**International Journal of Sciences, Engineering & Environmental Technology (IJOSEET), 16(16): 120-127, 2024**
**ISSN 0749-9650**

www.repcomseet.org

# Comparative Analysis on the Performance of some Supervised Classification Methods

[1]Adekunle Nurudeen Masopa, [2] Saheed Lekan Rasheed, [3]Waheed Yinka Olatidoye, [4]Rasheedat Adenike Ibrahim

[1,3,4]Department of Statistics, Federal Polytechnic Ede, Osun State. Nigeria

[2]Department of Statistics, Federal Polytechnic Ayede, Oyo State. Nigeria

email: Masopaadekunlenurudeen@gmail.com

***Abstract:*** *The goal of this work is to examine the performance of some classification methods under varying partitioning ratio for the training and test samples and the determine the best partitioning ratio for each of the classifiers. Classification methods such as Logistic regression, Linear Discriminant Analysis, K-Nearest Neighborhood and Glmnet were considered for train: test ratio ranging from 50:50 to 90: 10 respectively. Performance assessment relies on indicators such as Correct Classification Rate (CCR), Sensitivity, Specificity, Area Under the Curve (AUC) and Precision. The results of the analysis revealed that the performance improve as the percentage of training dataset increases across the classifiers though with some inconsistencies in performance rating. The performance indicators show that Glmnet outperformed other classifiers while k-NN shows the weakest performance. The Binary Logistic Regression (BLR) outperformed the Glmnet and others for training samples of 60%,70% and 80% for some indicators and clearly across all indicators at 80% and 90% training samples. However, k-NN showed high precision and specificity when the training proportion is 70% with other indicators showing a weak performance.*

*Keywords: Classifiers, Correct Classification Rate, Precision, Sensitivity, Specificity, Train and Test Samples.*

## 1.0 Introduction

Artificial intelligence as area of study consists of machine learning component that provides veritable approaches which focus on developing algorithms that learn from data and make predictions based on available information (Sarker, 2021; Ahmed, Mohamed, Zeeshan, & Dong,2020). Rudin et al. (2022) posited that machine learning rests on statistical principles and techniques to identify patterns and relationships within datasets. The prominent categories include unsupervised and supervised learning algorithms. Naeem, Ali,Anam and Ahmed(2023) asserted that unsupervised learning algorithms interact with unlabeled data to discover patterns or groupings without prior knowledge of outcomes. Nanga et al. (2021) highlighted that unsupervised technique is suitable for clustering similar items and dimensional reduction. Supervised learning algorithms are trained on labeled datasets where the input and output are known (Mahesh, 2020; Sen, Hajra, & Ghosh, 2020; Suyal, & Goyal, 2022). The procedure for supervised learning algorithms makes it suitable for regression and classification tasks. Neu, Lahann, and Fettke, (2022) opined that the goal of regression is to predict continuous values while classification is concerned with identification of specific group for each unit. Classification methods in machine learning are essential techniques used to categorize data into predefined classes based on input features (Chen, Dewi, Huang, & Caraka, (2020). The principal objective of classification is to assign a categorical label to input data (Sen, Hajra, & Ghosh, 2020; Dahouda, & Joe,2021). Various classification algorithms exist, each with unique strengths and applications. Common methods include Binary logistic regression(BLR), Linear discriminant analysis(LDA), K-Nearest Neighbor (k-NN) and Regularized Binary Logistics.

Binary logistic regression is a statistical method used to model the relationship between a set predictors and response variable which can take dichotomous outcomes (Harris, 2021; Srimaneekarn, Hayter, Liu, & Tantipoj,2022). Gomila (2021) posited that binary logistic regression is a generalized version of linear model that predicts the probability of a binary outcome on the basis of one or more predictor variables. The binary logistic regression differs with linear regression which predicts continuous outcomes by estimating the odds of the dependent variable being in one category versus another (Schober & Vetter,2021; Halvorson, McCabe, Kim, Cao, & King, 2022). Saha (2020) posited that the logistic regression model uses the logistic function to transform its output into a probability. Rahim et al. (2023) adopted binary logistic regression on dental study to

evaluate factors influencing hyperglycemia. Moulaei, Sharifi, Bahaadinbeigy, Haghdoost, and Nasiri, (2023) established binary logistic regression proved effective in identifying significant factors associated with hepatitis.

