EARLY DETECTION OF PARKINSON'S DISEASE USING MACHINE LEARNING

Aliviya Jana, P Subanaveen, Sanjana Lakkimsetty, M Lahari Priya,
aliviya_jana@srmap.edu.in, subanaveen_p@srmap.edu.in, sanjana_l@srmap.edu.in, laharipriya_m@srmap.edu.in
SRM University AP, Mangalagiri -Mandal, Neeru Konda, Amaravati, Andhra Pradesh 522502

Under the guidance of Dr. Ashok Kumar Pradhan, Department of Computer Science Engineering, SRM University AP.

***Abstract*—In this research paper, we tackle the challenge of accurately diagnosing Parkinson's disease (PD) using machine learning (ML) techniques, with a specific focus on addressing imbalanced datasets. We employ Adaptive Synthetic Sampling (ADASYN) to intelligently balance class representation, ensuring that minority groups, which are crucial for precise PD detection, are included. Additionally, we utilize min-max scaling to rescale features and incorporate various ML models, such as XGBoost, to leverage their unique strengths. Our findings underscore the effectiveness of this integrated approach in accurately identifying Parkinson's disease. Evaluation metrics, including accuracy, precision, recall, and F1 score, demonstrate the robust performance of our model. Visualization tools like the Confusion Matrix and Receiver Operating Characteristic (ROC) curve provide detailed insights into the capabilities of our model and areas for improvement. Significantly, our model achieves exceptional accuracy (97.44%) and precision (100%) in detecting Parkinson's disease, surpassing alternative algorithms. The incorporation of ADASYN to address class imbalance greatly enhances the performance of our model, highlighting its suitability for applications that require precise identification of minority classes. Overall, our research contributes to the advancement of ML-based diagnosis of Parkinson's disease, with potential implications for improved patient outcomes and personalized treatment strategies.**

***Index Terms*—Parkinson's Disease, Machine Learning, ADASYN, Min-Max, XGBoost, Cross-folds, Ensemble Learning.**

I. Introduction

Parkinson's disease(PD) is a complex neurodegenerative disease that impacts millions across the world, leading to gradual deterioration and eventual death of the brain nerve cells and peripheral nervous system. The effects of PD are multifaceted which often results in a range of symptoms such as diminished motor function, tremors, balance impairments and more primarily attributed to the gradual decline of the levels of dopamine in the brain. The onset of PD commonly occurs during early old age with initial manifestations which include a combination of motor and non-motor symptoms like olfactory dysfunction, drooling, constipation, gastrointestinal issues, restless leg syndrome, abnormal gait [7], [13], Bradykinesia and tremors.

In 1967, Hoehn and Yahr introduced a classification system dividing PD into five stages: stages one and two represent the early phase stages, two and three symbolize the middle phase and, stages four and five characterize the advanced phase of the PD. Stage one manifests as mild symptoms that do not significantly hinder daily life such as slight alterations in posture facial expressions [23] and gait, as the progress of the disease goes to the second stage symptoms tend to intensify potentially affecting daily activities and the body's mid-line. The mid-stage or stage three is characterized by an increasing loss of balance and the potential for disability. In stage four, individuals often rely on assistive devices like canes or walkers while stage five represents a severe stage where leg stiffness becomes pronounced and mobility without assistance becomes nearly impossible.

PD presents a preclinical phase where the degeneration of dopamine-producing neurons initiates [3] , though clinical symptoms are not yet evident. The prodromal phase involves the presence of certain symptoms that are insufficient for a definitive PD diagnosis. The clinical phase is marked by the clear and identifiable manifestation of PD symptoms. While several hypotheses associate the development of PD with various factors such as exposure to pesticides, the only confirmed cause remains genetic inheritance. When PD occurs without a genetic link, it is categorized as idiopathic, indicating that its origins remain unknown. Despite ongoing research efforts, there is currently no definitive method for predicting the onset of PD. Diagnosis typically occurs during the clinical phase, involving an array of techniques ranging from blood tests, computerized tomography (CT) scans, genetic testing, and magnetic resonance imaging (MRI), to positron emission tomography (PET) scans. [10], [15] And in our paper voice data has been taken for consideration. [4], [5], [19]

To ensure the dataset is conducive to model training, a comprehensive preprocessing phase is undertaken. Leveraging the Pandas library, the Parkinson's disease dataset is meticulously loaded, and the features and corresponding labels are extracted. To standardize the scale of the various features, the application of the Min-Max scaling technique is important. [8] This step facilitates the transformation of the dataset's values to a standardized range between -1 and 1, effectively averting any singular feature from disproportionately influencing the process. Furthermore, the division of the dataset into distinct training and testing sets enables evaluation of the model's performance on unseen data, thus ensuring its generalizability.

