Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary : Rapid Review : Open Access Journal Vol. 6, No. 2, May 2024, pp.78-84 e-ISSN: 2656-8624

RESEARCH ARTICLE OPEN ACCESS

Manuscript received December 18, 2023; revised January 20, 2024; accepted January 12, 2024; date of publication Mei 23, 2024
Digital Object Identifier (**DOI**): https://doi.org/10.35882/ijeeemi.v6i2.351

How to cite: Devi Aprilya Dinanthi, Elisa Ramadanti, Christian Sri Kusuma Aditya, and Didih Rizki Chandranegara, "Diabetes Detection Using Extreme Gradient Boosting (XGBoost) with Hyperparameter Tuning", Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 6, no. 2, pp. 78-84, Mei. 2024.

Diabetes Detection Using Extreme Gradient Boosting (XGBoost) with Hyperparameter Tuning

Devi Aprilya Dinanthi, Elisa Ramadanti, Christian Sri Kusuma Aditya, and Didih Rizki Chandranegara
Informatics Study Program, Faculty of Engineering, Universitas Muhammadiyah Malang, Indonesia

Corresponding author: Christian Sri Kusuma Aditya (christianskaditya@umm.ac.id).

ABSTRACT Diabetes is a serious condition that can lead to fatal complications and death due to metabolic disorders caused by a lack of insulin production in the body. This study aims to find the best classification performance on diabetes dataset using Extreme Gradient Boosting (XGBoost) method. The dataset used has 768 rows and 9 columns, with target values of 0 and 1. In this study, resampling is applied to overcome data imbalance using SMOTE, and hyperparameter optimization is performed using GridSearchCV and RandomSearchCV. Model evaluation was performed using confusion matrix as well as metrics such as accuracy, precision, recall, and F1-score. The test results show that the use of GridSearchCV and RandomSearchCV for hyperparameter tuning provides good results. The application of data resampling also managed to improve the overall model performance, especially in the XGBoost method that has been optimized using GridSearchCV, which achieved the highest accuracy of 85%, while XGBoost with RandomSearchCV optimization showed 83% accuracy performance.

INDEX TERMS Diabetes, XGBoost, SMOTE, Hyperparameter Tuning, GridSearchCV, RandomSearchCV

I. INTRODUCTION

Diabetes is a metabolic disorder caused by problems in insulin production in the body [1]. In this condition, the pancreas produces an insufficient amount of insulin, which results in an imbalance in blood sugar levels and increases blood sugar concentration [2]. Diabetes can be caused by a variety of factors, including genetic factors such as family history, and lifestyle factors such as smoking, unhealthy diet, lack of physical activity, stress management, and being overweight [3].

Diabetes is a significant global health problem. According to a 2021 report by the International Diabetes Federation (IDF), about 537 million people aged 20-79 worldwide have diabetes [4]. In Indonesia, the number of people with diabetes continues to increase every year. By 2021, the number of diabetes cases among adults will reach 19 million out of a total adult population of 179 million, with a prevalence of 10.8% [5]. Diabetes can be very dangerous if it causes complications in the sufferers such as nervous system damage (neuropathy), kidney system damage (nephropathy), eye damage (retinopathy), as well as microvascular and macrovascular complications. These complications can negatively affect the various organ systems of the human body over a period of time and can lead to death [6].

With the increasing number of diabetes cases, efforts to prevent and manage diabetes are becoming increasingly important. By intervening appropriately in diabetes, it becomes very important to reduce the risk of complications and the negative impact it causes. Currently, with the advancement of technology, diabetes identification can be done using machine learning algorithms with available data to analyze. One of the machine learning algorithms that can be applied is Extreme Gradient Boosting (XGBoost), which can provide a new foundation in diabetes prevention and control efforts.