Linear Discriminant Analysis (LDA) is a classification and dimensionality reduction method often used in supervised machine learning (Ali, Hussain, & Abd,2020; Fabiyi, Murray, Zabalza, & Ren, 2021). It aims to find a linear combination of features that best separates multiple classes. Zhao, Zhang, Yang, Zhou and Xu (2024) posited that linear discriminant analysis operates under the assumption of multivariate normality and equal covariance among classes making it effective for problems where these conditions are satisfied. This technique is closely related to Fisher's linear discriminant which focuses on maximizing the distance between class means while minimizing variance within each class. Wahid et al. (2022) pointed out that a major drawback for linear discrimant analysis is its ineffectiveness on small dataset. Tang, Chen and Li (2021) asserted that linear discriminant analysis is capable of adjusting for matching and covariates.

K-Nearest Neighbor (k-NN) is another supervised learning algorithm that performs both classification and regression tasks (Boateng, Otoo & Abaye,2020; Bansal, Goyal & Choudhary 2022). It operates on the principle of proximity to determine the category a new data point based on the classes of its nearest neighbors in the feature space. The k-NN algorithm classifies a new data point by examining the 'k' closest points in the training dataset based on certain metrics (Grover, & Toghi, 2020; Uddin, Haque Lu, Moni & Gide, 2022). Yang and Sung (2023) observed that identification of the k nearest neighbors gives the algorithm insight to adopt voting mechanism for classification by assigning new data to closest class.

Regularized binary logistic regression is an extension of binary logistic regression that incorporates a penalty function to avoid the possible problem of overfitting and improve model generalization (Tian & Zhang, 2022 ; Wang & Thrampoulidis, 2022).Chan et al.(2022) emphasized that the method of regularization of the model is particularly useful when dealing with high-dimensional datasets where the number of predictors exceeds the number of observations or when multicollinearity among predictors is present. Bukhari et al. (2022) mentioned that regularization involves inclusion of a penalty term to the logistic regression loss function to cater for problem of overfitting. A common way of incorporating the penalty function is through addition of absolute values of the coefficients as a penalty term or/and addition of the squared values of the coefficients as a penalty term (Seng & Li, 2022; McDonald & Wang,2024).

Algan and Ulusoy (2021) posited that the quality and quantity of data plays a prominent function in classification tasks. Issues of size of data in terms sample size and dimensionality continue to pose challenges in machine learning. Data partitioning is a crucial step in the machine learning involving division of a dataset into distinct subsets for training, validation and testing. Data partition is a technique often used to optimize information from a given dataset (Mahmud, Huang, Salloum, Emara & Sadatdiynov,2020; Joseph & Vakayil, 2022). This process ensures that models are trained on one portion of the data and tested on different subset. Oymak, Li, & Soltanolkotabi (2021). highlighted that the process of data partition is capable of preventing overfitting and guarantee adequate generalization.

The focus of this study is to examine the performance of some classification techniques under varying partitioning ratio. The rest of this paper is arranged as follow; section 2.0 deals with methodology,3.0 deals with results and discussions, 4.0 presents the conclusions for the study and 5.0 is on recommendation

**2.0 Methodology**

Two datasets namely the heart disease and diabetes data obtained from UCI Machine Learning Repository are used for this study. The heart disease dataset consists of 303 samples with 13 attributes with 165 units with heart disease and 138 units without heart disease. The diabetes dataset consists of 768 samples with 9 attributes with 500 positive and 268 negative cases.

In this study, four classification methods are considered namely Binary Logistic Regression (BLR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (k-NN) and Regularized Binary Logistics (Glmnet). The performance of the classification methods is examined on the basis of some indicators. The performance indices used for this study include Correct Classification Rate (CCR), Sensitivity (SEN), Specificity (SPEC), Precision (PREC), Area Under Curve (AUC), and Balanced Accuracy (BA).

