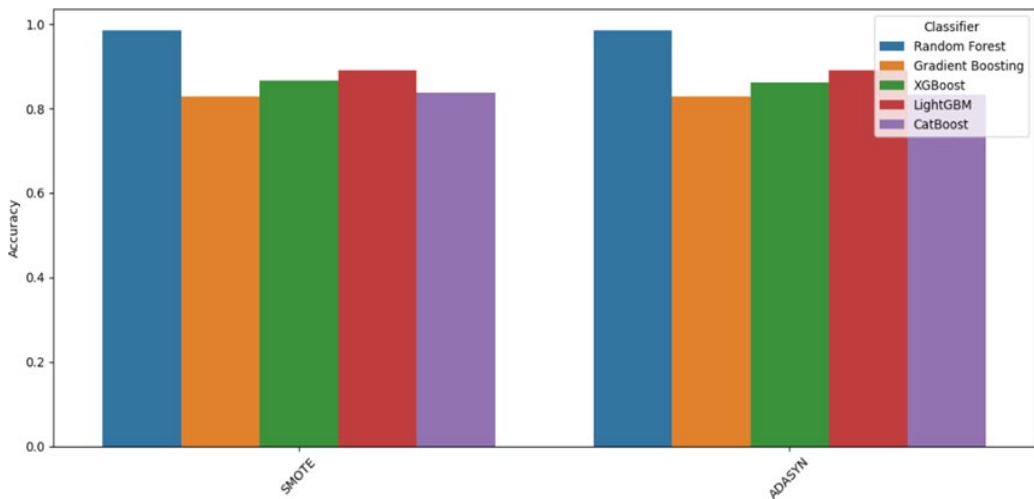


Figure 4. Comparison of classifier performance across the developed models.

5. CONCLUSIONS

This study examined how the performance of ensemble classifiers used to identify IoT anomalies was affected by the combination of SMOTE and ADASYN. The research implemented resampling techniques together with advanced classifiers to manage IoT_Modbus class imbalance thus improving detection accuracy of minority-class threats. The RF approach set itself apart as the premier performer due to its data-imbalanced reliability features although XGBoost along with other boosting models achieved solid outcomes. SMOTE and ADASYN yield equal improvements in classification performance when compared to each other according to the research findings. Future studies will benefit from conducting new sampling techniques [26], in addition to improving mechanisms of feature selection, and adjusting vital system parameters to maximize model effectiveness. Security systems that operationalize these optimized classifiers will boost their ability to detect and counter cyber-attacks in real-life IoT environments. Through the use of data balancing techniques in the creation of strong cybersecurity solutions, this study offers crucial insights that support attempts to enhance intrusion detection systems in IoT contexts.

REFERENCES

- [1] El-Sofany, H., El-Scoud, S. A., Karam, O. H. et al. Using ML algorithms to enhance IoT system security. *Scientific Reports*, vol. 14, no. 12077, 2024. <https://doi.org/10.1038/s41598-024-62861-y>
- [2] Leevy, J. L., Khoshgoftaar, T. M., Peterson, J. M. Mitigating class imbalance for IoT network intrusion detection: A survey. In *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, Oxford, United Kingdom, 2021, pp. 143–148. <https://doi.org/10.1109/BigDataService52369.2021.00023>
- [3] Musthafa, M. B., Huda, S., Kodera, Y., Ali, M. A., Araki, S., Mwaura, J., Nogami, Y. Optimizing IoT intrusion detection using balanced class distribution, feature selection, and ensemble ML techniques. *Sensors*, vol. 24, no. 13, 2024, p.4293. <https://doi.org/10.3390/s24134293>
- [4] Mazhar, T., Talpur, D. B., Shloul, T. A., Ghadi, Y. Y., Haq, I., Ullah, I., Ouahada, K., Hamam, H. Analysis of IoT security challenges and its solutions using artificial intelligence. *Brain Sciences*, vol. 13, no. 4, 2023, p. 683. <https://doi.org/10.3390/brainsci13040683>
- [5] Maria Vlahova-Takova, Milena Lazarova. CNN based multi-label image classification for presentation recommender system. *International Journal on Information Technologies and Security*, vol.16, no.4, 2024, pp. 73-84. <https://doi.org/10.59035/PUYE7368>
- [6] Tahir, U., Abid, M. K., Fuzail, M., Aslam, N. Enhancing IoT security through ML-driven anomaly detection. *VFAST Transactions on Software Engineering*, vol. 12, no. 2, 2024, pp.1-13. <https://doi.org/10.21015/vtse.v12i1.1766>
- [7] Peng, H., Wu, C., Xiao, Y. CBF-IDS: Addressing class imbalance using CNN-BiLSTM with focal loss in network intrusion detection system. *Applied Sciences*, vol. 13, no. 21, 2023, p.11629. <https://doi.org/10.3390/app132111629>