Previously, various approaches have been taken to detect diabetes. In 2022, research was conducted on diabetes classification using the logistic regression method which achieved an accuracy of 77% [7]. Then in 2021, another study predicted diabetes using the fuzzy logic method to detect diabetes with an accuracy of 96.47% [8]. In 2023, research was also conducted to detect diabetes using the Support Vector Machine and Random Forest methods. The research applied the SMOTE (Synthetic Minority Oversampling Technique) technique to balance the diabetes data. The results showed that the Random Forest method with the application of the SMOTE technique achieved the highest accuracy, which was 95.8% [9].

Extreme Gradient Boosting (XGBoost) is a method developed by Chen and Guestrin (2016) that applies the concept of Gradient Boosting (GB) which is efficient, fast, and scalable. XGBoost uses a level-wise growth approach, forming a set of decision trees where the model depends on the previous model. The first model in XGBoost tends to be weak in initializing the prediction value, and is then strengthened through updating the weights in each model formed [10]. Previous research in 2023 has shown the superiority of XGBoost in predicting cardiovascular disease (CVD) compared to several other algorithms such as Decision Tree, K-Nearest Neighbors, Naïve Bayes, and Random Forest. In the study, XGBoost achieved the highest accuracy of 92.34%.[11]. Another study in 2024 also used the XGBoost method to detect and analyze breast cancer. The accuracy result obtained is 94.74% [12]. Selecting the right set of hyperparameters is important in terms of classification model performance and accuracy. Hyperparameter tuning allows tweak model performance for optimal results with try out different combinations of parameters and values by iteratively [13].

This research has contribution including:

1. Based on its proven performance and excellence in disease prediction, this study selected XGBoost for diabetes classification.
2. In addition, this research will apply data resampling techniques using SMOTE to overcome data imbalance.
3. Hyperparameter tuning is also added to achieve optimal classification results.

In the research process, there are several stages from preprocessing, normalizing data using the Min-Max Normalization technique to classifying using the Extreme Gradient Boosting (XGBoost) method. This research is expected to produce an effective classification model to identify diabetes in patients.

II. RESEARCH METHOD

In this study, classification will be carried out on the diabetes disease dataset using the XGBoost method which consists of several stages. The flow of this research will be shown in FIGURE 1.

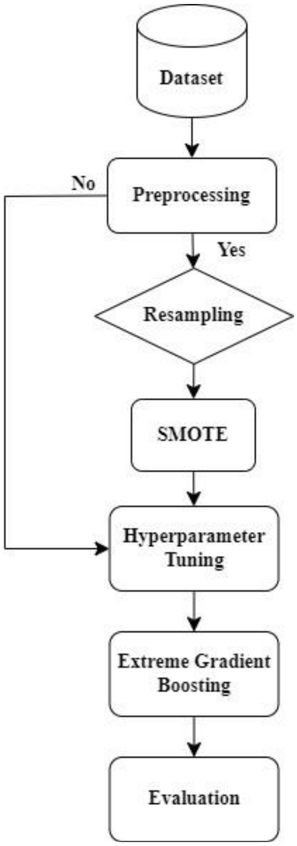

FIGURE 1. Research Flow

A. DATASET

The dataset used in this study is a diabetes dataset obtained from the kaggle website. Here is the link https://www.kaggle.com/datasets/mathchi/diabetes-data-set. This dataset consists of 768 rows and 9 columns, which are the attributes of the dataset. This dataset consists of two target classes, namely a class representing patients with diabetes and a class representing patients who do not have diabetes. This dataset is imbalanced with a ratio of 500 to 268, where 500 rows represent the class of patients who do not have diabetes and 268 rows represent the class of diabetic patients. To overcome this imbalance, data resampling is carried out using SMOTE so that the number of rows in each class becomes balanced with a ratio of 500 to 500. This dataset will be used for classification with the aim of predicting diabetes based on the attributes in the dataset. In this research, the XGBoost algorithm will be used to perform the classification. The following dataset details can be seen in TABLE 1.

