Diabetes Disease Detection Classification Using Light Gradient Boosting (LightGBM) With Hyperparameter Tuning

Elisa Ramadanti[1], Devi Aprilya Dinathi[2], Christian Sri Kusuma Aditya[3]*, Didih Rizki Chandranegara [4]
[1,2,3,4] Universitas of Muhammadiyah Malang, Malang, Indonesia
[1]elisaramadanti11@webmail.umm.ac.id, [2]dinanthi140402@webmail.umm.ac.id,
[3]chistianskaditya@umm.ac.id, [4]didihrizki@umm.ac.id

Submitted :Feb 23, 2024 | **Accepted** : Mar 25, 2024 | **Published** : Apr 1, 2024

Abstract: Diabetes is a condition caused by an imbalance between the need for insulin in the body and insufficient insulin production by the pancreas, causing an increase in blood sugar concentration. This study aims to find the best classification performance on diabetes datasets with the LightGBM method. The dataset used consists of 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 perform hyperparameter optimization. Model evaluation is performed using confusion matrix and various metrics such as accuracy, recall, precision and f1-score. This research conducted several tests. In hyperparameter optimization tests using GridSearchCV and RandomSearchCV, the LightGBM method showed good performance. In tests that apply data resampling, the LightGBM method achieves the highest accuracy, namely the LightGBM method with GridSearchCV optimization with the highest accuracy reaching 84%, while LightGBM with RandomSearchCV optimization reaches 82% accuracy.

Keywords: Detection Disease; Light Gradient Boosting; GridSearchCV; RandomSearchCv; SMOTE; Hyperparameter Tuning; Diabetes

INTRODUCTION

Diabetes is a condition caused by an imbalance between the need for and production of the hormone insulin (Afandi & Marpaung, 2019). Insufficient insulin production by the pancreas leads to increased blood sugar concentrations (Silalahi, 2019). Insulin is a hormone used to regulate glucose sugar levels in the blood and will help the body use sugar as energy (Hardianto, 2021a). Diabetes can be triggered by several risk factors, namely risk factors that can be changed by humans and risk factors that cannot be changed by humans (Alya Azzahra Utomo, Andira Aulia R, Sayyidah Rahmah, 2020). Risk factors that cannot be changed are age and genetic factors. Changeable risk factors include lifestyle, including food intake, rest patterns, physical activity and stress management.

These risk factors, if not controlled, can be dangerous and cause serious complications, potentially increasing the risk of premature death for sufferers. In Indonesia, there is a continuous rise in the prevalence of diabetes year after year.. According to a report from the International Diabetes Federation (IDF) organization, by 2020, around 6% of the 172 million adults will have diabetes (Tanoey & Becher, 2021). Meanwhile, worldwide, according to a report from IDF in 2021, About 537 million individuals aged 20 to 79 years old are dealing with diabetes. (Fauzi & Yunial, 2022).

Therefore, this increase should be a serious concern for Indonesians to change lifestyles with a healthy lifestyle. Preventing diabetes and reducing the risk of complications are important goals for everyone. Early prediction of diabetes can help with proper intervention (Hardianto, 2021b). With the development of current technology, machine learning algorithms such as Light Gradient Boosting (LightGBM) can be used to detect diabetes through analysis of available data. LightGBM is one of the Gradient Boosting Decision Tree (GBDT) based machine learning methods that can be used to predict and classify data (Wardhani & Akbar, 2022). In this case, the utilization of LightGBM can provide a new foundation in diabetes prevention and control efforts.

Previously, research has been carried out related to the classification of diabetes. In 2021 research, diabetes data classification was carried out using two methods with the results showing that Modified Balanced Random Forest (MBRF) has a better performance of 97.8% than Support Vector Machine (SVM) with an accuracy of 87.94% (Purbolaksono et al., 2021). In the same year, research with normalization methods in the Random Forest

* Corresponding author

algorithm showed that Min-Max normalization improved classification performance by 95.45%, better than other normalization methods (Gde Agung Brahmana Suryanegara, Adiwijaya, 2021). In 2021, a study compared the ADASYN-SVM and SMOTE-SVM methods, which showed that the ADASYN-SVM method had a higher accuracy of 87.3%, compared to SMOTE-SVM which reached 85.4% in detecting type 2 diabetes (Ramadhan, 2021). In 2020, research by comparing the Support Vector Machine (SVM) and Naive Bayes methods using diabetes datasets and showing that the SVM method has a higher accuracy of 78.4% compared to naïve bayes of 76.98% (Maulidah et al., 2021). In 2022, a study used logistic regression and SMOTE to handle data imbalance, which resulted in an increase in accuracy from 77% to 82% after the application of hyperparameter tuning with gridSearch (Erlin et al., 2022).