Let TP be the true positive, TN be true negative, FN be false negative and FP be the false positive rates. The estimates of the performance indices can be obtained as follows;

$$\mathrm{CCR} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TN}+\mathrm{TP}+\mathrm{FN}+\mathrm{FP}}$$

$$\mathrm{SEN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$

$$\text{SPEC} = \frac{\text{TN}}{\text{TN+FP}}$$
$$\text{BA} = \frac{\text{SEN+SPEC}}{2}$$
$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

## 3.0 **Result and Discussions**

The results presented in Table 1 and Table 2 provide a detailed analysis of the performance of various classifiers (Binary Logistic Regression (BLR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (k-NN), and Glmnet) across different data partitioning strategies (50:50, 60:40, 70:30, 80:20, and 90:10). The performance metrics evaluated include Correct Classification Rate (CCR), Sensitivity (SEN), Specificity (SPEC), Precision (PREC), Area Under Curve (AUC), and Balanced Accuracy (BA).

Table 1 shows the results of analysis for the classification methods under the partitioning ratio. For 50:50 split BLR achieves a CCR of 0.8212, with high sensitivity 0.9024 but lower specificity 0.7246. This indicates that while it is good at identifying positive cases, it misclassifies a significant number of negative cases. LDA performs similarly to BLR with a CCR of 0.8146 showing comparable sensitivity but slightly lower specificity. k-NN shows poor performance with a CCR of 0.649 indicating it struggles significantly with this dataset. Glmnet matches LDA's performance with a CCR of 0.8146 demonstrating its effectiveness. 60:40 Split Performance improves across all classifiers as more data is allocated for training. BLR and LDA both achieve a CCR of 0.8347, with BLR showing slightly better sensitivity 0.9242 compared to LDA 0.9394.k-NN continues to underperform, with a CCR dropping to 0.6116, indicating that it may not be suitable for this dataset. Glmnet also shows improvement with a CCR of 0.8278 maintaining strong sensitivity.

Table 1: results of classification methods on heart diseases data

| *Partitioning Algorithm* | *Classifiers* | *CCR* | *SEN* | *SPEC* | *PREC* | *AUC* | *BA* |
|---|---|---|---|---|---|---|---|
| 50:50 | *BLR* | **0.8212** | **0.9024** | **0.7246** | **0.7957** | 0.8916 | **0.8135** |
| | *LDA* | 0.8146 | **0.9024** | 0.7101 | 0.7872 | 0.8905 | 0.8062 |
| | *k-NN* | 0.649 | 0.6829 | 0.6087 | 0.6746 | 0.6887 | 0.6458 |
| | *Glmnet* | 0.8146 | **0.9024** | 0.7101 | 0.7872 | **0.8989** | 0.8062 |
| 60:40 | *BLR* | **0.8347** | 0.9242 | **0.7273** | **0.8026** | **0.9137** | **0.8257** |
| | *LDA* | 0.8347 | **0.9394** | 0.7091 | 0.7948 | 0.9099 | 0.8242 |
| | *k-NN* | 0.6116 | 0.5596 | 0.6818 | 0.6338 | 0.6155 | 0.6045 |
| | *Glmnet* | 0.8278 | 0.9268 | 0.7101 | 0.7917 | 0.9132 | 0.8185 |
| 70:30 | *BLR* | **0.8444** | 0.9184 | **0.7561** | **0.8181** | 0.9203 | **0.8372** |
| | *LDA* | **0.8444** | **0.9388** | 0.7317 | 0.8070 | 0.9193 | 0.8352 |
| | *k-NN* | 0.6111 | 0.7143 | 0.4878 | 0.6250 | 0.6444 | 0.6010 |
| | *Glmnet* | **0.8444** | 0.9184 | **0.7561** | **0.8181** | **0.9263** | **0.8372** |
| 80:20 | *BLR* | **0.8500** | **0.9697** | **0.7037** | **0.8000** | 0.9226 | **0.8367** |
| | *LDA* | 0.8333 | **0.9697** | 0.6667 | 0.7804 | **0.9292** | 0.8181 |
| | *k-NN* | 0.5833 | 0.6970 | 0.4444 | 0.6053 | 0.5959 | 0.5707 |
| | *Glmnet* | 0.8333 | **0.9697** | 0.6667 | 0.7805 | 0.9270 | 0.8181 |
| 90:10 | *BLR* | **0.8621** | **1.0000** | **0.6923** | **0.8000** | 0.9038 | 0.8461 |
| | *LDA* | **0.8621** | **1.0000** | **0.6923** | **0.8000** | 0.9038 | **0.8462** |
| | *k-NN* | 0.5862 | 0.7500 | 0.3846 | 0.6000 | 0.6226 | 0.5673 |
| | *Glmnet* | 0.8276 | **1.0000** | 0.6154 | 0.7619 | **0.9182** | 0.8076 |