Given the commonplace occurrence of class imbalance

within medical datasets, a pivotal aspect of this study involves the implementation of the Adaptive Synthetic Sampling (ADASYN) algorithm. [14] This technique is instrumental in the generation of synthetic samples for the minority class, primarily focusing on individuals diagnosed with Parkinson's disease. By oversampling the underrepresented class, the resulting balanced training dataset equips the model to learn from both the majority and minority classes, thereby significantly reducing the likelihood of biased predictions during the model training phase.

The focal point of our research revolves around the utilization of an XGBoost classifier, configured with an array of appropriate parameters such as the evaluation metric, label encoding utilization, the defined objective function, the determined number of estimators, and the specified maximum depth. The model undergoes rigorous training on the meticulously curated and balanced dataset resulting from the application of the ADASYN algorithm. Rigorous evaluation of the model's performance is executed utilizing standard classification metrics including accuracy, precision, recall, and the F1 score. This evaluation process provides deep insights into the model's efficacy in accurately discerning between individuals afflicted with Parkinson's disease and those who are not.

For the robustness of the model and to effectively mitigate potential instances of overfitting, the implementation of the Stratified K-Fold cross-validation methodology is paramount. [16] This cross-validation approach facilitates the meticulous division of the dataset into 'k' subsets [2], [5], [6], whilst concurrently ensuring the preservation of the class distribution within each fold. Within the confines of each fold, the ADASYN algorithm is once again applied to effectively address any existing class imbalance, subsequently culminating in the training and evaluation of a fresh instance of the XGBoost model. This multifaceted process not only provides an all-encompassing evaluation of the model's performance across diverse segments of the dataset but also serves to enable more reliable estimation of its overarching generalizability and predictive capability.

The empirical results derived from our comprehensive study underscore the profound effectiveness of the XGBoost algorithm in the accurate prediction and classification of Parkinson's disease, leveraging a diverse array of clinical features. [**?**] The seamless integration of the ADASYN algorithm to effectively tackle class imbalance, coupled with the utilization of cross-validation methodologies, serves to underscore the robustness and reliability of our model's performance evaluation. [17] The implications of our findings hold the potential to significantly augment the capabilities of machine learning algorithms in assisting clinicians with the early and precise diagnosis of Parkinson's disease, thereby paving the way for the development of effective interventions and highly personalized treatment strategies for individuals afflicted with this debilitating condition.

II. Contribution

The main focus of this paper is to detect Parkinson's Disease (PD) at an early stage by analyzing voice recordings. [4], [6], [20] This is achieved by utilizing key attributes such as jitter, shimmer, fo, fhi, and flo. To enhance the accuracy of PD detection, machine learning techniques are integrated, resulting in the following notable contributions:

1. In order to normalize feature ranges and expedite the model training process, we have adopted the min-max scaling technique.

2. To tackle the issue of data imbalance, we have employed the ADASYN technique. This technique generates synthetic samples for the minority class.

3. Our approach leverages the XGBoost algorithm by utilizing this algorithm, we are able to achieve effective predictive modeling, leading to improved accuracy and reliability.

4. We have conducted a thorough evaluation of the models, considering critical performance metrics such as accuracy, precision, recall, and F1-score. This evaluation provides a comprehensive understanding of the effectiveness of the models.

These contributions collectively advance the field of early PD detection. The subsequent sections of the paper delve into the PD dataset (Section IV) and provide a concise overview of the proposed machine learning models (Section VI). These sections offer detailed insights into the dataset used for analysis and the methodologies employed for PD detection using machine learning techniques.