Based on the background and description of previous research, this study aims to get the best classification results on diabetes datasets using the Light Gradient Boosting method. The selection of the method is based on its ability to perform prediction and classification. In addition, this research will implement a resampling technique using SMOTE (Synthetic Minority Oversampling Technique), which is needed to overcome imbalances and hyperparameter tuning to improve the performance of the model used. In the research process, there are several stages from the preprocessing stage which aims to overcome missing values by replacing or deleting data. Furthermore, data normalization is carried out using the Min-Max Normalization Technique. The next stage is to perform classification using the Light Gradient Boosting method. This research is expected to provide efficient classification results in identifying diabetes conditions in patients.

LITERATURE REVIEW

Light Gradient Boosting Machine (LightGBM) was developed to overcome the shortcomings of Gradient Boosting Decision Tree (GBDT) in handling big data by speeding up the training process up to 20 times without sacrificing almost the same accuracy (Hartanto et al., 2023). LightGBM has two important techniques which are gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) (Febriantoro et al., 2023). GOSS is a sampling method designed for Gradient Boosted Decision Trees (GDBT) that achieves a balance by simultaneously decreasing the data sample size while preserving accuracy in the trained decision tree. (Li et al., 2021). In addition, LightGBM applies a leaf-wise growth strategy, which is one of the characteristics of LightGBM that distinguishes it from other methods that have the ability to limit the depth of the model so as to avoid the use of large computing power and overfitting (Rachmadi et al., 2021). So, this research will focus on using the Light Gradient Boosting (LightGBM) method.

Previous research has been conducted related to the issues raised and used the LightGBM method for classification and prediction of various diseases. For example, research conducted for the classification of stroke disease which resulted in an accuracy of 98% (Kurniadi & Larasati, 2022). In addition, the prediction of cardiovascular heart disease using the LightGBM method shows good results with an accuracy of 68% (Nugraha, 2021). Based on the research that has been done, it is proven that the use of the LightGBM method in the problem of disease classification and prediction has succeeded in getting a high level of accuracy in predicting and classifying diseases.

METHOD

In this study, the classification of diabetes datasets using the Light Gradient Boosting (LightGBM) method consists of several stages. The flow of this research will be shown by figure 1.

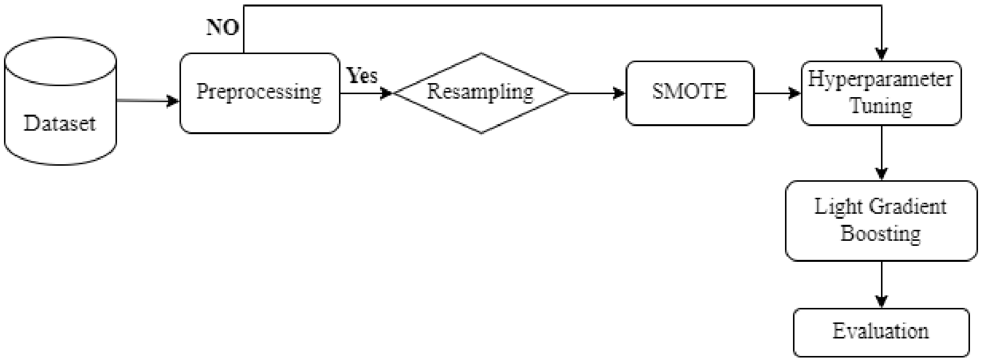

Fig. 1 Research Flow

Dataset

The dataset used in this study is a diabetes dataset obtained from the kaggle website. 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

* Corresponding author

is imbalanced with a ratio of 500 to 268. To overcome this imbalance, data resampling is carried out using the SMOTE method so that the number of rows in each class becomes balanced with a ratio of 500 to 500. This aims to improve data accuracy and cleanliness. The resampled dataset will be used for classification, with the aim of predicting diabetes based on existing attributes. In this research, the LightGBM algorithm is chosen to perform the classification. Table 1 shows the details of the dataset.

