IJCCS (Indonesian Journal of Computing and Cybernetics Systems)
Vol.19, No.3, July 2025, pp. 327~336
ISSN (print): 1978-1520, ISSN (online): 2460-7258
DOI: https://doi.org/10.22146/ijccs.107881

Implementation of Chi-Square Feature Selection for Parkinson's Disease Classification Using LightGBM

Annisa Salsabila Ahdyani[1], Irwan Budiman*[2], Dwi Kartini[3], Andi Farmadi[4], Muhammad Itqan Mazdadi[5]
[1,2,3,4,5] Department of Computer Science, Faculty of Mathematics and Natural Science, Lambung Mangkurat University, Banjarbaru, Indonesia
e-mail: [1]annisasalsabila60@gmail.com, ***[2]irwan.budiman@ulm.ac.id**, [3]dwikartini@ulm.ac.id, [4]andifarmadi@ulm.ac.id, [5]mazdadi@ulm.ac.id

Abstrak

Penyakit Parkinson merupakan gangguan kerusakan sel-sel saraf di otak dan termasuk penyakit yang berkembang paling cepat secara global. Upaya yang dapat dilakukan untuk mencegah meningkatnya kasus penyakit parkinson adalah diagnosis melalui pendekatan klasifikasi menggunakan algoritma pembelajaran mesin. Penelitian ini mengimplementasikan metode seleksi fitur Chi-Square yang dikombinasikan dengan algoritma Light Gradient Boosting Machine (LightGBM) untuk klasifikasi penyakit Parkinson. Tujuan dari seleksi fitur Chi-Square adalah untuk menghilangkan fitur-fitur yang tidak terlalu relevan sehingga dapat meningkatkan hasil kinerja model. Selain itu, diterapkan metode SMOTE untuk menangani ketidakseimbangan data serta hyperparameter tuning untuk menentukan kombinasi parameter optimal. Pengujian menggunakan sepuluh variasi jumlah fitur hasil seleksi, mulai dari 5 hingga 250 fitur. Hasil terbaik diperoleh pada penggunaan 200 fitur dengan nilai akurasi sebesar 96,05%. Penerapan Chi-Square berhasil meningkatkan performa pada model LightGBM dibandingkan dengan model yang tidak menggunakan seleksi fitur Chi-Square. Penerapan kombinasi metode ini secara signifikan mampu meningkatkan performa model klasifikasi dan berpotensi diaplikasikan dalam sistem pendukung diagnosis penyakit Parkinson.

Kata kunci*—Penyakit Parkinson, Chi-Square, LightGBM*

Abstract

Parkinson's disease refers to neurological disorder caused by damage to brain's nerve cells and is among the most rapidly increasing diseases globally. One way that can be done to prevent the increasing cases related to Parkinson's disease is diagnosis through an algorithmic learning approach classification method. This study implements the Chi-Square technique for approach for selecting relevant features in conjunction with the Light Gradient Boosting Machine (LightGBM) algorithm for Parkinson's disease classification. Chi-Square feature selection aims to reduce less relevant features so as to improve model performance results. In addition, the SMOTE method is applied to handle data imbalance and hyperparameter tuning to determine the optimal parameter combination. The test used ten variations in the number of selected features, ranging from 5 to 250 features. The best result was obtained when using 200 features, achieving an accuracy of 96.05%. Utilizing the Chi-Square method enhanced the performance of the LightGBM model when compared to its performance without feature selection. The application of this combination of methods can significantly improve the performance of the classification model and has the potential to be applied in the Parkinson's disease diagnosis support system.

Keywords—*Parkinson Disease, Chi-Square, LightGBM*

Received June 13th,2025; Revised July 29th, 2025; Accepted July 30th, 2025

1. INTRODUCTION

A condition known as Parkinson's disease damages the brain's or central nervous system's nerve cells. It ranks among the most rapidly increasing diseases globally, with the number of cases expected to increase in the next two decades. In general, Parkinson's disease usually appears at the age of 55 to 65 years and is more common in men than women [1]. An algorithm can be used to diagnose Parkinson's disease using a classification procedure in an attempt to stop the number of cases from rising.