TABLE 1. Dataset Details

Attribute	Description
Pregnancies	Number of pregnancies the patient has had.
Glucose	Plasma glucose concentration 2 hours after oral glucose tolerance.
Blood Presure	A measure of blood pressure.
SkinThickness	Thickness of the skin fold in the triceps region (mm).
Insulin	Insulin concentration in serum 2 hours after oral glucose tolerance test.
BMI	Body mass index, which is calculated as weight (kg) divided by height (m).
DiabetesPredigreeFunction	Family history of diabetes.
Age	Patient age (years).
Outcome	The target class, with values 0 and 1, where 0 represents non-diabetic patients and 1 represents non-diabetic patients.

B. PREPROCESSING

1) DATA CLEANING

Preprocessing is an important stage in the data analysis process [14]. The purpose of the preprocessing stage is to prepare and improve the quality of data before entering the analysis and classification model building stage so that the analysis results can be more effective and accurate [15]. The preprocessing stages carried out include the following:

2) DATA NORMALIZATION

Data normalization is the process of re-scaling the attribute values of the data so that it can make processing easier [16]. In this research, the data normalization used is Min-Max Scaling. Min-Max Scaling is used to transform data with a range of 0 and 1 and ensure there are no features with unbalanced values [17]. The following equation is like Eq. (1) :

$$x' = \left(\frac{x - \min(x)}{\max(x) - \min(x)}\right) \tag{1}$$

2) RESAMPLING

Resampling is a statistical technique used to overcome class imbalance in a dataset, such as in the classification process, by manipulating the training data to balance the data distribution [18]. In this research, the resampling technique used is SMOTE (Synthetic Minority Oversampling Technique), which is an oversampling technique that produces synthetic samples by combining minority samples and their neighbors [19]. The following are the results of the resampling process in this study can be seen through FIGURE 2 and FIGURE 3.

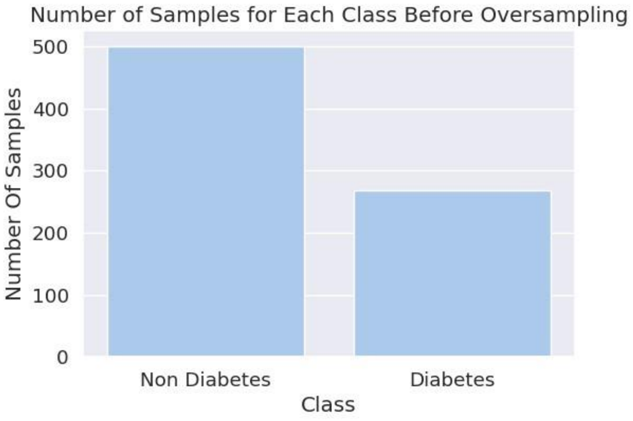

FIGURE 2. Before Resampling

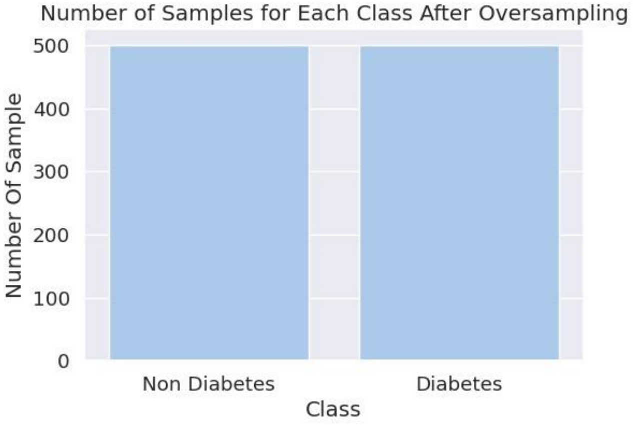

FIGURE 3. After Resampling

C. HYPERPARAMETER TUNING

In this study, parameter tuning is performed to evaluate how well the applied model can provide accurate results. It is also part of the training process using training data and validation data to achieve the optimal level of accuracy [20]. In this study, the hyperparameter tuning used is GridSearchCV and RandomSearchCV to select the most optimal combination of parameters with the aim of improving the performance of the XGBoost model.