Table 1. Dataset Information

attribute	Descriptiona
Pregnancies	Number of pregnancies the patient has had
Glucose	Plasma glucose concentration 2 hours after oral glucose tolerance test
BloodPressure	Blood pressure measures
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) squared.
DiabetesPredigreeFunction	Family history of diabetes
Age	Patient age (years)
Outcome	The target class, with values of 0 and 1, where 0 represents patients who do not have diabetes and 1 represents patients who have diabetes

Preprocessing

Preprocessing is an important part of data mining (Anggrawan & Mayadi, 2023) with the aim of improving data quality so that analysis results are more accurate, information is more meaningful, and classification models are better (Liang et al., 2023). This process involves removing invalid data, managing missing values, normalizing, and transforming data so that the data is ready to be used in the next stage.

Data Cleaning

Data cleaning is the initial stage that aims to ensure that the dataset does not contain incorrect or erroneous data (Lee et al., 2021). In this research, blank values and duplication in the data are checked to make the data cleaner, more accurate and reliable to get more valid data analysis results.

Data Normalization

In this study, data normalization was carried out, to rescale the data (Pneumonia et al., 2022) using Min-Max Scaling, which is a normalization method that is often used to overcome the problem of differences in value scales between features that have too far a distance (Wijayanti et al., 2018). The following equation is like formula 1.

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

Description:

x' : Data Normalization Result
x : Data that has not been normalized
min(x) : Minimum value of all data
max(x) : Maximum value of all data

Data Resampling

Resampling is a technique used to handle class imbalance in datasets to avoid problems in the classification process. In the process, it will manipulate the training data to balance the data distribution (Handayani & Erni, 2023). In this study, resampling was carried out using SMOTE (Synthetic Minority Oversampling Technique), which is a development of the oversampling method (Moh. Badris Sholeh Rahmatullah et al., 2022). The following are the results of the resampling stage in this study as shown in figure 2 and figure 3.

* Corresponding author

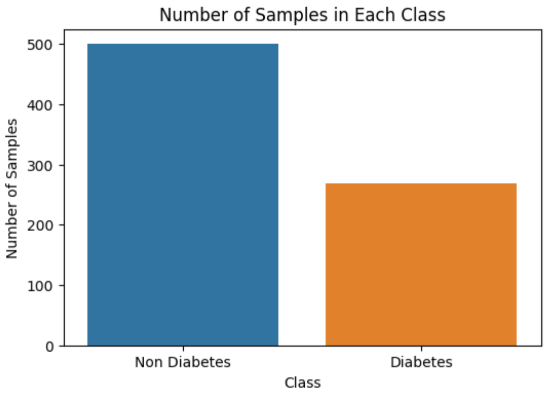

Fig. 2 Before Resampling

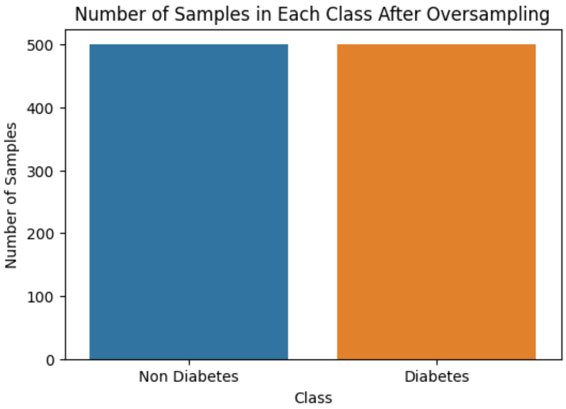

Fig. 3 After Resampling

Hyperparameter Tuning Optimization

In this study, hyperparameter tuning optimization was performed using GridSearchCV and RandomSearchCV to select the most optimal combination of parameters to improve model performance (Marlim et al., 2022). GridSearchCV is a method to find the best combination of a predefined set of hyperparameter values (Kohli & Joshi, 2021), while RandomSearchCV is an alternative method to find the best parameters in a model, using a random combination of selected hyperparameters to train the model (T. A. E. Putri et al., 2023). The parameters that are often used are Max_depth, num_leaves, n_estimator, learning_ratem and many more.