**Note:** *BLR (Binary Logistic Regression); LDA (Linear Discriminant Analysis); k-NN (K-Nearest Neighbor); and Glmnet (Regularized Binary Logistic Regression).*

70:30 split further improvements are observed, particularly for BLR and LDA, both achieving a CCR of 0. 8444.Sensitivity remains high for both models, but k-NN continues to struggle with a CCR of only 0.6111, highlighting its limitations in this scenario.80:20 Split BLR reaches its highest CCR so far at 0.8500, with perfect sensitivity 0.9697 but lower specificity 0.7037. This indicates that it is highly proficient at detecting positive cases but may still misclassify some negatives. LDA maintains strong performance with a CCR of 0.8333 while k-NN drops further to 0.5833 indicating persistent issues.90:10 Split at this split, both BLR and LDA achieve the highest CCR of 0.8621, along with perfect sensitivity 1.0000. This indicates that they can accurately identify all positive cases in this scenario. However, k-NN continues to perform poorly with a CCR of only 0.5862, showing that it is not well-suited for this dataset.

The result in table 2 shows that 50:50 Split BLR achieves a CCR of 0.75, with a sensitivity of 0.5896, indicating it identifies about 59% of actual positive cases. Its specificity is relatively high at 0.8360 suggesting it effectively identifies negative cases. LDA has a slightly lower CCR of 0.7552 with sensitivity at 0.5671 and specificity at 0.8560. This indicates that LDA is also effective at identifying negatives but less so for positives compared to BLR.

Table 2: results of classification methods on diabetes data

| *Partitioning Algorithm* | *Classifiers* | *CCR* | *SEN* | *SPEC* | *PREC* | *AUC* | *BA* |
|---|---|---|---|---|---|---|---|
| 50:50 | *BLR* | 0.75 | 0.5896 | 0.8360 | 0.6583 | 0.8114 | 0.7127 |
| | *LDA* | 0.7552 | 0.5671 | **0.8560** | 0.6785 | 0.8126 | 0.8062 |
| | *k-NN* | 0.6979 | 0.5298 | 0.7880 | 0.5725 | 0.7676 | 0.6589 |
| | *Glmnet* | **0.8278** | **0.9268** | 0.7101 | **0.7916** | **0.8160** | **0.8184** |
| 60:40 | *BLR* | **0.7687** | **0.5794** | 0.8700 | 0.7045 | **0.8409** | **0.7247** |
| | *LDA* | 0.759 | 0.5607 | 0.8650 | 0.6896 | 0.8401 | 0.7128 |
| | *k-NN* | 0.7296 | 0.6355 | 0.7800 | 0.6071 | 0.7959 | 0.7077 |
| | *Glmnet* | 0.7655 | 0.5233 | **0.8950** | **0.7272** | 0.8407 | 0.7091 |
| 70:30 | *BLR* | **0.5739** | **0.3375** | 0.7000 | 0.3750 | **0.5576** | 0.5187 |
| | *LDA* | 0.5522 | 0.3250 | 0.6733 | 0.3466 | 0.5567 | 0.4991 |
| | *k-NN* | 0.6435 | 0.2000 | **0.8800** | **0.4705** | 0.5568 | **0.5400** |
| | *Glmnet* | 0.5609 | 0.2625 | 0.7200 | 0.3333 | 0.5490 | 0.4913 |
| 80:20 | *BLR* | **0.7582** | **0.5660** | **0.8600** | **0.6818** | **0.8437** | **0.7130** |
| | *LDA* | **0.7582** | 0.5471 | 0.8700 | 0.6904 | 0.8432 | 0.7085 |
| | *k-NN* | 0.7124 | **0.5660** | 0.7900 | 0.5882 | 0.7888 | 0.6780 |
| | *Glmnet* | 0.7712 | 0.5094 | 0.9100 | 0.7500 | 0.8575 | 0.7097 |
| 90:10 | *BLR* | **0.8289** | **0.7692** | **0.8600** | **0.7407** | **0.8830** | **0.8146** |
| | *LDA* | **0.8289** | **0.7692** | **0.8600** | **0.7407** | 0.8808 | **0.8146** |
| | *k-NN* | 0.7105 | 0.6923 | 0.7200 | 0.5625 | 0.8073 | 0.7062 |
| | *Glmnet* | 0.8276 | 1.000 | 0.6153 | 0.7619 | 0.8800 | 0.8077 |