D. EXTREME GRADIENT BOOSTING (XGBOOST)

Extreme Gradient Boosting (XGBoost) is a decision tree-based algorithm that belongs to the tree other Gradient Boosting methods, both for classification and regression problems [21]. In the context of regression trees, the internal nodes represent the values used to test the attributes, while the leaf nodes produce scores that represent the decision [22]. The final prediction is obtained from the sum of the scores predicted by the K tree, the following equation is shown in Eq. (2):

$$obj(\theta) = \sum_{i=1}^{n} l\ (y_i\,, \hat{y}_l) + \sum_{k}^{k} \Omega\ (fk) \tag{2}$$

Where is the rquation $\sum_{i=1}^{n} l\ (y_i, \hat{y}_l)$ is a loss function that can be calculated to measure how well the model fits the training dataset and $\sum_{k}^{k} \Omega\ (fk)$ is an equation that determines the complexity of the model.

Before building the XGBoost prediction model, the first step is to perform optimal hyperparameter tuning. XGBoost has many large hyperparameters that can affect its performance. These hyperparameters are divided into three categories: general parameters, booster parameters, and learning task parameters [23].

D. EVALUATION

In this research, the model will be evaluated using Confusion Matrix, a technique used to evaluate classification models [24]. Confusion matrix measures accuracy, recall, precision, and F1-score [25]. Confusion matrix involves several evaluation components, such as TP (True Positive) which represents correctly predicted positive data, TN (True Negative) which represents correctly predicted negative data, FP (False Positive) which indicates negative data that was incorrectly predicted as positive, and FN (False Negative) which reflects positive data that was incorrectly predicted as negative [26]. The following is the formula for calculating the confusion matrix used:

1) ACCURACY

Accuracy describes how much test data is successfully predicted correctly by the model [27]. The following equation is like Eq. (3):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

2) PRECISION

Precision is a metric that shows the number of true positive predictions [28]. The following equation is like Eq. (4):

$$precision = \frac{TP}{TP + FP} \tag{4}$$

3) RECALL

Recall is a metric that calculates the number of successful positive predictions [29]. The following equation is like Eq. (5):

$$recall = \frac{TP}{TP + FN} \tag{5}$$

4) F1-SCORE

F1-Score is an evaluation metric that combines the results of precision and recall values [30]. The following equation is like Eq. (6):

$$F1 - Score = 2\ \times \frac{precision\ \times recall}{precision + recall} \tag{6}$$

III. RESULT

The research used three test scenarios. The first test was conducted to optimize hyperparameter tuning with GridSearchCV. The GridSearchCV method is part of the scikit-learn library that performs validation for more than a few models while automatically and systematically providing hyperparameters for each model [31]. The following is the best hyperparameter tuning of the XGBoost method obtained using GridSearchCV can be seen in TABLE 2.

TABLE 2. Hyperparameter GridSearchCV

Hyperparameter	Hyperparameter Value	Best Hyperparameter Value	Description
Max_depth	5, 6, 7	6	Organize into max. decision trees.
min_child_weight	1, 2, 3	1	Determine the min. number of samples in each branch of the decision tree.
n_estimator	200, 300, 400	50	Determine the number of decision trees in the model.
Learning_rate	0.01, 0.05, 0.1	0.1	Measures the relative contribution of the model to the next model at each iteration.
Colsample_bytree	0.7, 0.8, 0.9	0.7	Controls the fraction of features used in each iteration when building the tree.

subsample	0.7, 0.8, 0.9	0.9	Controls the fraction of training data used in each iteration.

The results of the XGBoost method with GridSearchCV parameter settings show an accuracy of 0.75 or 75%, which will be displayed in the form of a classification report on TABLE 3.

TABLE 3. Classification Report with GridSearch CV

Model		Accuracy	Precision	Recall	F1-Score
	0		0.79	0.84	0.82
XGBoost	1	0.75	0.64	0.56	0.60

The second test was conducted to perform hyperparameter tuning optimization with RandomSearchCV. Parameter search using RandomizedSearchCV involves randomly combining elements from a predefined parameter space [32]. This approach allows for a more thorough exploration of the parameter space and helps find effective solutions to optimize the parameters of the XGBoost algorithm. The following is the best hyperparameter tuning of the XGBoost method obtained using RandomizedSearchCV can be seen in TABLE 4.