Light Gradient Boosting (LightGBM) Algorithm

Light Gradient Boosting (LightGBM) is a Gradient-Boosting Decision tree (GBDT) based machine learning method that can be used for data prediction and classification. The method is designed to achieve optimal efficiency, with several advantages such as faster training speed, ability to handle large data volumes and minimal memory usage (Wardani et al., 2023). LightGBM has two strategies that can be used, namely gradient-based one-side sampling (GOSS) and leaf-wise growth (Ju et al., 2019). The LightGBM method trains models with T trees by applying additive training proces where each new model learns to predict the rest of the previous model (Chen et al., 2019). To build a LightGBM model with T trees can be done by applying the equation as in formula 2.

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2}$$

Where in the equation, $\hat{y}_i^{(t)}$ is the prediction for the i example at the-t iteration. The function f_t The function ft is the function trained for the-t decision tree. Each iteration, the model $\hat{y}_i$ remains in use and a new function f or trained residuals are added in the model. LightGBM can use parameters to improve its capabilities and performance, some parameters that can be used are such as learning_rate, min_data_in_leaf, bagging_fraction and many more parameters that can be used (Wang & Wang, 2020).

* Corresponding author

Evaluation

Evaluation is done to measure the ability of the model either on the training set, validation set or test set (Rajagede, 2021). This research evaluates the model using confusion matrix with several metrics such as accuracy, recall, precision and f1-score. Confusion matrix consists of several components to evaluate, namely TP (true positive) value or positive data that is predicted correctly, TN (true negative) value or negative data that is predicted correctly, FP (false positive) value or negative data detected as positive data, and FN (false negative) value or positive data but detected as negative data (Nikmatun & Waspada, 2019). Accuracy is a description of test data that is correctly predicted by the model (Arifah et al., 2023). The following equation is like formula 3.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (3)$$

Recall is a metric that counts the number of positive predictions that have been made. The following equation is like formula 4.

$$recall = \frac{TP}{TP+FN} \qquad (4)$$

Precision is a metric that indicates the number of positive predictions that are actually positive. The following equation is like formula 5.

$$precision = \frac{TP}{TP+FP} \qquad (5)$$

F1-Score is an evaluation metric that combines the results of precision and recall values. The following equation is like formula 6.

$$F1 - Score = 2 \times \frac{precision \times recall}{precision+recall} \qquad (6)$$

RESULT

In this study, results were obtained from three test scenarios conducted. First, the objective is to perform hyperparameter tuning optimization using GridSearchCV. Second, it aims to perform hyperparameter tuning optimization using RandomSearchCV. Third, to evaluate the effect of resampling data with oversampling techniques. In the first test, the application of GridSearchCV with the best hyperparameter of the LightGBM method, obtained through GridSearchCV, is shown in Table 2.

Table 2. Best hyperparameters of LightGBM method with GridSearchCV

Hyperparameter	hyperparameter value	Best hyperparameter
Max_depth	3, 5, 10, 15	5
num_leaves	31, 50, 100, 150	31
n_estimator	50, 100, 200, 300	200
Learning_rate	0.01, 0.1, 0.2	0.01

With the application of GridSearchCV, the classification results show that the accuracy obtained reaches 0.77 or 77%. Table 3 will show the classification report obtained.

Table 3. Clasification Report LightGBM with GridSearchCV

Model		Accuracy	Precision	Recall	F1-Score
	0		0.87	0.82	0.82
LightGBM	1	0.77	0.67	0.68	0.67

In the Second Test, RandomSearchCV was applied with the best hyperparameters from the LightGBM method, which were obtained through RandomSearchCV as shown in Table 4.

Table 4. Best hyperparameters of LightGBM method with RandomSearchCV

Hyperparameter	hyperparameter value	Best hyperparameter
Max_depth	3, 5, 10, 15	10
num_leaves	31, 50, 100, 150	50
n_estimator	50, 100, 200, 300	300
Learning_rate	0.01, 0.1, 0.2	0.01

With the application of GridSearchCV, the classification results show that the accuracy obtained reaches 0.76 or 76%. Table 5 will show the classification report obtained by LightGBM with RandomSearchCV.

Table 5. Classification Report LightGBM with RandomSearchCV

Model		Accuracy	Precision	Recall	F1-Score
	0		0.83	0.79	0.81
LightGBM	1	0.76	0.64	0.70	0.67

In the Third Test, the aim is to evaluate the effect of data resampling in the classification of diabetes disease detection using the LightGBM machine learning method that has undergone optimization with GridSearchCV and RandomSearchCV. The following are the results obtained after resampling the data, which are shown in Table 6.