III. DATASET

The dataset considered contains a wide range of data having 180 data values and 20 attributes associated with individuals, including their names and several acoustic measurements related to voice characteristics [12], [19], [20]. These acoustic features encompass various fundamental frequency parameters, such as MDVP:Fo(Hz), MDVP:Fhi(Hz), and MDVP:Flo(Hz), which provide insights into different aspects of the voice's fundamental frequency. Additionally, there are measurements related to voice jitter (MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, Jitter:DDP) and voice shimmer (MDVP:Shimmer, MDVP:Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA), which offer valuable information about the stability and variability of the voice signal. Furthermore, parameters like NHR (Noise-to-Harmonics Ratio), HNR (Harmonics-to-Noise Ratio), and various nonlinear dynamical features (status, RPDE, DFA, spread1, spread2, D2, PPE) contribute to a comprehensive set of attributes. These features are particularly useful for voice and speech analysis, as they enable researchers to explore potential patterns or correlations with the 'status' attribute, which likely serves as an indicator of the individuals' health status refer

IV. ALGORITHM

1) The Parkinson's disease dataset is obtained from PPMI UCI and is pre-processed for analysis.

No	Description	Voice Measure
1	Mean absolute difference of pitch	MDVP:Fo(Hz)
2	High frequency component	MDVP:Fhi(Hz)
3	Low frequency component	MDVP:Flo(Hz)
4	Frequency variation in voice	MDVP:Jitter(%)
5	Absolute jitter	MDVP:Jitter
6	Pitch period entropy	MDVP:RAP
7	Pitch period perturbation quotient	MDVP:PPQ
8	Jitter difference between consecutive periods	Jitter: DDP
9	Amplitude variation in voice	MDVP:Shimmer
10	Shimmer in decibels	MDVP:Shimmer(dB)
11	Amplitude perturbation quotient, 3rd period	Shimmer:APQ3
12	5th period	Shimmer:APQ5
13	Mean absolute perturbation quotient	MDVP:APQ
14	Shimmer difference	Shimmer:DDA
15	Noise to harmonics ratio	NHR
16	Harmonics to noise ratio	HNR
17	Result status	Status
18	Recurrence period density entropy	RPDE
19	Detrended fluctuation analysis	DFA
20	Frequency spread 1	Spread1
21	Frequency spread 2	Spread2
22	Correlation Dimension	D2
23	Pitch period entropy	PPE

2) Features (independent variables) and labels (dependent variables) are identified within the dataset, the foundation for classification.
3) Feature values are normalized using Min-Max scaling to ensure a consistent range, optimizing between -1 to 1.
4) The dataset is partitioned into training and testing subsets with an 80-20 split ratio.
5) ADASYN is applied on the training data to balance the class, generating synthetic samples for the minority class.
6) An XGBoost classifier is trained using the training dataset which acts as a power for gradient boosting.
7) The trained model is evaluated on the testing set, and key performance metrics including accuracy, precision, recall, and F1 score are computed.

V. METHODOLOGY

As the dataset doesn't contain any missing values, the data is preprocessed using ADASYN to handle imbalance and Min-Max Scling to range it accordingly. Then, the aim of classification is to find a model that describes and, at the same time, distinguishes classes of data, and then uses it to predict the class to which an unclassified object will belong. Classification is a process that allows data to be divided into given classes based on their properties. The classification process takes place in the following steps, Training based on the analysis of the classification model created in the training set; and Testing, that is, evaluation of the quality of the created model using test data.

A. ADASYN

The adaptive synthetic sampling algorithm known as ADASYN was created to solve the problem of unbalanced datasets in machine learning [16]. ADASYN aims to balance the class distribution by creating synthetic examples for the minority class.The adaptive nature of ADASYN lies in its focus on producing more artificial samples for more challenging cases to learn.

$$x(syn) = x(min) + \lambda * (\text{nearest neighbour} - x(min))$$

Here,
• x(syn) is the synthetic sample generated.
• x(min) is the minority class we want to generate the synthetic sample for.
• λ is a random number between 0 to 1 which controls the amount of synthetic data generated.
• The nearest neighbour is randomly selected of x(min).

Finding the dataset's level of imbalance is the first step taken by ADASYN[14]. Based on the separations between each instance in the minority class and its k-nearest neighbours, it computes a density distribution for each instance. Less dense instances—which are found in denser areas of the feature space—are thought to be more difficult to learn [9]. Next, using the determined density distribution as a guide, synthetic samples are created for the minority class instances. By concentrating on instances with lower densities, the generating process highlights more difficult situations.