TABLE 4. Hyperparameter RandomSearchCV

Hyperparameter	Hyperparameter Value	Best Hyperparameter Value
Max_depth	5, 6, 7	5
Min_child_wight	1, 2, 3	1
N_estimator	200, 300, 400	300
Learning_rate	0.01, 0.05, 0.1	0.01
Colsample_bytree	0.7, 0.8, 0.9	0.8
subsample	0.7, 0.8, 0.9	0.7

The results of the XGBoost method with RandomSearchCV parameter tuning show an accuracy of 0.73 or 73%, which will be displayed in the form of a classification report.on TABLE 5.

TABLE 5. Classification Report with RandomSearchCV

Model		Accuracy	Precision	Recall	F1-Score
XGBoost	0		0.79	0.81	0.80
	1	0.73	0.60	0.58	0.59

In the third scenario, it was carried out to observe the effect of using data resampling using SMOTE on the classification of diabetes detection using the XGBoost method which had previously been optimized using GridSearchCV and RandomSearch. The following are the results obtained after resampling the data, which are shown in TABLE 6.

TABLE 6. Results using Data Resampling

Model	GridSearchCV After Resampling	RandomSearchCV After Resampling
XGBoost	0.85	0.83

Based on TABLE 6, the results of using data resampling on GridSearchCV show that the accuracy reaches 0.85 or 85%, while on RandomSearchCV the accuracy reaches 0.83 or 83%. The following is a comparison of the results obtained before and after resampling data on GridSearchCV and RandomSearchCV, which is shown in FIGURE 4.

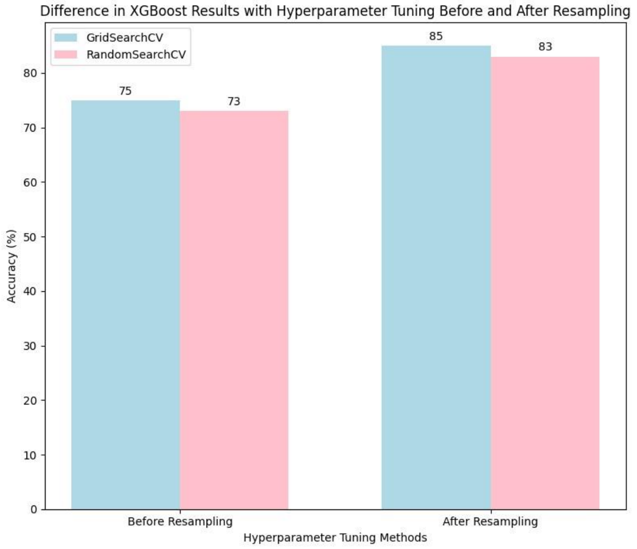

FIGURE 4. Difference in XGBoost result with Hyperparameter Tuning Before and After Resampling

In FIGURE 4, the difference in results before and after the application of data resampling is shown. Before data resampling was applied, GridSearchCV reached 75%, while after application, it reached 85%. Meanwhile, RandomSearchCV reached 73% before data resampling was applied, but after application, it reached 83% accuracy.

IV. DISCUSSION

The results of this study show that the application of parameter optimization through GridSearchCV and RandomSearchCV in the XGBoost method can affect the performance of the diabetes classification model. Based on the experimental results, it can be seen that the use of GridSearchCV produces an accuracy rate of 75%, while RandomSearchCV produces an accuracy rate of 73%. Although this difference is not significant, GridSearchCV tends to provide slightly superior performance in finding the best parameters for the XGBoost model.