In implementing the classification process, there is often a problem of many features in the dataset that are not all relevant. Therefore, feature selection is performed to select the most influential attributes. By finding the best features, this feature selection process aims to improve model performance outcomes by removing features that aren't important. [2]. One statistical theory-based feature selection technique for determining a term's link to its category is chi square. Research conducted by [3], comparing Chi-Square feature selection with Mutual Information, showed that Chi-Square is more effective because it excels in average recall and computation time. One example of research by [4], also proved that Chi-Square feature selection improved accuracy from 77.08% to 83.33%.

In the data processing process, it was found that the dataset obtained had a class imbalance. Class imbalance occurs where one of the classes has a more dominant number of samples, causing the classification process to tend to predict the dominant class. Techniques that can be used to handle class imbalance can be done with SMOTE (Synthetic Minority Oversampling Technique) [5]. Research by [6] compared SMOTE and ADASYN, showing that SMOTE with SVM model attained a higher accuracy of 0.8901 compared to ADASYN which only reached 0.8871.

A well-known classification algorithm with computational speed and efficiency in handling large and complex data is LightGBM. The LightGBM algorithm is superior with faster computation time compared to the CatBoost and XGBoost algorithms [7]. Based on research by [8] demonstrates that, when it comes to the post-min aggregation approach, the LightGBM algorithm performs the best, with an accuracy of 0.858.

Improper selection of hyperparameters could affect the classification model’s output. Therefore, hyperparameter tuning is utilized to identify the ideal set of parameters. In this research, the Grid Search method is used as an optimization method that divides the parameters into equally spaced grid points. Using Random Search and Grid Search techniques, research by [9] compared the algorithms for Naive Bayes, Decision Tree and Random Forest. The findings indicate that hyperparameter tuning with Grid Search produces the best model with the greatest f1-score performance value of 0.868 compared to other methods.

Several studies have not specifically combined the LightGBM algorithm with the Chi-Square feature selection method for Parkinson's disease classification, according to previous research descriptions and findings. To enhance the model's performance, this research employs a feature reduction technique aimed at eliminating insignificant attributes by identifying those that are strongly associated with the target. The dataset utilized in this research was sourced from the UCI Machine Learning Repository, consisting of 756 records with 753 features. The classification target is to distinguish between persons diagnosed with Parkinson's disease and those who are healthy. Therefore, this research aims to implement the combination to improve classification performance with a more optimal accuracy value. It is anticipated that this research will also aid in the creation of a more precise and effective Parkinson's disease diagnosis assistance system.

2. METHODS

This research method provides a detailed description of the design of the methods used in this research. This research uses the LightGBM algorithm approach for parkinson's disease

classification with an emphasis on Chi-Square feature selection. The figure below demonstrates the phases of the study.

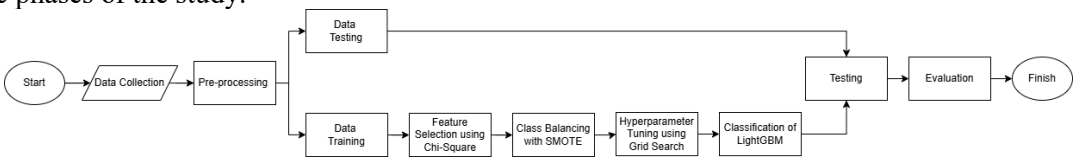

Figure 1 The Flowchart for Research

2.1 Data Collection

The dataset used for classifying Parkinson's Disease cases, was sourced from the UCI repository that houses machine learning datasets and is available through the following URL : https://archive.ics.uci.edu/dataset/470/parkinson+s+disease+classification [10]. The data collection includes a total of 756 records and 753 features. The dataset is the result of extraction from patient voice recordings including TWQT features, Wavelet Transform based Features, Time Frequency Features, Baseline Features, Vocal Fold Features, and Mel Frequency Cepstral Coefficients (MFCCs). The data is classified into two classes: class 0 for individuals who are not Parkinson's patients and class 1 for Parkinson's patients, with 192 and 564 records respectively.

