

Article

Performance evaluation of supervised machine learning algorithms for diabetes prediction

Babagana Modu^{1*} and Kale Kawu Kale²

¹ Department of Mathematics and Statistics, Yobe State University, Damaturu, Nigeria

² Department of Mathematics and Statistics Computational Laboratory, Yobe State University, Damaturu, Nigeria; kalekk@gmail.com

* Correspondence: modubabagana70@yahoo.com

Abstract: This study presents a performance evaluation of six machine learning algorithms for the prediction of diabetes using a publicly available dataset from Kaggle, which includes relevant clinical and demographic features. Comprehensive preprocessing procedures were undertaken to address missing values, outliers, and a pronounced class imbalance (38:1 ratio of non-diabetic to diabetic cases), which poses significant challenges to model performance. The evaluated algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, and Artificial Neural Network (ANN)—were assessed using key performance metrics, including balanced accuracy, precision, recall, F1-score, sensitivity, specificity, and detection rate. Among the models, Random Forest achieved the highest balanced accuracy (93%), followed by SVM (83%) and Decision Tree (82%), demonstrating superior predictive performance. The findings underscore the potential of machine learning in enhancing diabetes diagnosis and management and conclude that a careful balance of performance metrics, especially when dealing with imbalanced healthcare datasets should guide model selection.

Keywords: machine learning; diabetes prediction; random forest; class imbalance; performance evaluation.

Received: 25 Sept. 2025; Revised: 18 November 2025; Accepted: 10 January 2026; Published: 27 January 2026



Copyright: ©2026 the Author(s). Published by JSSCI. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Journal Abbreviation: J. Stat. Sci. Comput. Intell.

1. Introduction

Recent advancements in data science and machine learning have had a profound impact on healthcare, particularly in the realms of early disease detection and personalized treatment. Supervised machine learning algorithms are increasingly employed to predict diabetes by analyzing a variety of patient data, including demographic information, clinical history, and behavioral patterns. Prominent models utilized in diabetes prediction include Decision Trees, Support Vector Machines (SVM), Logistic Regression, k-Nearest Neighbors (k-NN), Random Forests, and Neural Networks, each offering distinct advantages in terms of predictive accuracy, interpretability, and computational complexity. The selection of an appropriate algorithm must be

guided by the specific characteristics of the dataset and the trade-off between model performance and interpretability. Furthermore, evaluating the performance of these models through metrics such as accuracy, precision, recall, F1-score, and Area Under the Curve (AUC) is essential for assessing their robustness and applicability in clinical settings.

The application of machine learning (ML) for diabetes prediction has emerged as a burgeoning area of research, with numerous studies assessing the efficacy of various ML models in forecasting diabetes and its associated complications. [1] demonstrated the efficacy of ten ML algorithms using a dataset from Frankfurt Hospital, where the Gaussian Process (GP) classifier achieved superior performance with an accuracy of 98%. Other models, including Random Forest (RF) and Artificial Neural Networks (ANN), also exhibited high accuracy, emphasizing the potential of ML techniques to reduce mortality rates, particularly in developing nations. Similarly, [2] compared the performance of K-Nearest Neighbors (KNN) and Naive Bayes algorithms, with Naive Bayes achieving an accuracy of 76%, underscoring the importance of selecting suitable algorithms for distinct datasets. In a study by [3], three ML algorithms—Gaussian Naïve Bayes (GNB), Linear Support Vector Machine (LSVM), and Random Forest (RF)—were evaluated, with LSVM delivering the highest accuracy of 78%. [4] explored various models, finding that a Neural Network (NN) with two hidden layers attained an accuracy of 89%. A systematic review conducted by [5], encompassing 32 studies, concluded that while Random Forest consistently outperformed other models, further external validation is essential prior to clinical implementation. This view was similarly reflected by [6], who identified Random Forest as the most accurate model, achieving an accuracy of 77%. [7] evaluated multiple algorithms across two datasets, showing that Support Vector Machine (SVM) performed best on the Pima Indian Diabetes dataset (74%), whereas KNN and RF excelled with the German dataset (99%), highlighting the impact of dataset characteristics on model performance. In support of these findings, [8] reaffirmed the dominance of Random Forest, achieving an accuracy of 88%. Furthermore, [9] demonstrated that Random Forest and Naïve Bayes performed consistently well across both the Pima and Bangladeshi datasets, illustrating the versatility of these models across diverse healthcare settings. [10] suggested that ontology-based classifiers and SVM exhibited strong performance, adding an additional layer of complexity to algorithm selection. [11] developed a smart web application integrating multiple ML algorithms, significantly enhancing prediction accuracy through appropriate pre-processing techniques. [12] further advanced this approach by combining SVM and ANN models, achieving an impressive 95% prediction accuracy. Additionally, [13][14] underscored the effectiveness of Random Forest in classification accuracy, particularly highlighting the advantages of flexible ML methods such as XGBoost and feedforward networks for specific outcomes. In conclusion, while Random Forest, Naïve Bayes, and SVM models consistently demonstrate superior performance in diabetes prediction, the selection of the optimal algorithm is contingent on the specific dataset and prediction objectives. This study is necessary to validate these models in real-world clinical settings and to explore the integration of multiple models, thereby improving prediction accuracy and contributing to personalized healthcare solutions.