Table 6. Results Using Data Resampling

Model	GridSearchCV Before resampling	GridSearchCV After Resampling	RandomSearchCV before Resampling	RandomSearcgCV After Resampling
LightGBM	0.77	0.84	0.76	0.82

Based on the results in Table 6, LightGBM with GridSearchCV after resampling gets higher accuracy which reaches 0.84 or 84% compared to RandomSearchCV with an accuracy of 0.82 or 82%.

DISCUSSIONS

The test results show that the use of hyperparameter tuning optimization with GridSearchCV and RandomSearchCV significantly affects the performance of the model in diabetes detection classification using the LightGBM machine learning method. The application of GridSearchCV resulted in a classification accuracy of 77%, while the use of RandomSearchCV resulted in an accuracy of 76%. Although the difference is not significant, GridSearchCV tends to provide slightly better performance in determining the best hyperparameters for the LightGBM model.

Furthermore, evaluation of the effect of resampling the data showed a marked improvement in model performance. After resampling the data, the classification accuracy increased to 84% with GridSearchCV and 82% with RandomSearchCV. This indicates that the use of oversampling techniques can help improve the model's ability to better classify diabetes cases.

Based on Table 6, using a combination of hyperparameter tuning optimization and data resampling significantly improved the performance of the model in predicting diabetes detection using the LightGBM method. This finding shows that after applying data resampling, the LightGBM method optimized with GridSearchCV yields the highest accuracy.

CONCLUSION

Based on research conducted using the LightGBM method to detect diabetes, it can be concluded that the Light Gradient Boosting (LightGBM) method has good performance. The evaluation results show that LightGBM has good accuracy in classifying diabetic disease datasets. This research was conducted with three test scenarios that resulted in a comparison of different results. In the first scenario, tests were conducted to optimize hyperparameters with GridSearchCV, to find the best combination of parameters that significantly improved model performance. Evaluation of model performance showed the accuracy of the LightGBM method was 0.77 or 77%. Then, in the second scenario, tests were conducted to optimize the hyperparameters with RandomSearchCV with the aim of finding the best parameters by performing a random combination of the selected hyperparameters to train the model. The model performance evaluation shows that the accuracy of LightGBM reaches 0.76 or 76. In the third scenario, tests were conducted to evaluate the effect of data resampling on the LightGBM method that has performed gridSearchCV and RandomSearchCV optimization. The evaluation results showed that the LightGBM method that has performed gridSearchCV optimization, achieved an accuracy of 0.84 from the previous 0.77. While the lightGBM method that has performed randomSearchCV optimization, achieved an accuracy of 0.82 from the previous 0.76.

From the three test scenarios, it can be concluded that the LightGBM 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 LightGBM model in detecting diabetes.

ACKNOWLEDGMENT

The acknowledgments are given at the end of the research paper and should at a minimum name the sources of funding that contributed to the article. You may also recognize other people who contributed to the article or data contained in the article but at a level of effort that does not justify their inclusion as authors. You may also state the research grant contract number if any.

REFERENCES

Afandi, M. R., & Marpaung, F. R. (2019). Correlation Between Apoprotein B/Apoprotein a-I Ratio With Homa Ir Value (Homeostatic Model Assesment Insulin Resistance) in Type 2 Diabetes Mellitus. *Journal of Vocational Health Studies*, *3*(2), 78. https://doi.org/10.20473/jvhs.v3.i2.2019.78-82

Alya Azzahra Utomo, Andira Aulia R, Sayyidah Rahmah, R. A. (2020). FAKTOR RISIKO DIABETES MELLITUS TIPE 2: A SYSTEMATIC REVIEW. *AN-Nur: Jurnal Kajian Dan Pengembangan Kesehatan Masyarakat*, *1*(1), 44–52. https://doi.org/10.31101/jkk.395

Anggrawan, A., & Mayadi, M. (2023). Application of KNN Machine Learning and Fuzzy C-Means to Diagnose Diabetes. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *22*(2), 405–418. https://doi.org/10.30812/matrik.v22i2.2777

Chen, T., Xu, J., Ying, H., Chen, X., Feng, R., Fang, X., Gao, H., & Wu, J. (2019). Prediction of Extubation Failure for Intensive Care Unit Patients Using Light Gradient Boosting Machine. *IEEE Access*, *7*, 150960–150968. https://doi.org/10.1109/ACCESS.2019.2946980