2.2 Preprocessing

Before beginning the modelling process, the preprocessing step is completed to guarantee the quality of the data. The data preprocessing step carried out is checking for duplicate data which is then removed so as not to affect the data distribution and model training results. Following the completion of the cleaning phase, the data normalization process is also carried out. This normalization uses Min-Max to ensure all features are in the same range. The normalization stage is important to avoid the dominance of certain features that have a larger range of values in the model training process [11] . Equation (1) displays the formula of Min-Max normalization.

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

X_{sc} represents the normalization value, X indicates the value of the value in the dataset, X_{min} is a feature's lowest value, whereas X_{max} represents the feature's highest value.

2.3 Splitting Data

For this investigation, training and testing datasets are separated, with 90% going to training data and 10% going to testing data. The outcomes of classification performance are impacted by the choice of the proportion of data used for testing and training [12]. Inappropriate percentages can reduce model performance. When the proportions of testing and training data are compared in study [13], the 90%:10% ratio has the highest accuracy value of 78.40%.

2.4 Selection of Features Chi-square

Selection of features includes the process of selecting relevant features from a dataset. This process is used on high-dimensional data in order to find optimal features [14]. The Chi-Square test is one of the feature selection techniques employed. Machine learning models perform better when noise attributes are reduced using the Chi-Square feature selection method. This technique helps select the most suitable features by calculating the Chi-Square score for each feature and selecting features with high scores. Chi-Square is a tool for measuring the degree of statistical association between variables and evaluating the relationship between features in a dataset and specific classes or categories. Equation 2 is the Chi-Square formula according to reference:

$$x^2(t,c) = \frac{\mathrm{N}\,(\mathrm{AD}-\mathrm{CB})^2}{(\mathrm{A}+\mathrm{C})\,(B+D)\,(A+B)\,(C+D)} \quad (2)$$

Where t denotes the term (i.e., the feature being evaluated), c represents the class/category, A refers to the count of documents within class c that include term t, B represents the quantity of documents outside of class c but still containing term t, C denotes the count of documents in category c where term t is absent, D stands for the total documents outside class c that also do not include term t. N denotes the total count of documents used in the training process [15].

2.5 SMOTE

In this process the data will be balanced using SMOTE (Synthetic Minority Oversampling Technique) by creating synthetic data. The feature space's k-nearest neighbors are interpolated to create the synthetic or artificial data [16]. This SMOTE process aims to increase the data from the non-patient class because the number of samples is less than the Parkinson's disease patient class. This is done so that the amount of data from both classes in the dataset becomes equal. The function of the SMOTE method is expressed in the form SMOTE (X, N, k), where X is data from the less frequent class, N defines the amount of synthetic data required, and k represents the count of nearest samples to the target instance. The calculation of the separation between individual data is carried out the Euclidean Distance approach, which can be written in equation (3) below:

$$dist = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2} + \cdots + (X_n - Y_n)^2 \quad (3)$$

After calculating the closest distance using Euclidean Distance, the next step is to create replicate data from the closest new sample according to equation (4).

$$X_{syn} = X_i + (X_{knn} - X_i) \times \sigma \quad (4)$$

Where X_{syn} refers to synthetic data generated through replication, X_i represents the i-th instance form the underrepresented class, X_{knn} represents the closest neighboring sample from the same minority class to X_i, σ is a random number between 0 and 1.

2.6 LightGBM

One algorithm used in machine learning is called Light Gradient Boosting Machine recognized for its fast data processing capabilities and efficiency in managing large-scale datasets. LightGBM utilizes the gradient boosting technique by using a leaf-wise approach in growing the decision tree vertically [17]. This approach significantly speeds up the training process. In addition, LGBM does not use the actual data values directly. Instead, it builds a histogram of feature values. This technique not only reduces memory usage, but also enhances the training process efficiency, making it well-suited for handling large volumes of data.