Despite the growing use of supervised machine learning techniques in diabetes prediction (e.g., [1]–[3], [10]–[15]), a significant research gap remains in the comprehensive evaluation of these models across varying healthcare domains. Existing studies [7]–[9] often overlook the critical impact of dataset characteristics—such as data imputation, outlier treatment, and class imbalance—on model performance. Hence, this study addresses these limitations that influence the effectiveness of supervised learning algorithms and evaluates their computational performance in real-world healthcare scenarios. The findings contribute to the development of more robust and generalizable predictive models for diabetes diagnosis.

The remaining sections of this paper are organized as follows: Section 2 outlines the data sources and methodological framework; Section 3 presents and interprets the analytical results; and Section 4 summarizes

the key findings and suggests directions for future research.

2. Methodology

This section provides a detailed description of the materials and methodologies utilized for data preprocessing, the development of supervised learning models, and the subsequent evaluation of their performance.

2.1. Source of data

The dataset utilized in this study was obtained from the Kaggle platform via the following link <https://www.kaggle.com/datasets>, a reputable and widely used online repository that hosts a broad range of open-access datasets for data science and machine learning research. Specifically, the selected dataset contains clinically relevant features associated with diabetes, including demographic variables (such as age and gender), lifestyle indicators, and key physiological measurements (e.g., glucose levels, blood pressure, and body mass index). These attributes make the dataset particularly suitable for developing and evaluating supervised machine learning models aimed at predicting diabetes outcomes. The use of a publicly available and well-documented dataset ensures reproducibility and generalizability of the results.

2.2. Preprocessing

To develop a robust and effective supervised learning model, it is imperative to perform comprehensive data preprocessing. This process involves several critical steps, including the identification and treatment of outliers to mitigate their impact on model accuracy, imputation of missing values to preserve data integrity, addressing class imbalance to ensure fair and unbiased predictions, and partitioning the dataset into training and testing subsets to enable reliable model evaluation and generalization.

2.2.1. Expository Data Analysis

Outliers were identified through the boxplot illustrated in Figure 1, with notable prominence observed in variables such as cholesterol and glucose levels relative to other features. To mitigate the influence of these outliers, the Z-score method was applied [15]. Furthermore, approximately 14% of the dataset contained missing values, which were addressed using the Multiple Imputation by Chained Equations (MICE) technique [15].