**Note:** *BLR (Binary Logistic Regression); LDA (Linear Discriminant Analysis); k-NN (K-Nearest Neighbor); and Glmnet (Regularized Binary Logistic Regression).*

k-NN shows a CCR of 0.6979, with low sensitivity 0.5298 and relatively high specificity 0.7880. This suggests that while it can identify some negatives, it struggles significantly with positives. Glmnet performs best in this partitioning with a CCR of 0.8278 high sensitivity 0.9268 and specificity 0.7101. This indicates Glmnet is very

effective at identifying positive cases. 60:40 SplitPerformance improves for most classifiers as more data is allocated for training. BLR achieves a CCR of 0.7687, with sensitivity slightly decreasing to 0.5794 but specificity improves to 0.8700.LDA maintains a similar trend with a CCR of 0.759, showing slight improvements in specificity 0.8650 but lower sensitivity 0.5607.k-NN shows improvement in sensitivity 0.6355 but remains lower in overall performance with a CCR of 0.7296.Glmnet's performance drops slightly with a CCR of 0.7655 but its specificity increases significantly to 0.8950 indicating better identification of negative cases. 70:30 Split a notable decline in performance is observed across all classifiers, particularly for BLR and LDA. BLR drops to a CCR of 0.5739 with very low sensitivity 0.3375 and moderate specificity 0.7000. LDA also declines to a CCR of 0.5522 with sensitivity dropping further to 0.3250. k-NN shows some improvement in CCR 0.6435 but very low sensitivity 0. 2000.Glmnet performs poorly as well, with a CCR of only 0.5609 and low sensitivity 0.2625.80:20 Split Performance improves again for most classifiers. Both BLR and LDA achieve the same CCR of 0.7582, maintaining moderate sensitivity and high specificity. k-NN shows slight improvement with a CCR of 0.7124 indicating better overall performance than previous splits. Glmnet achieves the highest CCR at this split with 0.7712, although its sensitivity remains low at 0.5094.90:10 split at this partitioning, both BLR and LDA achieve the highest CCR of 0.8289, along with high sensitivity 0.7692 and good specificity 0.8600. This indicates that they are effective at identifying both classes in this scenario.k-NN performs reasonably well compared to earlier splits, achieving a CCR of 0.7105 but still has lower sensitivity compared to BLR and LDA. Glmnet maintains strong performance with a CCR of 0.8276, achieving perfect sensitivity 1.000 but lower specificity 0.6153.