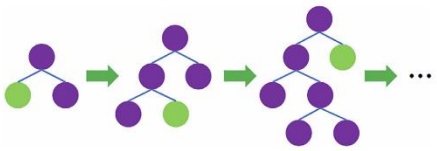

Figure 2 Depth-based tree growth in the LightGBM algorithm

2.7 Hyperparameter Tuning

The best hyperparameter combination is found through hyperparameter tuning during the data training phase. Choosing the right hyperparameters can affect training speed, prediction accuracy, and how effectively the model can apply its learning to novel data. By finding the optimal combination, the model can work more efficiently and produce more accurate predictions.

In the process of training datasets, this technique is very important because it is useful for obtaining optimal machine learning models depending on the attributes of the information being used [18].

In this research, the search for the best hyperparameter values was conducted using grid search. Grid Search is a hyperparameter optimization technique that performs a thorough search through a combination of predetermined values for each hyperparameter [19]. The hyperparameters used along with the description and range of test values are explained in Table 1.

Table 1 Hyperparameter LightGBM

Hyperparameter	Description	Value Tested
n_estimators	The quantity of separate trees that the tree boosting model uses	[100, 200, 300]
learning_rate	The magnitude of weight adjustment applied at each iteration during model updating	[0.01, 0.05, 0.1]
num_leaves	The most leaves that each decision tree can have	[31, 50, 70]
max_depth	The decision tree's deepest point	[3, 5, 7]

2.8 Evaluation

The classification model's effectiveness in this research is evaluated through the use of a Confusion Matrix. This matrix displays a comparison between the actual labels and the model's predicted classifications in a tabular format. The model prediction outcomes, which include TP (True Positive), FN (False Negative), TN (True Negative), and FP (False Positive), are shown in this table. The classification model's performance is assessed using the confusion matrix function, which displays the number of examples the model correctly or incorrectly predicts [20]. This can be shown in Table 2.

Table 2 Confusion Matrix

Actual Class	Predicted Class	
	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

A situation for which the reference data has a positive label and is classified as positive is known as a TP (True Positive). An instance that has an indication of positivity in the reference data but a negative label from the classifier is known as a FN (False Negative). False Positive (FP) is a case that has a negative label in the reference data but a positive label from the classifier. While, TN (True Negative) is a case that is classified as negative and also has a negative label in the reference data [21].

Classification accuracy has a prediction of the correct value (positive value and negative value). A classification's overall quality is gauged by its accuracy. Equation (5) can be applied to ascertain the accuracy value.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (5)$$

3. RESULTS AND DISCUSSION

The discussion will focus on each stage that has been carried out, starting from the data preprocessing process, data division, feature selection using the Chi-Square method, data balancing with SMOTE, to the application of the LightGBM algorithm combined with the

hyperparameter tuning process. Each result obtained will be analyzed to see the effect of each stage on model performance, especially in improving the accuracy value as the main evaluation metric.

3.1 Preprocessing

The initial stage of preprocessing begins with a check for the presence of duplicate data in the dataset. This was done using Python's 'duplicated()' function, which allows identification of rows with similar values in all columns. The result showed that there was 1 duplicate data record. To prevent potential bias as well as the risk of overfitting that could affect model performance, the duplicate data was removed from the dataset.

Data normalization comes next. Each feature value in the dataset is normalized in this study to fall between 0 and 1 using the min-max scaler approach. Table 3 presents a sample of the original dataset used in this study before the normalization process. The dataset representing various biomedical voice measurements and signal processing attributes. These features are relevant for detecting vocal and motor impairments commonly observed in Parkinson's disease patients.

Table 3 Parkinson's Disease Dataset

Id	Gender	PPE	DFA	...	tqwt_dec_36
0	1	0.85247	0.71826	...	18.9405
0	1	0.76686	0.69481	...	45.1780
0	1	0.85083	0.67604	...	4.7666
1	0	0.41121	0.79672	...	4.0603
1	0	0.32790	0.79782	...	6.1164
...	...	...	...	...	...
251	0	0.81304	0.76471	...	3.1527

This is after normalizing the data using the min-max scaler in table 4:

Table 4 Parkinson's Disease Dataset Min-Max Scaler Normalization Result

Gender	PPE	DFA	...	tqwt_dec_36
1.0	0.936	0.565	...	0.107
1.0	0.837	0.489	...	0.277
1.0	0.934	0.428	...	0.015
0.0	0.426	0.819	...	0.011
0.0	0.330	0.822	...	0.024
...	...	...	...	...
0.0	0.890	0.715	...	0.005

3.2 Feature Selection

This study employs the Chi-Square Test technique to select relevant features by eliminating features that are less significant to the classification process. The feature selection process is carried out by setting various numbers of features. Overall, 10 experiments were conducted with variations in the number of features ranging from 5 to 250. Table 5 displays the findings of the comparison of the accuracy values acquired according to the quantity of characteristics.