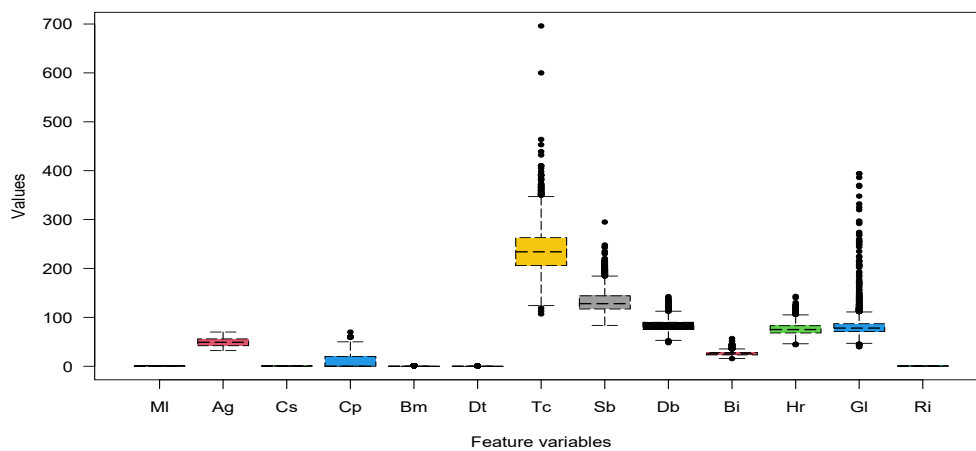


Figure 1. Boxplot for outlier detection.

Table 1. Numerical summary with bootstrapped confidence interval for median

Variables	Min	Q1	Median	Bootstrapped CI	Q3	Max
Male (Ml)	0.00	0.00	0.00	0.00 – 0.00	1.00	1.00
Age (Ag)	32.00	42.00	49.00	48.00 – 49.00	56.00	70.00
currentSmoker (Cs)	0.00	0.00	0.00	0.00 – 1.00	1.00	1.00
cigsPerDay (Cp)	0.00	0.00	0.00	0.00 – 1.00	20.00	70.00
BPMeds (Bm)	0.00	0.00	0.00	0.00 – 0.00	0.00	1.00
diabetes (Dt)	0.00	0.00	0.00	0.00 – 0.00	0.00	1.00
totChol (Tc)	107.00	206.00	234.00	232.00 – 235.00	263.00	696.00
sysBP (Sb)	83.50	117.00	128.00	128.00 – 129.00	144.00	295.00
diaBP (Db)	48.00	75.00	82.00	81.00 – 82.00	90.00	142.50
BMI (Bi)	15.54	23.07	25.40	25.26 – 25.54	28.04	56.80
heartRate (Hr)	44.00	68.00	75.00	75.00 – 75.00	83.00	143.00
glucose (Gl)	40.00	71.00	78.00	78.00 – 78.00	87.00	394.00
Risk (Ri)	0.00	0.00	0.00	0.00 – 0.00	1.00	1.00

2.2.2. Class Imbalanced

Class imbalance arises in binary classification tasks when one class is substantially more prevalent than the other, often resulting in models biased toward the majority class. This imbalance can lead to deceptively high accuracy while yielding poor recall for the minority class. The degree of imbalance is commonly quantified using the Imbalance Ratio (IR) [16].

$$IR = \frac{N_{\beta}}{N_{\alpha}} \tag{1}$$

where N_{α} and N_{β} represent the number of samples in the minority and majority classes, respectively.

2.3. Algorithms Performance Index

In the context of imbalanced classification problems, the performance metrics presented in Table 2 will be utilized to assess the effectiveness of the supervised learning models examined in this study. These metrics are especially appropriate for evaluating model performance under unequal class distributions, as they ensure a comprehensive assessment of both minority and majority class predictions.

Table 2. Algorithms performance expression

Measures	Mathematical Expression	Definition and Result Interpretation
Sensitivity	$\frac{TP}{(TP + FN)}$	The proportion of actual positive cases that are correctly identified by the model
Specificity	$\frac{TN}{(TN + FP)}$	Measures the model's ability to correctly identify negative cases
Precision	$\frac{TP}{(TP + FP)}$	High precision indicates a low rate of false positive predictions
Recall	$\frac{TP}{(TP + FN)}$	A high recall value indicates that the model produces a low number of false negatives
F1-Score	$2(\text{Recall}/\text{Recall} + \text{Prec.})$	The trade-off between recall and precision to enhance overall model performance
Detection Rate	$\frac{TP}{(TP + FN)}$	High detection rate reflects strong model performance in recognizing true positives
Balance Accuracy	$(\text{Sens.} + \text{Spec.})/2$	The metric evaluates performance across both classes

2.4. Computational Software

The R statistical software environment, supplemented with specialized packages including *caret*, *randomForest*, and *e1071*, was employed to implement, train, and evaluate a range of supervised machine learning models for the purpose of diabetes prediction. These packages provided robust tools for data preprocessing, model tuning, performance assessment, and cross-validation, thereby ensuring methodological rigor and reproducibility in the analytical workflow [18].