Erlin, Yulvia Nora Marlim, Junadhi, Laili Suryati, & Nova Agustina. (2022). Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression. *Jurnal Nasional Teknik Elektro Dan Teknologi Informasi*, *11*(2), 88–96. https://doi.org/10.22146/jnteti.v11i2.3586

Fauzi, A., & Yunial, A. H. (2022). JEPIN (Jurnal Edukasi dan Penelitian Informatika) Optimasi Algoritma Klasifikasi Naive Bayes, Decision Tree, K-Nearest Neighbor, dan Random Forest menggunakan Algoritma Particle Swarm Optimization pada Diabetes Dataset. *(JEPIN) Jurnal Edukasi Dan Penelitian Informatika*, *8*(3), 470–481.

Febriantoro, E., Setyati, E., & Santoso, J. (2023). PEMODELAN PREDIKSI KUANTITAS PENJUALAN MAINAN MENGGUNAKAN LightGBM. *SMARTICS Journal*, *9*(1), 7–13. https://ejournal.unikama.ac.id/index.php/jst/article/view/8279

Gde Agung Brahmana Suryanegara, Adiwijaya, M. D. P. (2021). Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(1), 114–122. https://doi.org/10.29207/resti.v5i1.2880

Handayani, K., & Erni, E. (2023). Penerapan Light Gradient Boosting Dalam Prediksi Rasio Klik Tayang. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(1), 13–18. https://doi.org/10.36040/jati.v7i1.6010

Hardianto, D. (2021a). Insulin: Produksi, Jenis, Analisis, dan Rute Pemberian. *Bioteknologi Dan Biosains Indonesia*, *8*(2), 321–331. http://ejurnal.bppt.go.id/index.php/JBBI

Hardianto, D. (2021b). Telaah Komprehensif Diabetes Melitus: Klasifikasi, Gejala, Diagnosis, Pencegahan, Dan Pengobatan. *Jurnal Bioteknologi & Biosains Indonesia (JBBI)*, *7*(2), 304–317. https://doi.org/10.29122/jbbi.v7i2.4209

Hartanto, A. D., Nur Kholik, Y., & Pristyanto, Y. (2023). Stock Price Time Series Data Forecasting Using the Light Gradient Boosting Machine (LightGBM) Model. *JOIV : International Journal on Informatics Visualization*, *7*(4), 2270–2279. https://doi.org/10.30630/joiv.7.4.1740

Ju, Y., Sun, G., Chen, Q., Zhang, M., Zhu, H., & Rehman, M. U. (2019). A model combining convolutional neural network and lightgbm algorithm for ultra-short-term wind power forecasting. *IEEE Access*, *7*, 28309–28318. https://doi.org/10.1109/ACCESS.2019.2901920

Kohli, S., & Joshi, P. (2021). *" A Brief Study on Random Forest Using Python ."* *3*(6), 2063–2069. https://doi.org/10.35629/5252-030620632069

Kurniadi, F. I., & Larasati, P. D. (2022). Light Gradient Boosting Machine untuk Deteksi Penyakit Stroke. *Jurnal*

* Corresponding author

SISKOM-KB (Sistem Komputer Dan Kecerdasan Buatan), *6*(1), 67–72. https://doi.org/10.47970/siskom-kb.v6i1.328

Lee, G. Y., Alzamil, L., Doskenov, B., & Termehchy, A. (2021). *A Survey on Data Cleaning Methods for Improved Machine Learning Model Performance*. 1–6. http://arxiv.org/abs/2109.07127

Li, L., Lin, Y., Yu, D., Liu, Z., Gao, Y., & Qiao, J. (2021). A Multi-Organ Fusion and LightGBM Based Radiomics Algorithm for High-Risk Esophageal Varices Prediction in Cirrhotic Patients. *IEEE Access*, *9*, 15041–15052. https://doi.org/10.1109/ACCESS.2021.3052776

Liang, D., Jin, X., Yuan, Y., & Zou, R. (2023). Performance Analysis of Machine Learning Methods. *Journal of Physics: Conference Series*, *2428*(1), 481–490. https://doi.org/10.1088/1742-6596/2428/1/012039

Marlim, Y. N., Suryati, L., & Agustina, N. (2022). *Early Detection of Diabetes Using Machine Learning with Logistic Regression Algorithm*. *11*(2), 88–96.