Table 5 Accuracy Value Compared to Feature Selection Number

Features	Accuracy (%)
5	78,95%
10	78,95%
20	81,58%
30	89,47%
40	89,47%
50	93,42%
100	90,79%
150	94,74%
200	96,05%
250	93,42%

The accuracy value obtained is affected by variations in the number of characteristics, according to the test results shown in Table 5. With 200 features, the best accuracy value is 96.05%. Meanwhile, using 250 features, accuracy decreased to 93.42%. his indicates that the excessive addition of features can introduce noise and less relevant information, thereby reducing the model's performance. Therefore, selecting 200 features represents the optimal balance between maximizing accuracy and minimizing feature redundancy.

3.3 SMOTE Result

At this stage, data imbalance is handled using the SMOTE method. This method creates artificial data by interpolating from the feature space's k-nearest neighbors [16]. This research uses the Imblearn library available in Python. The SMOTE approach ensures that the model is not biased in support of the majority class. The class distribution before and after using the SMOTE approach is shown in the accompanying picture.

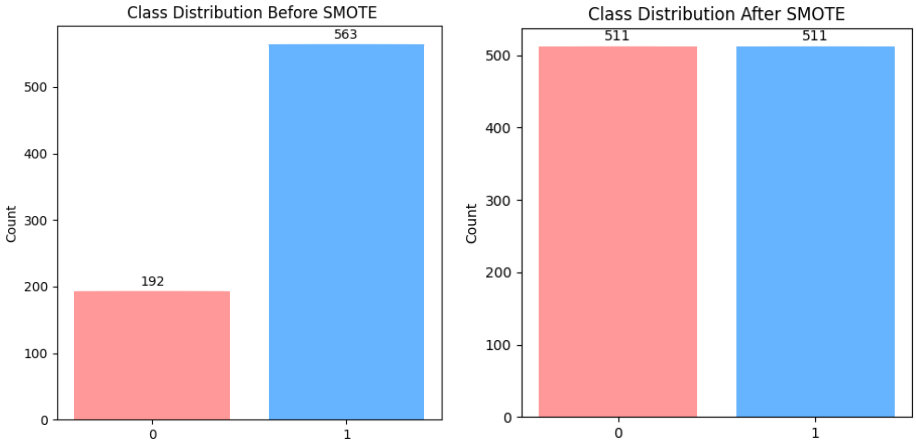

Figure 3 Class distribution comparison before and after SMOTE implementation

3.4 Hyperparameter Tuning Result

The hyperparameter tuning process in this study uses Grid Search with the aim of obtaining the optimal parameter combination that can maximize model performance. Grid Search works by evaluating each predefined parameter combination, then the parameters are divided into points in the same grid-shaped search space. Each combination is tested through the classification model training process. The hyperparameter tuning results are shown in Table 6 below.

Table 6 Best Hyperparameters for LightGBM

Parameter	Value
n_estimator	300
learning_rate	0.1
num_leaves	31
max_depth	7

3.5 Model Comparison

Testing this model using the accuracy value, Figure 4 presents a comparison of model performance from several previous studies and the results of this study.