3. Results Presentation and discussion

In this section, we present the results of our analysis and provide a detailed discussion of the key findings, highlighting their significance, implications.

The dataset underwent a comprehensive preprocessing phase to ensure data quality and enhance the reliability of subsequent analyses. Initially, the dataset was examined for missing values and potential outliers. A total of 580 missing entries were identified, representing approximately 14% of the entire dataset. To address this issue, missing values were imputed using the Multiple Imputation by Chained Equations (MICE) method, which is a robust statistical technique that creates multiple plausible estimates for missing data based on the relationships among observed variables. This approach helps to preserve the inherent structure and variability within the dataset [15]. Outlier detection and treatment were also conducted to minimize their adverse effects on model training and prediction. Identified outliers were standardized using the Z-score method, which transforms data values based on the mean and standard deviation, allowing for the detection of extreme values that deviate significantly from the normal distribution.

In addition, the dataset was found to exhibit a substantial class imbalance, with an imbalance ratio (IR) of 38:1. In this context, the majority class (label = 0) represents individuals without diabetes, whereas the minority class (label = 1) represents individuals diagnosed with diabetes. This high degree of imbalance poses a significant challenge to the performance of supervised learning algorithms, as it can lead to biased predictions that favor the majority class while underrepresenting the minority class. Consequently, models trained on imbalanced data may achieve high overall accuracy while performing poorly in detecting the minority class—diabetic cases—that is of primary interest in medical diagnosis. To facilitate model training and evaluation, the preprocessed dataset was partitioned into training and testing subsets in an 80:20 ratio. This data split ensures that the model is trained on a sufficient number of observations while preserving an independent test set for unbiased performance assessment. In the context of supervised machine learning, especially for imbalanced classification problems, traditional performance metrics such as overall accuracy can be misleading, as they do not account for the distribution of class labels [19]. To provide a more comprehensive evaluation of model performance, several specialized metrics are employed—such as precision, recall, F1-score, sensitivity, specificity, balance accuracy, and detection rate—as detailed in Table 3.

Table 3. Supervised learning algorithms performance evaluation.

Algorithms Performance Metrics	ANN	DT	LoR	NB	RF	SVM
Sensitivity	0.5227	0.6477	0.5000	0.3750	0.8625	0.6932
Specificity	0.9994	0.9973	0.9991	0.9876	0.9981	0.9699
Precision	0.9583	0.8636	0.9362	0.9987	0.8649	0.9959
Recall	0.5227	0.6477	0.5000	0.3750	0.7520	0.6932
F1-Score	0.6765	0.7403	0.6519	0.5455	0.8660	0.8188
Detection Rate	0.0136	0.0168	0.0130	0.0097	0.0259	0.0180

Balance Accuracy	0.7611	0.8225	0.7496	0.6813	0.9303	0.8316
------------------	--------	--------	--------	--------	--------	--------

The performance of six supervised machine learning algorithms—Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, and Artificial Neural Network (ANN)—was systematically evaluated using a comprehensive set of performance metrics. These metrics included balanced accuracy, precision, recall, F1-score, sensitivity, specificity, and detection rate. Such metrics were selected to ensure a robust and meaningful assessment of model effectiveness, particularly given the imbalanced nature of the dataset used for diabetes prediction. Among the evaluated models, the Random Forest classifier demonstrated the highest overall performance, achieving a balanced accuracy of 93% (see Figure 2). The Support Vector Machine (SVM) followed this and Decision Tree classifiers, which attained balanced accuracy scores of 83% and 8%, respectively. These results highlight the superior capability of ensemble methods, such as Random Forest, in handling complex and high-dimensional healthcare data, particularly under conditions of class imbalance. The superior performance of Random Forest can be attributed to its ability to aggregate the outputs of multiple decision trees, thereby reducing over fitting and improving generalization. In contrast, while models such as Naive Bayes and Logistic Regression offered relatively faster training times and interpretability, their predictive performance was comparatively lower, especially in detecting the minority (diabetic) class. These findings reinforce the growing potential of machine learning techniques in supporting the early diagnosis and effective management of diabetes.