Furthermore, the effect of data resampling on improving model performance shows a significant improvement. After applying data resampling, the model accuracy increased to 85% for GridSearchCV and 83% for RandomSearchCV, indicating that the application of data resampling using SMOTE effectively improves the model's ability to classify diabetes cases. Based on the visualization in FIGURE 4, the improvement after the application of data resampling is quite high, with a difference of 10% in GridSearchCV (previously reached 75%) and 10% in RandomSearchCV (previously reached 73%). The results of the combination of hyperparameter tuning optimization and data resampling can be an effective strategy in overcoming class imbalance and improving prediction accuracy on datasets that have unbalanced class distributions, such as in this diabetes disease detection case.

In the previous study, diabetes detection analysis was conducted using the Logistic Regression machine learning algorithm. Initially, the accuracy obtained was 77%, but increased to 82% after hyperparameter tuning using GridSearchCV. In contrast to the current study, where the highest accuracy was obtained using the XGBoost method which has gone through GridSearchCV hyperparameter tuning and applied data resampling, with accuracy reaching 85%. Before resampling, the accuracy obtained was 75%. The following comparison between previous research and current research will be shown in FIGURE 5.

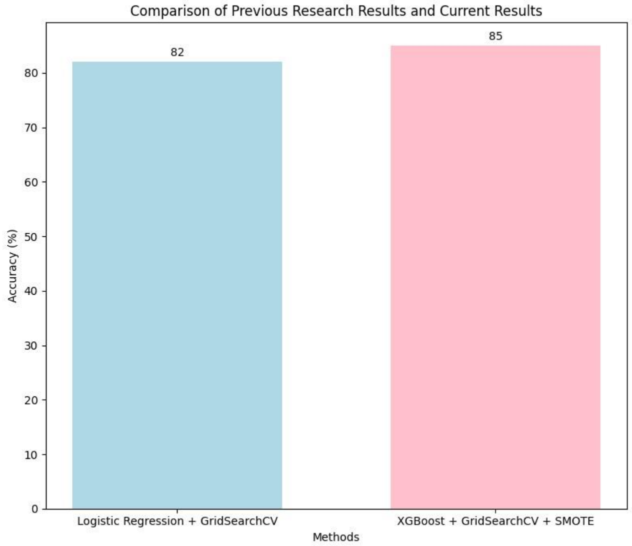

FIGURE 5. Comparison of Previous Research Result and Current Result

V. CONCLUSION

Based on the research conducted using the XGBoost method to detect diabetes, it can be concluded that the approach using Extreme Gradient Boosting (XGBoost) shows satisfactory performance in the detection of the disease. The results show that XGBoost displays good accuracy in classifying diabetic disease datasets. This research involved three testing scenarios that resulted in the comparison of different results. In the first scenario, tests were conducted to optimize the hyperparameters using GridSearchCV to find the best combination of parameters that significantly improved model performance. Evaluation of the model performance showed that the accuracy of XGBoost reached 75%. In the second scenario, tests were conducted to optimize hyperparameters using RandomSearchCV with the aim of finding the best parameters by performing random combinations of the selected hyperparameters to train the model. Evaluation of the model performance showed an increase in XGBoost accuracy to 73%. In the third scenario, tests were conducted to evaluate the use of data resampling in diabetes detection classification using the XGBoost method that has been optimized with GridSearchCV and RandomSearchCV. The evaluation results showed an increase in XGBoost accuracy to 85% from the previous 75%. Meanwhile, XGBoost that has been optimized with RandomSearchCV achieved 83% accuracy from the previous 73%.

From the three test scenarios, it can be concluded that the XGBoost method with GridSearchCV optimization shows the highest accuracy, and the use of data resampling further improves the performance of the model. In addition, the application of data resampling significantly improves the performance of the XGBoost model in detecting diabetes.

As for the limitations and weaknesses of this study, although there is a significant improvement in accuracy in comparison with previous studies using Logistic Regression after implementing GridSearchCV, the difference may be due to other factors such as variations in the features used. Although the improvement in accuracy shows positive results, further evaluation regarding the generalization of the model to new unknown datasets is still needed. Therefore, it is important to ensure that the developed model can be effectively implemented in real-world situations for diabetes detection.