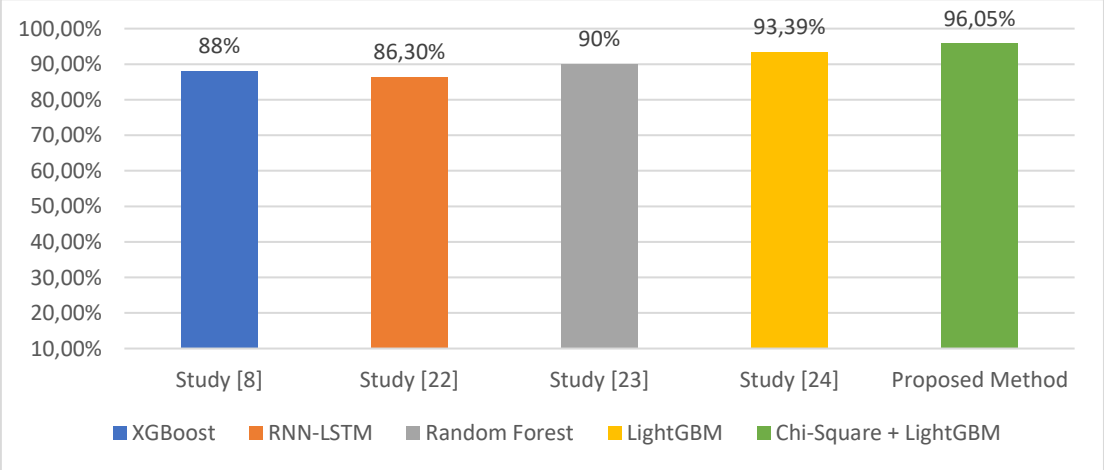

Figure 4. Model Comparison

Figure 4 illustrates that the suggested model, which combines the LightGBM method with Chi-Square feature selection, performs best, with an accuracy value of 96.05%. This figure exceeds the findings of study by [8] using XGBoost with an accuracy of 88%, and research [22] using the RNN-LSTM algorithm with an accuracy value of 86.30%. In addition, research [23] the Random Forest algorithm yielded a 90% accuracy rate and research [24] an accuracy value of 93.39% was obtained by utilizing LightGBM alone, without feature selection.

These results indicate that the implementation of the Chi-Square feature selection process effectively contributes to reducing irrelevant features. This approach not only achieves higher accuracy but also reduces computational complexity, thereby enhancing the model's performance in Parkinson's disease classification. However, this study has limitations, particularly the use of a single dataset and reliance on only one feature selection method. For future directions, it is recommended to validate the proposed model using diverse datasets and explore other feature selection techniques.

4. CONCLUSIONS

The application of a combination of feature selection utilizing the Chi-Square Test with the LightGBM algorithm offers improved performance in the categorization of Parkinson's disease, according to the findings and discussions that have been conducted. Classification testing on variations in the number of features, namely 5, 10, 20, 30, 40, 50, 100, 150, 200, 250, shows that the use of 200 features provides the highest accuracy value of 96.05%. In addition, the comparison results between using additional Chi-Square-based feature reduction and not applying it demonstrate that selecting features through the Chi-Square method leads to better performance than using LightGBM without any feature selection. As a future development, it is recommended to explore the use of other classification algorithms as well as different feature selection methods

to obtain a more optimal combination of features, so as to further enhance the model's capability for Parkinson's disease classification.

However, this study has several limitations, such as the use of a single dataset and reliance on one feature selection technique, which may affect the generalization of the results. As a direction for future development, it is recommended to validate the model on larger and multiple datasets, and to explore the use of different classification algorithms as well as alternative feature selection methods. These efforts aim to obtain a more optimal feature combination and further enhance the model's performance in Parkinson's disease classification.

ACKNOWLEDGMENTS

For the direction and counsel given during the study process, the author would like to sincerely thank the supervising instructor. Additionally, gratitude is given to Lambung Mangkurat University for all of their assistance that made this research possible.

REFERENCES

[1] I. S. Zein And K. Khairunnisa, "Parkinson Disease," *Jurnal Riset Rumpun Ilmu Kedokteran (Jurrike)*, Vol. 2, No. 2, Pp. 50–63, Oct. 2023, Doi: 10.55606/Jurrike.V2i2.1701.