Importantly, the study underscores that model selection in medical prediction tasks should not rely solely on a single metric such as overall accuracy. Instead, it should be guided by a holistic evaluation of multiple performance measures, especially in the presence of imbalanced datasets where underrepresentation of critical classes—such as patients with diabetes—can lead to suboptimal clinical decision-making. Ultimately, careful model selection and validation are essential to ensure that machine-learning tools contribute reliably and ethically to healthcare delivery.

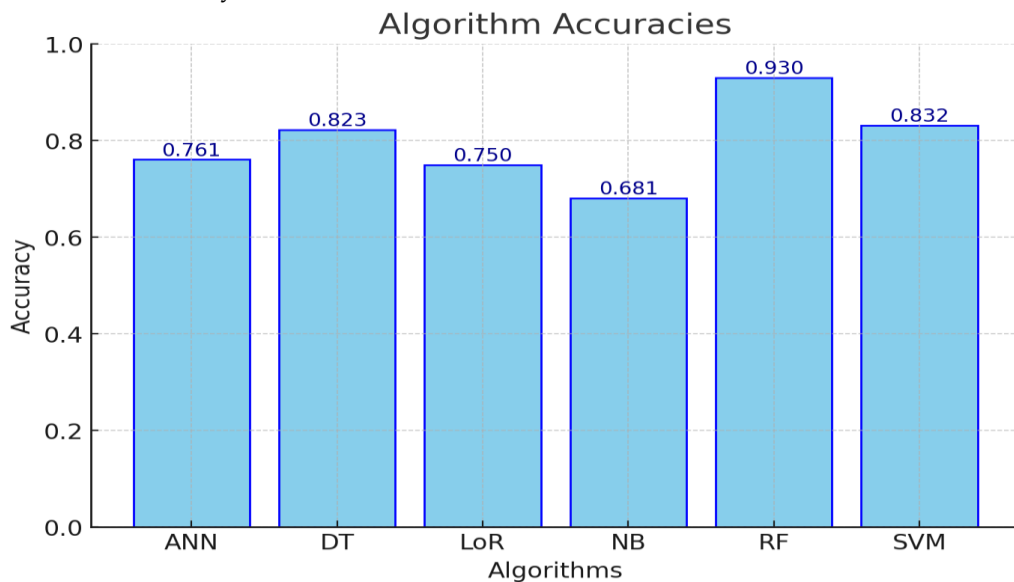


Figure 2. Algorithms vs performance accuracies.

4. Conclusions

This study evaluated the performance of six supervised machine learning models to predict diabetes using a medical dataset. Among the models investigated, Random Forest and Support Vector Machine demonstrated superior accuracy and generalization, emphasizing the potential of machine learning techniques for early

disease diagnosis and the reduction of healthcare costs through preventive interventions. However, the study is subject to several limitations, including the use of a small dataset that may not be representative of the broader population, the exclusion of additional relevant clinical features, and reliance on a single dataset, which raises concerns about potential overfitting and bias. Furthermore, algorithm performance may vary across different datasets or real-world contexts. Future research should incorporate cross-validation across diverse healthcare settings and regions to better evaluate the robustness of these models. Additionally, collaboration with real-time clinical data and longitudinal patient information could enhance the accuracy and personalization of diabetes risk predictions.

Author contributions: Conceptualization was carried out by K.K.; B.M. performed methodology and data analysis; manuscript review and editing were undertaken by B.M.

Funding Statement: This research received no external funding.

Data Availability: The data that support the findings are available from <https://www.kaggle.com/datasets>.

Acknowledgments: The authors sincerely thank the editorial team for their valuable guidance, constructive feedback, and efficient handling of the manuscript throughout the review process.

Conflict of interest: The authors declare no conflict of interest.