- [8] Zhang, Y., et al. An efficient large-margin multiclass boosting framework for imbalanced data classification. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, 2021, pp. 3125–3138.
- [9] Qawqzeh, Y. K., Ashraf, M. A fraud detection system using decision trees classification in an online transaction. In *Proceedings of the 2023 12th International Conference on Software and Computing Applications (ICSCA '23)*, New York, NY, USA, 2023, pp. 212–217.
- [10] Wang, Z., et al. A Hybrid SMOTE-CNN framework for IoT anomaly detection. *IEEE Internet of Things Journal*, vol. 10, no. 4, 2023, pp. 5678–5690.
- [11] Varotto, G., Susi, G., Tassi, L., Gozzo, F., Franceschetti, S., Panzica, F. Comparison of resampling techniques for imbalanced datasets in ML: Application to epileptogenic zone localization from interictal intracranial EEG recordings in patients with focal epilepsy. *Frontiers in Neuroinformatics*, vol. 15, 2021. <https://doi.org/10.3389/fninf.2021.715421>
- [12] Koziarski, M., Woźniak, M., Krawczyk, B. Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise. *Knowledge-Based Systems*, vol. 204, 2020, p.106223. <https://doi.org/10.1016/j.knosys.2020.106223>
- [13] Waleed, O., Ali, N. The effects of resampling on classifying imbalanced datasets. *IEEE Xplore*, vol. 204, 2-22, pp. 1–6.
- [14] Tanha, J., Abdi, Y., Samadi, N. et al. Boosting methods for multi-class imbalanced data classification: An experimental review. *Journal of Big Data*, vol. 7, no. 70, 2020. <https://doi.org/10.1186/s40537-020-00349-y>
- [15] Ahmed, S., Khan, M. A., Rehman, A., Alazab, M. Feature selection and ADASYN for efficient IoT intrusion detection. *Computers & Security*, 114, 2022, pp. 102–115. <https://doi.org/10.1016/j.cose.2022.102115>
- [16] Radi Romansky. Networking and need of complex literacy in the digital age. *International Journal on Information Technologies and Security*, vol.16, no.4, 2024, pp. 3-14. <https://doi.org/10.59035/VPKK5580>
- [17] Khan, A., Malim, A. H. Comparative studies on resampling techniques in ML and deep learning models for drug-target interaction prediction. *Molecules*, vol. 28, no. 4, 2023, p.1663. <https://doi.org/10.3390/molecules28041663>
- [18] Zhang, Y., Li, X., Liu, J., & Wang, Z. An efficient large-margin multiclass boosting framework for imbalanced data classification. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, 2021, pp. 3125–3138.
- [19] Collell, G., Prelec, D., Patil, K. R. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*, vol. 275, 2018, pp. 330–340. <https://doi.org/10.1016/j.neucom.2017.08.035>
- [20] Qawqzeh, Y., Reaz, M., Ali, M., Gan, K., Zulkifili, Z., Noraidatulakma, A. Assessment of atherosclerosis in erectile dysfunction subjects using second derivative of photoplethysmogram. *Scientific Research Essays*, vol. 7, 2012. https://academicjournals.org/article/article1380790451_Qawqzeh%20et%20al.pdf
- [21] Rezvani, S., Wang, X. A broad review on class imbalance learning techniques. *Applied Soft Computing*, vol. 43, 2023, p. 110415. <https://doi.org/10.1016/j.asoc.2023.110415>

- [22] Qawqzeh, Y., Alourani, A., Ghwanmeh, S. An improved breast cancer classification method using an enhanced Adaboost classifier. *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 1, 2023. <https://doi.org/10.14569/IJACSA.2023.0140151>
- [23] Yang, H., Xu, J., Xiao, Y., Hu, L. SPE-ACGAN: A resampling approach for class imbalance problem in network intrusion detection systems. *Electronics*, vol. 12, no. 15, 2023, p. 3323. <https://doi.org/10.3390/electronics12153323>
- [24] Ootom, M. M., Jemmali, M., Qawqzeh, Y., Sa, K. N., Al Fay, F. Comparative analysis of different machine learning models for estimating the population growth rate in data-limited area. *International Journal of Computer Science and Network Security*, vol. 19, 2019, p.96.
- [25] Zhao, Z., Cui, T., Ding, S., Li, J., Bellotti, A. G. Resampling techniques study on class imbalance problem in credit risk prediction. *Mathematics*, vol. 12, no. 5, p. 2024.
- [26] Krishnan, D., Shrinath, P. Enhancing energy efficiency and imbalance handling in botnet detection in IoT networks: a multi-stage feature reduction and weighted approach. *International Journal of Information Technology*, vol. 17, 2015, pp.811–822. <https://doi.org/10.1007/s41870-024-02219-9>.
- [27] Alharbi, F., Ouarbya, L., Ward, J. A. Comparing sampling strategies for tackling imbalanced data in human activity recognition. *Sensors*, vol.22, no. 4, 2022 <https://doi.org/10.3390/s22041373>
- [28] Alharbi, F., Ouarbya, L., Ward, J. A. Synthetic sensor data for human activity recognition. *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, 2020, pp. 1-9, doi: 10.1109/IJCNN48605.2020.9206624.
- [29] Alharbi, F., Farrahi, K. A convolutional neural network for smoking activity recognition. *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, Ostrava, Czech Republic, 2018, pp. 1-6, doi: 10.1109/HealthCom.2018.8531148.

Information about the author:

Fayez Alharbi – Fayez is an Assistant Professor in the IT Department at the College of Computer Sciences and Information Technology, Majmaah University. He earned his PhD in Computer Science from the University of London – Goldsmiths in 2022, building on his academic foundation with a Master of Science in Networking and Systems Administration from the Rochester Institute of Technology (2013) and a Bachelor of Science in Computer Information Technology from Indiana University Purdue University (2010). Since the beginning of his career Fayez dedicated both his research and professional practice to data science and machine learning which specifically involves deep learning and generative models and the Internet of Things (IoT). The expertise of Fayez appears through his multiple international journal and conference publications which demonstrate his dedication to forward progress in advanced technologies. Through his consultant work at artificial intelligence and data science and digital transformation projects Fayez assists organizations to use technology for innovative growth solutions.

Manuscript received on 12 February 2025