When working with unbalanced datasets, ADASYN may greatly enhance a machine learning model's performance. The model gains improved generalisation and prediction skills by learning the features of the underrepresented class through the creation of synthetic examples for the minority class [16], [18].

B. XGBoost

Also known as eXtreme Gradient Boosting, or XGBoost, is a well-liked and potent machine learning algorithm that falls under the group of ensemble learning techniques. As an ensemble learning method, integrates the predictions of several different models to provide a final forecast that is more reliable and accurate. It belongs to the class of boosting techniques, in which every model in the ensemble fixes the mistakes made by the model before it. [21] Regularisation techniques are used into XGBoost to manage model complexity and avoid over-fitting. It regulates the depth of each decision tree in the ensemble using a method known as tree pruning. By keeping the trees from becoming too large and overfitting the training set, pruning helps.It comes with built-in features to deal with missing data. It automatically picks up handling missing values during training, negating the requirement for preprocessing operations to impute or remove missing data. [12], [18].

C. Hyperparameter Tuning

A large number of hyperparameters are available in XGBoost that may be adjusted to maximise the performance of the model. The number of trees in the ensemble, tree depth, and learning rate are examples of common hyperparameters. Optimising hyperparameters is essential to get optimal

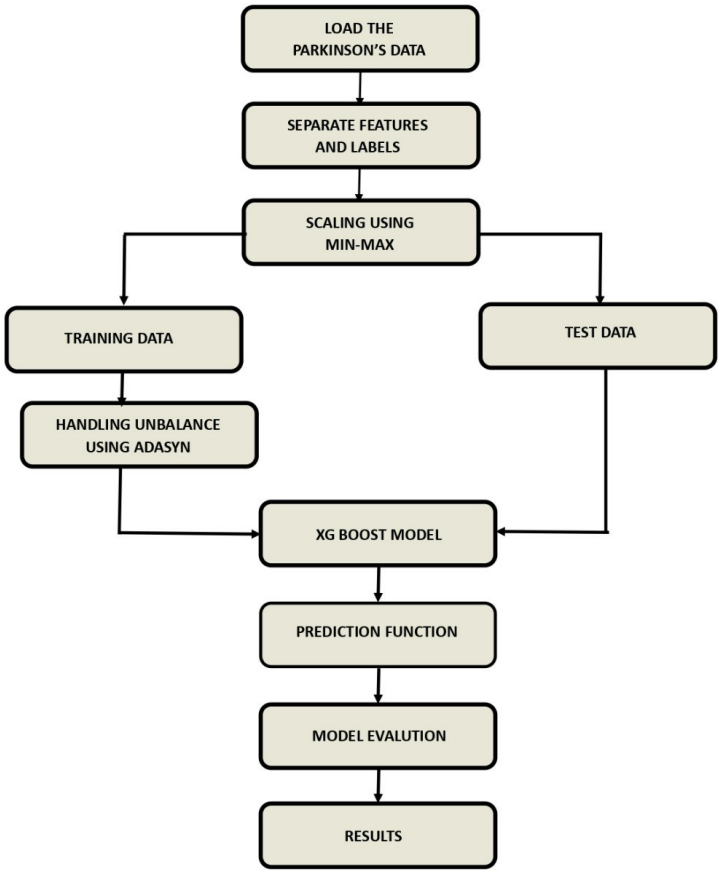

Fig. 1. Flowchart of the algorithm

predictive efficacy.

Choice of Evaluation Metrics:

Precision: Precision guides the navigation of positive forecasts' accuracy.

Recall: Recall replicates the model's ability to capture all forms of Parkinson's disease by taking real positive examples.

F1-Score: A balance for false positives and false negatives, the F1-Score is of recall and precision.

AUC-ROC or Area Under the ROC Curve It is done with positive and negative examples.

Specificity: It complements recall for the negative class smoothly.

VI. RESULTS

Evaluating the performance of a model designed to detect Parkinson's disease involves several key metrics, each providing valuable insights into its effectiveness. So, if our model has high accuracy, precision, recall, and F1 score, it suggests that it is performing well in identifying whether an individual has Parkinson's disease or not.