REFERENCES

[1] Y. Mukhtar, A. Galalain, and U. Yunusa, "a Modern Overview on Diabetes Mellitus: a Chronic Endocrine Disorder," *Eur. J. Biol.*, vol. 5, no. 2, pp. 1–14, 2020, doi: 10.47672/ejb.409.

[2] M. R. Afandi and F. R. Marpaung, "Correlation Between Apoprotein B/Apoprotein a-I Ratio With Homa Ir Value (Homeostatic Model Assesment Insulin Resistance) in Type 2 Diabetes Mellitus," *J. Vocat. Heal. Stud.*, vol. 3, no. 2, p. 78, 2019, doi: 10.20473/jvhs.v3.i2.2019.78-82.

[3] L. Ismail, H. Materwala, and J. Al Kaabi, "Association of risk factors with type 2 diabetes: A systematic review," *Comput. Struct. Biotechnol. J.*, vol. 19, pp. 1759–1785, 2021, doi: 10.1016/j.csbj.2021.03.003.

[4] M. J. Hossain, M. Al-Mamun, and M. R. Islam, "Diabetes mellitus, the fastest growing global public health concern: Early detection should be focused," *Heal. Sci. Reports*, vol. 7, no. 3, pp. 5–9, 2024, doi: 10.1002/hsr2.2004.

[5] S. Webber, *International Diabetes Federation*, vol. 102, no. 2. 2013. doi: 10.1016/j.diabres.2013.10.013.

[6] B. CURCHOD and J. P. DAEPPEN, "Complications of diabetes mellitus," *Praxis (Bern. 1994).*, vol. 48, no. 26, pp. 602–603,

1959.

[7] Erlin, Y. N. Marlim, Junadhi, L. Suryati, and N. Agustina, “Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 11, no. 2, pp. 88–96, 2022.

[8] K. M. Aamir, L. Sarfraz, M. Ramzan, M. Bilal, J. Shafi, and M. Attique, “A fuzzy rule-based system for classification of diabetes,” *Sensors*, vol. 21, no. 23, 2021, doi: 10.3390/s21238095.

[9] H. Hairani and D. Priyanto, “A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 8, pp. 585–590, 2023, doi: 10.14569/IJACSA.2023.0140864.

[10] Y. Sitinjak, M. Nababan, and M. City, “Liver Disease Classification Analysis,” vol. 7, no. 1, pp. 132–141, 2023.

[11] F. Kanwal, M. K. Abid, M. S. Maqbool, D. N. Aslam, and M. Fuzail, “Optimized Classification of Cardiovascular Disease Using Machine Learning Paradigms,” *VFAST Trans. Softw. Eng.*, vol. 11, no. 2, pp. 140–148, 2023, doi: 10.21015/vtse.v11i2.1527.

[12] Rahmanul Hoque, Suman Das, Mahmudul Hoque, and Mahmudul Hoque, “Breast Cancer Classification using XGBoost,” *World J. Adv. Res. Rev.*, vol. 21, no. 2, pp. 1985–1994, 2024, doi: 10.30574/wjarr.2024.21.2.0625.

[13] Liao, L., Li, H., Shang, W., & Ma, L. (2022). An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks. ACM Transactions on Software Engineering and Methodology (TOSEM), 31(3), 1-40.

[14] A. Anggrawan and M. Mayadi, “Application of KNN Machine Learning and Fuzzy C-Means to Diagnose Diabetes,” *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 2, pp. 405–418, Mar. 2023, doi: 10.30812/matrik.v22i2.2777.

[15] F. Putra, H. F. Tahiyat, and R. M. Ihsan, “Application of K-Nearest Neighbor Algorithm Using Wrapper as Preprocessing for Determination of Human Weight Information Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia,” vol. 4, no. January, pp. 273–281, 2024.

[16] F. N. S. Inggih Permana, “The Effect of Data Normalization on the Performance of the Classification Results of the Backpropagation Algorithm,” *IJIRSE Indones. J. Inform. Res. Softw. Eng.*, vol. 2, no. 1, pp. 67–72, 2022.