## 4.0 Conclusions

The analysis of the heart disease dataset reveals that both Binary Logistic Regression (BLR) and Linear Discriminant Analysis (LDA) are effective classifiers, particularly as the training data proportion increases. These classifiers consistently outperform K-Nearest Neighbors (k-NN), which struggles across all partition ratios, indicating its unsuitability for this classification task. The high sensitivity of BLR and LDA highlights their effectiveness in identifying positive cases. In the diabetes dataset analysis, Glmnet also performs well across various data partitioning strategies, excelling in sensitivity, while k-NN continues to underperform. Notably, a decline in performance metrics at the 70:30 split suggests that model effectiveness is sensitive to the training-test data ratio, underscoring the importance of sufficient training data for optimal model learning. Overall, BLR and LDA demonstrate strong classification capabilities, especially with larger datasets (80:20 and 90:10 splits). The findings stress the importance of selecting appropriate classifiers based on dataset characteristics and partitioning strategies to enhance predictive accuracy in health-related classification tasks particularly in contexts with potential class imbalances.

## 5.0 Recommendations

BLR and LDA demonstrated strong performance particularly with larger training datasets (80:20 and 90:10 splits) it is recommended to prioritize BLR and LDA for classification tasks in health-related datasets making the methods suitable for critical applications such as disease diagnosis. The consistent underperformance of k-NN across various partition ratios indicates that it may not be suitable for these classification tasks. It is advisable to limit its use in favour of more robust classifiers like BLR and LDA. The study also revealed the relevance of adequate training data for model effectiveness. It is recommended to experiment with different data partitioning strategies to allocate sufficient data for training to enhance model learning and predictive accuracy. Also, in health-related classification tasks, where class imbalances can skew results, it is essential to implement techniques such as stratified sampling or synthetic data generation to ensure that minority classes are adequately represented during model training.

## References


Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database,2020*, baaa010.

Algan, G., & Ulusoy, I. (2021). Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, *215*, 106771.

Ali, A. H., Hussain, Z. F., & Abd, S. N. (2020). Big data classification efficiency based on linear discriminant analysis. *Iraqi Journal for Computer Science and Mathematics*, *1*(1), 7-12.

Bansal, M., Goyal, A., & Choudhary, A. (2022). A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal*, *3*, 100071.

Boateng, E. Y., Otoo, J., & Abaye, D. A. (2020). Basic tenets of classification algorithms K-nearest-neighbor, support vector machine, random forest and neural network: A review. *Journal of Data Analysis and Information Processing*, *8*(4), 341-357.

Bukhari, M. M., Ullah, S. S., Uddin, M., Hussain, S., Abdelhaq, M., & Alsaqour, R. (2022). An Intelligent Model for Predicting the Students' Performance with Backpropagation Neural Network Algorithm Using Regularization Approach. *Human-Centric Computing and Information Sciences*, *12*.

Chan, J. Y. L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z. W., & Chen, Y. L. (2022). Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics*, *10*(8), 1283.

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, *7*(1), 52. Gomila, R. (2021).

Dahouda, M. K., & Joe, I. (2021). A deep-learned embedding technique for categorical features encoding. *IEEE Access*, *9*, 114381-114391.

Fabiyi, S. D., Murray, P., Zabalza, J., & Ren, J. (2021). Folded LDA: extending the linear discriminant analysis algorithm for feature extraction and data reduction in hyperspectral remote sensing. *IEEE Journal of selected topics in applied earth observations and remote sensing*, *14*, 12312-12331.

Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, *150*(4), 700.

Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, *150*(4), 700.

Grover, D., & Toghi, B. (2020). MNIST dataset classification utilizing k-NN classifier with modified sliding-window metric. In *Advances in Computer Vision: Proceedings of the 2019 Computer Vision Conference (CVC), Volume 2 1* (pp. 583-591). Springer International Publishing.

Halvorson, M. A., McCabe, C. J., Kim, D. S., Cao, X., & King, K. M. (2022). Making sense of some odd ratios: A tutorial and improvements to present practices in reporting and visualizing quantities of interest for binary and count outcome models. *Psychology of Addictive Behaviors*, *36*(3), 284.

Harris, J. K. (2021). Primer on binary logistic regression. *Family medicine and community health*, *9*(Suppl 1).

Joseph, V. R., & Vakayil, A. (2022). SPlit: An optimal method for data splitting. *Technometrics*, *64*(2), 166-176.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR). [Internet]*, *9*(1), 381-386.