[2] D. Kurnia, M. Itqan Mazdadi, D. Kartini, R. Adi Nugroho, And F. Abadi, "Seleksi Fitur Dengan Particle Swarm Optimization Pada Klasifikasi Penyakit Parkinson Menggunakan Xgboost," *Jurnal Teknologi Informasi Dan Ilmu Komputer*, Vol. 10, Pp. 1083–1094, 2023, Doi: 10.2516/Jtiik.2023107252.

[3] A. Rahmadeyan And M. Mustakim, "Seleksi Fitur Pada Supervised Learning: Klasifikasi Prestasi Belajar Mahasiswa Saat Dan Pasca Pandemi Covid-19," *Jurnal Nasional Teknologi Dan Sistem Informasi*, Vol. 9, No. 1, Pp. 21–32, May 2023, Doi: 10.25077/Teknosi.V9i1.2023.21-32.

[4] A. Purnamawati, M. N. Winarto, And M. Mailasari, "Analisis Sentimen Aplikasi Tiktok Menggunakan Metode Bm25 Dan Improved K-Nn Fitur Chi-Square," *Jurnal Komtika (Komputasi Dan Informatika)*, Vol. 7, No. 1, Pp. 97–105, May 2023, Doi: 10.31603/Komtika.V7i1.8938.

[5] R. Aryanti, T. Misriati, And A. Sagiyanto, "Analisis Sentimen Aplikasi Primaku Menggunakan Algoritma Random Forest Dan Smote Untuk Mengatasi Ketidakseimbangan Data," *Journal Of Computer System And Informatics (Josyc)*, Vol. 5, No. 1, Pp. 218–227, Nov. 2023, Doi: 10.47065/Josyc.V5i1.4562.

[6] A. Nurhopipah And C. Magnolia, "Perbandingan Metode Resampling Pada Imbalanced Dataset Untuk Klasifikasi Komentar Program Mbkm," *Jurnal Publikasi Ilmu Komputer Dan Multimedia*, Vol. 1, No. 2, Pp. 9–22, 2022.

[7] M. Bahril Ilmi And Kusrini, "Perbandingan Kinerja Algoritma Machine Learning Dalam Deteksi Potensi Risiko Hiv," *Jurnal Buffer Informatika*, Vol. 11, No. 1, Apr. 2025, [Online]. Available: Https://Journal.Fkom.Uniku.Ac.Id/Buffer

[8] Z. Yang *Et Al.*, "Optimizing Parkinson's Disease Prediction: A Comparative Analysis Of Data Aggregation Methods Using Multiple Voice Recordings Via An Automated Artificial Intelligence Pipeline," *Data (Basel)*, Vol. 10, No. 1, Jan. 2025, Doi: 10.3390/Data10010004.

[9] J. Dwi Muthohhar And A. Prihanto, "Analisis Perbandingan Algoritma Klasifikasi Untuk Penyakit Jantung," *Journal Of Informatics And Computer Science*, Vol. 04, 2023.

[10] C. Sakar, G. Serbes, A. Gunduz, H. Nizam, And B. Sakar. "Parkinson's Disease Classification," Uci Machine Learning Repository. [Online]. Available: Https://Doi.Org/10.24432/C5ms4x

[11] G. T. Adewale, A. U. Victor, A. E. Sylvia, T. Sonubi, And A. O. Mesogboriwon, "Integrating Big Data And Machine Learning In Management Information Systems For

Predictive Analytics: A Focus On Data Preprocessing And Technological Advancements," *World Journal Of Advanced Research And Reviews*, Vol. 24, No. 2, Pp. 774–789, Nov. 2024, Doi: 10.30574/Wjarr.2024.24.2.3427.

[12] H. Bichri, A. Chergui, And M. Hain, "Investigating The Impact Of Train / Test Split Ratio On The Performance Of Pre-Trained Models With Custom Datasets," 2024. [Online]. Available: Www.Ijacsa.Thesai.Org

[13] B. N. Azmi, A. Hermawan, And D. Avianto, "Analisis Pengaruh Komposisi Data Training Dan Data Testing Pada Penggunaan Pca Dan Algoritma Decision Tree Untuk Klasifikasi Penderita Penyakit Liver," *Jtim : Jurnal Teknologi Informasi Dan Multimedia*, Vol. 4, No. 4, Pp. 281–290, Feb. 2023, Doi: 10.35746/Jtim.V4i4.298.