In aspects of the model's performance. Some of the common graphs are:

1.Confusion Matrix: It plays a pivotal role as a fundamental tool for assessing the model's performance. The table meticulously captures the intricate details of true positives, true negatives, false positives, and false negatives. Within the scope of our research, the model demonstrated precision by correctly identifying 27 cases as not having Parkinson's disease and accurately recognizing 9 cases where Parkinson's disease was present. Despite these successes, the model exhibited fallibility with 2 instances of misjudging Parkinson's disease and a regrettable oversight where it failed to detect 1 case of Parkinson's disease.

2. The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between the true positive rate and the false positive rate for different threshold values. Fig. 4

In terms of model evaluation metrics, our model exhibits exceptional performance, boasting an accuracy of 97.44 percent. Specifically, precision is at a perfect 100.00 percent, indicating that the model almost exclusively predicts the minority class when it is genuinely present. Additionally, the recall stands at an impressive 96.88 percent, highlighting the model's capability to capture a significant proportion of actual positive instances. The F1 score, a harmonized measure of precision and recall, further supports the model's balanced performance, registering at a notable 98.41 percent.[fig.6]

In a comparative analysis against other machine learning algorithms [17], our model stands out with its superior accuracy. While achieving a commendable 97.44 percent accuracy, competing models such as K-NN, Naïve Bayes, Random Forest, and SVM [1], [22] lag behind with accuracies of 92.31 percent, 69.23 percent,87.18 percent and 94.87 percent, respectively.[fig.7]

.