[17] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Appl. Soft Comput.*, vol. 97, no. xxxx, p. 105524, 2020, doi: 10.1016/j.asoc.2019.105524.

[18] R. Ghorbani and R. Ghousi, “Comparing Different Resampling Methods in Predicting Students’ Performance Using Machine Learning Techniques,” *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.

[19] D. Elreedy, A. F. Atiya, and F. Kamalov, “A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning,” *Mach. Learn.*, no. 0123456789, 2023, doi: 10.1007/s10994-022-06296-4.

[20] X. Du, H. Xu, and F. Zhu, “Understanding the Effect of Hyperparameter Optimization on Machine Learning Models for Structure Design Problems,” *CAD Comput. Aided Des.*, vol. 135, p. 103013, 2021, doi: 10.1016/j.cad.2021.103013.

[21] J. Han, K. Shu, and Z. Wang, “Predicting energy use in construction using Extreme Gradient Boosting,” *PeerJ Comput. Sci.*, vol. 9, pp. 1–14, 2023, doi: 10.7717/peerj-cs.1500.

[22] F. Nurrahman, H. Wijayanto, A. H. Wigena, and N. Nurjanah, “Pre-Processing Data on Multiclass Classification of Anemia and Iron Deficiency With the Xgboost Method,” *BAREKENG J. Ilmu Mat. dan Terap.*, vol. 17, no. 2, pp. 0767–0774, 2023, doi: 10.30598/barekengvol17iss2pp0767-0774.

[23] N. T. Tran, T. T. G. Tran, T. A. Nguyen, and M. B. Lam, “A new grid search algorithm based on XGBoost model for load forecasting,” *Bull. Electr. Eng. Informatics*, vol. 12, no. 4, pp. 1857–1866, 2023, doi: 10.11591/eei.v12i4.5016.

[24] R. Sistem *et al.*, “JURNAL RESTI MRI Image Based Alzheimer ’ s Disease Classification Using,” vol. 5, no. 158, pp. 18–25, 2024.

[25] Siti Khomsah, Rima Dias Ramadhani, and Sena Wijaya, “The Accuracy Comparison Between Word2Vec and FastText On Sentiment Analysis of Hotel Reviews,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 3, pp. 352–358, 2022, doi: 10.29207/resti.v6i3.3711.

[26] J. H. B, “Risk Level Prediction of Life Insurance Applicant using Machine Learning,” *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 2, pp. 2213–2220, 2020, doi: 10.30534/ijatcse/2020/199922020.

[27] D. Liang, X. Jin, Y. Yuan, and R. Zou, “Performance Analysis of Machine Learning Methods,” *J. Phys. Conf. Ser.*, vol. 2428, no. 1, 2023, doi: 10.1088/1742-6596/2428/1/012039.

[28] I. Imantoko, A. Hermawan, and D. Avianto, “Comparative analysis of support vector machine and k-nearest neighbors with a pyramidal histogram of the gradient for sign language detection,” *Matrix J. Manaj. Teknol. dan Inform.*, vol. 11, no. 2, pp. 107–118, 2021, doi: 10.31940/matrix.v11i2.2433.

[29] Darussalam and G. Arief, “Jurnal Resti,” *Resti*, vol. 1, no. 1, pp. 19–25, 2018.

[30] R. Irmanita, Sri Suryani Prasetiyowati, and Yuliant Sibaroni, “Classification of Malaria Complication Using CART (Classification and Regression Tree) and Naïve Bayes,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 1, pp. 10–16, 2021, doi: 10.29207/resti.v5i1.2770.

[31] S. Kohli and P. Joshi, “‘ A Brief Study on Random Forest Using Python ,’” vol. 3, no. 6, pp. 2063–2069, 2021, doi: 10.35629/5252-030620632069.

[32] D. Navon and A. M. Bronstein, “Random Search Hyper-Parameter Tuning: Expected Improvement Estimation and the Corresponding Lower Bound,” 2022, [Online]. Available: http://arxiv.org/abs/2208.08170