[14] T. Ernayanti, M. Mustafid, A. Rusgiyono, And A. R. Hakim, "Penggunaan Seleksi Fitur Chi-Square Dan Algoritma Multinomial Naïve Bayes Untuk Analisis Sentimen Pelangggan Tokopedia," *Jurnal Gaussian*, Vol. 11, No. 4, Pp. 562–571, Feb. 2023, Doi: 10.14710/J.Gauss.11.4.562-571.

[15] D. Chen Sami, A. Sugiharto, And F. Jie, "Chi Square Feature Selection For Improving Sentiment Analysis Of News Data Privacy Treats," *J Theor Appl Inf Technol*, Vol. 102, No. 18, 2024, [Online]. Available: Www.Jatit.Org

[16] Y. B. Wah *Et Al.*, "Machine Learning And Synthetic Minority Oversampling Techniques For Imbalanced Data: Improving Machine Failure Prediction," *Computers, Materials And Continua*, Vol. 75, No. 3, Pp. 4821–4841, 2023, Doi: 10.32604/Cmc.2023.034470.

[17] R. Sibindi, R. W. Mwangi, And A. G. Waititu, "A Boosting Ensemble Learning Based Hybrid Light Gradient Boosting Machine And Extreme Gradient Boosting Model For Predicting House Prices," *Engineering Reports*, Vol. 5, No. 4, Apr. 2023, Doi: 10.1002/Eng2.12599.

[18] D. S. Bhakti, A. Prasetyo, And P. Arsi, "Implementation Of Hyperparameter Tuning In Random Forest Algorithm For Loan Approval Prediction," *Jurnal Teknik Informatika (Jutif)*, Vol. 5, No. 4, Pp. 63–69, 2024, Doi: 10.52436/1.Jutif.2024.5.4.2032.

[19] D. M. Belete And M. D. Huchaiah, "Grid Search In Hyperparameter Optimization Of Machine Learning Models For Prediction Of Hiv/Aids Test Results," *International Journal Of Computers And Applications*, Vol. 44, No. 9, Pp. 875–886, 2022, Doi: 10.1080/1206212x.2021.1974663.

[20] G. M. Foody, "Challenges In The Real World Use Of Classification Accuracy Metrics: From Recall And Precision To The Matthews Correlation Coefficient," *Plos One*, Vol. 18, No. 10 October, Oct. 2023, Doi: 10.1371/Journal.Pone.0291908.

[21] S. Suryanto And W. Andriyani, "Sentiment Analysis Of X Platform On Viral 'Fufufafa' Account Issue In Indonesia Using Svm," *Ijccs (Indonesian Journal Of Computing And Cybernetics Systems)*, Vol. 19, No. 1, P. 95, Jan. 2025, Doi: 10.22146/Ijccs.104158.

[22] S. Chandrabhanu And S. Hemalatha, "Cgan Facilitated Data Augmentation Of Voice And Speech Parameters For Detecting Parkinson's Disease In The Prodromal Phase," *Brain (Bacau)*, Vol. 15, No. 3, Pp. 208–222, 2024.

[23] O. M. El-Habbak *Et Al.*, "Enhancing Parkinson's Disease Diagnosis Accuracy Through Speech Signal Algorithm Modeling," *Computers, Materials And Continua*, Vol. 70, No. 2, Pp. 2953–2969, 2022, Doi: 10.32604/Cmc.2022.020109.

[24] M. M. Nishat, T. Hasan, S. M. Nasrullah, F. Faisal, M. A. A. R. Asif, And M. A. Hoque, "Detection Of Parkinson's Disease By Employing Boosting Algorithms," In *2021 Joint 10th International Conference On Informatics, Electronics And Vision, Iciev 2021 And 2021 5th International Conference On Imaging, Vision And Pattern Recognition, Icivpr 2021*, Institute Of Electrical And Electronics Engineers Inc., 2021. Doi: 10.1109/Icievicivpr52578.2021.9564108.