Maulidah, N., Supriyadi, R., Utami, D. Y., Hasan, F. N., Fauzi, A., & Christian, A. (2021). Prediksi Penyakit Diabetes Melitus Menggunakan Metode Support Vector Machine dan Naive Bayes. *Indonesian Journal on Software Engineering (IJSE)*, *7*(1), 63–68. https://doi.org/10.31294/ijse.v7i1.10279

Moh. Badris Sholeh Rahmatullah, Aulia Ligar Salma Hanani, Akmal Muhammad Naim, Zamah Sari, & Yufis Azhar. (2022). Detection of Credit Card Fraud with Machine Learning Methods and Resampling Techniques. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *6*(6), 923–929. https://doi.org/10.29207/resti.v6i6.4213

Nikmatun, I. A., & Waspada, I. (2019). Implementasi Data Mining untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. *Jurnal SIMETRIS*, *10*(2), 421–432.

Nugraha, W. (2021). Prediksi Penyakit Jantung Cardiovascular Menggunakan Model Algoritma Klasifikasi. *Jurnal Managemen Dan Informatika*, *9*(2), 3–8.

Pneumonia, F., Mortality, T., Comparative, U., & Perceptron, M. (2022). *Jurnal resti*. *5*(158), 528–537.

Purbolaksono, M. D., Irvan Tantowi, M., Imam Hidayat, A., & Adiwijaya, A. (2021). Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes. *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, *5*(2), 393–399. https://doi.org/10.29207/resti.v5i2.3008

Putri, T. A. E., Widiharih, T., & Santoso, R. (2023). Penerapan Tuning Hyperparameter Randomsearchcv Pada Adaptive Boosting Untuk Prediksi Kelangsungan Hidup Pasien Gagal Jantung. *Jurnal Gaussian*, *11*(3), 397–406. https://doi.org/10.14710/j.gauss.11.3.397-406

Rachmadi, R. R., Sudarsono, A., & Santoso, B. (2021). Implementasi Metode LightGBM Untuk Klasifikasi Kondisi Abnormal Pada Pengemudi Sepeda Motor Berbasis Sensor Smartphone. *Jurnal Komputer Terapan*, *7*(2), 218–227.

Rajagede, R. A. (2021). Improving Automatic Essay Scoring for Indonesian Language using Simpler Model and Richer Feature. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *4*, 11–18. https://doi.org/10.22219/kinetik.v6i1.1196

Ramadhan, N. G. (2021). Comparative Analysis of ADASYN-SVM and SMOTE-SVM Methods on the Detection of Type 2 Diabetes Mellitus. *Scientific Journal of Informatics*, *8*(2), 276–282. https://doi.org/10.15294/sji.v8i2.32484

Silalahi, L. (2019). Hubungan Pengetahuan dan Tindakan Pencegahan Diabetes Mellitus Tipe 2. *Jurnal PROMKES*, *7*(2), 223. https://doi.org/10.20473/jpk.v7.i2.2019.223-232

Tanoey, J., & Becher, H. (2021). Diabetes prevalence and risk factors of early-onset adult diabetes: results from the Indonesian family life survey. *Global Health Action*, *14*(1). https://doi.org/10.1080/16549716.2021.2001144

Wang, Y., & Wang, T. (2020). Application of improved LightGBM model in blood glucose prediction. *Applied Sciences (Switzerland)*, *10*(9). https://doi.org/10.3390/app10093227

Wardani, B. S., Sa, S., & Nurjanah, D. (2023). *Measuring and Mitigating Bias in Bank Customers Data with XGBoost , LightGBM , and Random Forest Algorithm*. *9*(1), 142–155. https://doi.org/10.26555/jiteki.v9i1.25768

Wardhani, K. D. K., & Akbar, M. (2022). Diabetes Risk Prediction Using Extreme Gradient Boosting (XGBoost). *Jurnal Online Informatika*, *7*(2), 244–250. https://doi.org/10.15575/join.v7i2.970

Wijayanti, R. A., Furqon, M. T., & Adinugroho, S. (2018). Penerapan Algoritme Support Vector Machine Terhadap Klasifikasi Tingkat Risiko Pasien Gagal Ginjal. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *2*(10), 3500–3507. http://j-ptiik.ub.ac.id