Mahmud, M. S., Huang, J. Z., Salloum, S., Emara, T. Z., & Sadatdiynov, K. (2020). A survey of data partitioning and sampling methods to support big data analysis. *Big Data Mining and Analytics*, *3*(2), 85-101.

McDonald, E., & Wang, X. (2024). Generalized regression estimators with concave penalties and a comparison to lasso type estimators. *METRON*, *82*(2), 213-239.

Moulaei, K., Sharifi, H., Bahaadinbeigy, K., Haghdoost, A. A., & Nasiri, N. (2023). Machine learning for prediction of viral hepatitis: A systematic review and meta-analysis. *International journal of medical informatics*, *179*, 105243.

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*.

Nanga, S., Bawah, A. T., Acquaye, B. A., Billa, M. I., Baeta, F. D., Odai, N. A., ... & Nsiah, A. D. (2021). Review of dimension reduction methods. *Journal of Data Analysis and Information Processing*, *9*(3), 189-231.

Neeraj, K. N., & Maurya, V. (2020). A review on machine learning (feature selection, classification and clustering) approaches of big data mining in different area of research. *Journal of critical reviews*, *7*(19), 2610-2626.

Neu, D. A., Lahann, J., & Fettke, P. (2022). A systematic literature review on state-of-the-art deep learning methods for process prediction. *Artificial Intelligence Review*, *55*(2), 801-827.

Oymak, S., Li, M., & Soltanolkotabi, M. (2021). Generalization guarantees for neural architecture search with train-validation split. In *International Conference on Machine Learning* (pp. 8291-8301). PMLR.

Rahim, A., Hassan, S., Ullah, N., Noor, N., Ahmed, Rafique, R., ... & Afaq, S. (2023). Association and comparison of periodontal and oral hygiene status with serum HbA1c levels: a cross-sectional study. *BMC Oral Health*, *23*(1), 442.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, *16*, 1-85.

Saha, S., Saha, M., Mukherjee, K., Arabameri, A., Ngo, P. T. T., & Paul, G. C. (2020). Predicting the deforestation probability using the binary logistic regression, random forest, ensemble rotational forest, REPTree: A case study at the Gumani River Basin, India. *Science of the Total Environment*, *730*, 139197.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), 160.

Schober, P., & Vetter, T. R. (2021). Logistic regression in medical research. *Anesthesia & Analgesia*, *132*(2), 365-366.

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018* (pp. 99-111). Springer Singapore.

Seng, L., & Li, J. (2022). Structural equation model averaging: Methodology and application. *Journal of Business & Economic Statistics*, *40*(2), 815-828.

Srimaneekarn, N., Hayter, A., Liu, W., & Tantipoj, C. (2022). Binary response analysis using logistic regression in dentistry. *International Journal of Dentistry*, *2022*(1), 5358602.

Suyal, M., & Goyal, P. (2022). A review on analysis of K-nearest neighbor classification machine learning algorithms based on supervised learning. *International Journal of Engineering Trends and Technology*, *70*(7), 43-48.

Tang, Y., Chen, D., & Li, X. (2021). Dimensionality reduction methods for brain imaging data analysis. *ACM Computing Surveys (CSUR)*, *54*(4), 1-36.

Tian, Y., & Zhang, Y. (2022). A comprehensive survey on regularization strategies in machine learning. *Information Fusion*, *80*, 146-166.

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), 6256.

Wahid, J. A., Shi, L., Gao, Y., Yang, B., Wei, L., Tao, Y., ... & Yagoub, I. (2022). Topic2Labels: A framework to annotate and classify the social media data through LDA topics and deep learning models for crisis response. *Expert Systems with Applications*, *195*, 116562.

Wang, K., & Thrampoulidis, C. (2022). Binary classification of gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science*, *4*(1), 260-284.

Yang, N. C., & Sung, K. L. (2023). Non-intrusive load classification and recognition using soft-voting ensemble learning algorithm with decision tree, K-Nearest neighbor algorithm and multilayer perceptron. *IEEE Access*.

Zhao, S., Zhang, B., Yang, J., Zhou, J., & Xu, Y. (2024). Linear discriminant analysis. *Nature Reviews Methods Primers*, *4*(1), 70.