References

- [1] Nishat, M.M., Faisal, F., Mahbub, M.A., Mahbub, M.H., Islam, S., & Hoque, M.A. (2021). Performance assessment of different machine learning algorithms in predicting diabetes mellitus. *Bioscience, Biotechnology, and Research Communications*, 14(1), 74-82.
- [2] Febrian, M. E., Ferdinan, F. X., Sendani, G. P., Suryanigrum, K. M., & Yunanda, R. (2023). Diabetes prediction using supervised machine learning. *Procedia Computer Science*, 216, 21-30. [\[CrossRef\]](#)
- [3] Assegie, T. A., & Nair, P. S. (2020). The performance of different machine learning models on diabetes prediction. *International journal of scientific & technology research*, 9(01).
- [4] Khanam, J.J., & Foo, S.Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4), 432-439. [\[CrossRef\]](#)
- [5] Tan, K. R., Seng, J. J. B., Kwan, Y. H., Chen, Y. J., Zainudin, S. B., Loh, D. H. F., ... & Low, L. L. (2023). Evaluation of machine learning methods developed for prediction of diabetes complications: a systematic review. *Journal of diabetes science and technology*, 17(2), 474-489. [\[CrossRef\]](#)
- [6] Soni, M., & Varma, S. (2020). Diabetes prediction using machine learning techniques. *International Journal of Engineering Research & Technology (IJERT)*, 9(09), 2278-0181. [\[CrossRef\]](#)
- [7] Kangra, K., & Singh, J. (2023). Comparative analysis of predictive machine learning algorithms for diabetes mellitus. *Bulletin of Electrical Engineering and Informatics*, 12(3), 1728-1737. [\[CrossRef\]](#)
- [8] Nahzat, S., & Yag'anoğlu, M. (2021). Diabetes prediction using machine learning classification algorithms. *Avrupa Bilim ve Teknoloji Dergisi*, (24), 53-59. [\[CrossRef\]](#)
- [9] Pranto, B., Mehnaz, S. M., Mahid, E. B., Sadman, I. M., Rahman, A., & Momen, S. (2020). Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information*, 11(8), 374. [\[CrossRef\]](#)
- [10] El Massari, H., Sabouri, Z., Mhammedi, S., & Gherabi, N. (2022). Diabetes prediction using machine learning algorithms and ontology. *Journal of ICT Standardization*, 10(2), 319-337. [\[CrossRef\]](#)

- [11] Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., & Paul, B. K. (2021). Machine learning based diabetes prediction and development of smart web application. *International Journal of Cognitive Computing in Engineering*, 2, 229-241. [\[11\]](#) [\[SEP\]](#)
- [12] Haque, F., Bin Ibne Reaz, M., Chowdhury, M. E. H., Srivastava, G., Hamid Md Ali, S., Bakar, A. A. A., & Bhuiyan, M. A. S. (2021). Performance analysis of conventional machine learning algorithms for diabetic sensorimotor polyneuropathy severity classification. *Diagnostics*, 11(5), 801. [\[12\]](#) [\[SEP\]](#)
- [13] Li, L., Lee, C. C., Zhou, F. L., Molony, C., Doder, Z., Zalmover, E., ... & Wu, C. (2021). Performance assessment of different machine learning approaches in predicting diabetic ketoacidosis in adults with type 1 diabetes using electronic health records data. *Pharmacoepidemiology and drug safety*, 30(5), 610-618. [\[13\]](#) [\[SEP\]](#)
- [14] Kaur, H., & Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied computing and informatics*, 18(1/2), 90-100. [\[14\]](#) [\[SEP\]](#)
- [15] Modu, B., & Fika, I. A. (2025). Supervised Machine Learning Models for COVID-19 Prediction. *Asian Journal of Probability and Statistics*, 27(3), 13-23. [\[15\]](#) [\[SEP\]](#)
- [16] Wibbeke, J., Rohjans, S., & Rauh, A. (2025). Quantification of Data Imbalance. *Expert Systems*, 42(3), e13840. [\[16\]](#) [\[SEP\]](#)
- [17] Band, S. S., Yarahmadi, A., Hsu, C. C., Biyari, M., Sookhak, M., Ameri, R., ... & Liang, H. W. (2023). Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40, 101286. [\[17\]](#) [\[SEP\]](#)
- [18] Nandan Prasad, A. (2024). Data Quality and Preprocessing. In *Introduction to Data Governance for Machine Learning Systems: Fundamental Principles, Critical Practices, and Future Trends* (pp. 109-223). Berkeley, CA: Apress. [\[18\]](#) [\[SEP\]](#)
- [19] Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513, 429-441.



Disclaimer/Publisher's Note: The views, opinions, and content expressed in all articles are solely those of the respective author(s) and contributor(s) and do not necessarily reflect those of the JSSCI, its editors, or the publisher. JSSCI and its editorial team assume no responsibility for any harm or damage resulting from the use of information, methods, or products mentioned in the publication.