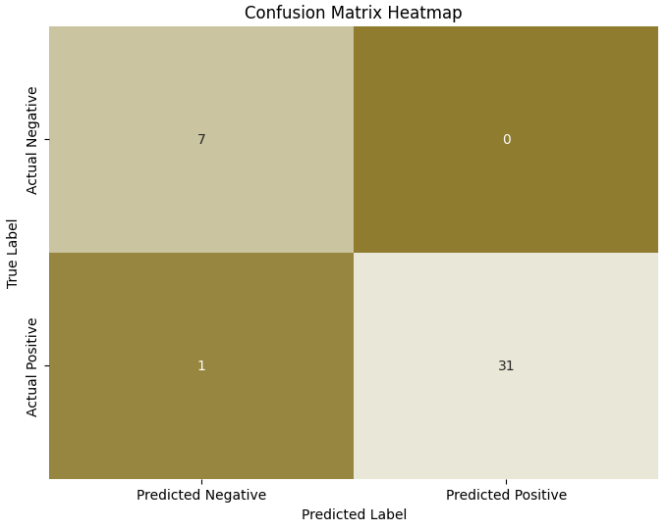

Fig. 2. Confusion Matrix of Predicted vs Actual

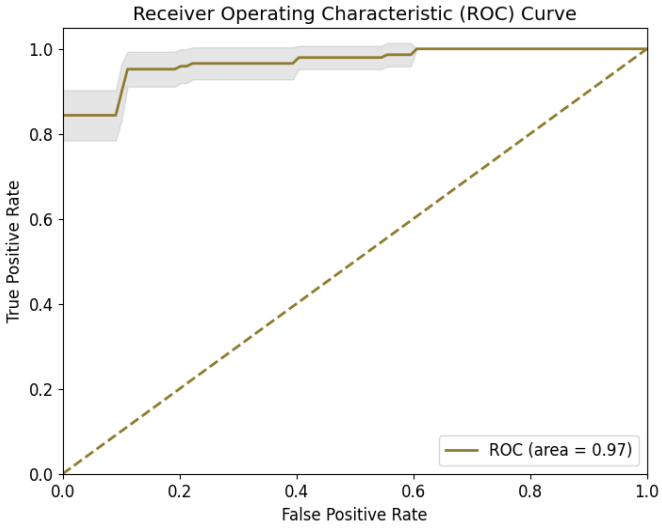

Fig. 3. ROC Curve

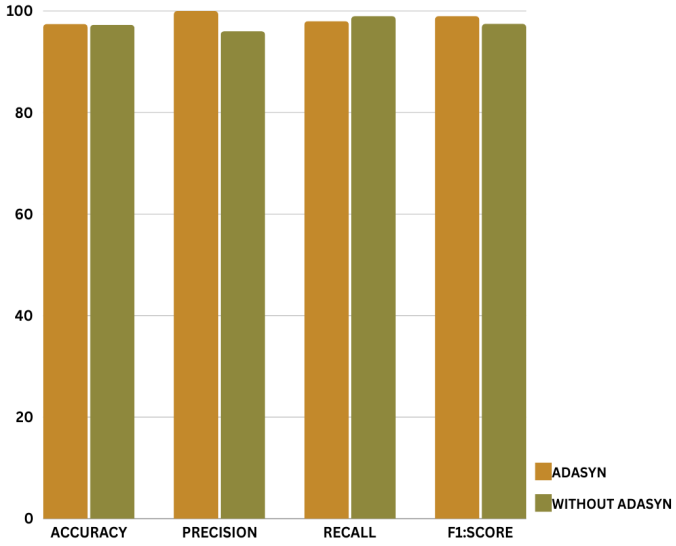

Fig. 4. Markers With ADASYN vs Without ADASYN

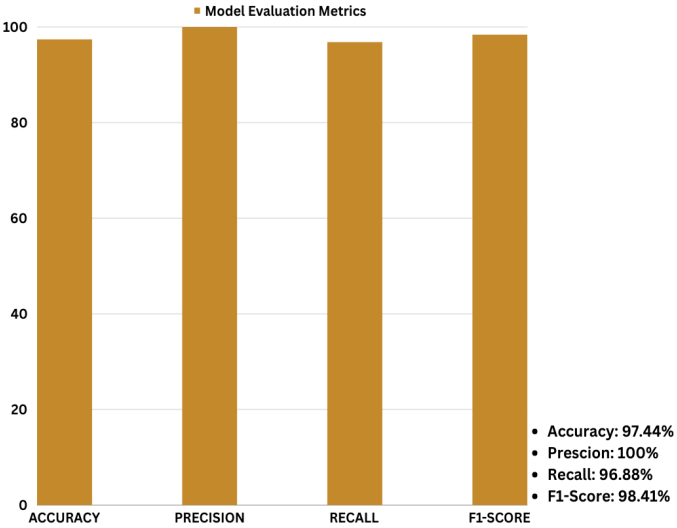

Fig. 5. Accuracy, Precision, Recall and F1 Score

VII. CONCLUSION

We would like to finally conclude this paper by discussing the overview of the whole process. In order to tackle the issue of imbalanced datasets in the pursuit of machine learning-based early Parkinson's disease detection, our study utilised the adaptive synthetic sampling technique, ADASYN. By producing synthetic cases for the minority class, this algorithm—which was created to address class imbalances in binary classification problems— helped create a more representative training set, which is crucial in medical contexts when disease instances are less common.

Our research showed excellent results when we combined the potent XGBoost algorithm with ADASYN. The efficieny was prominently improved. Our Parkinson's disease prediction model was able to be built on a solid foundation.

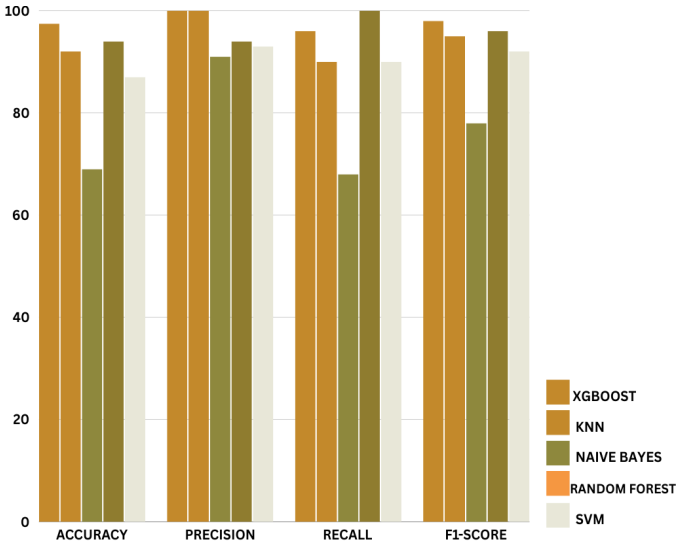

Fig. 6. Comparison over other ML Algorithms

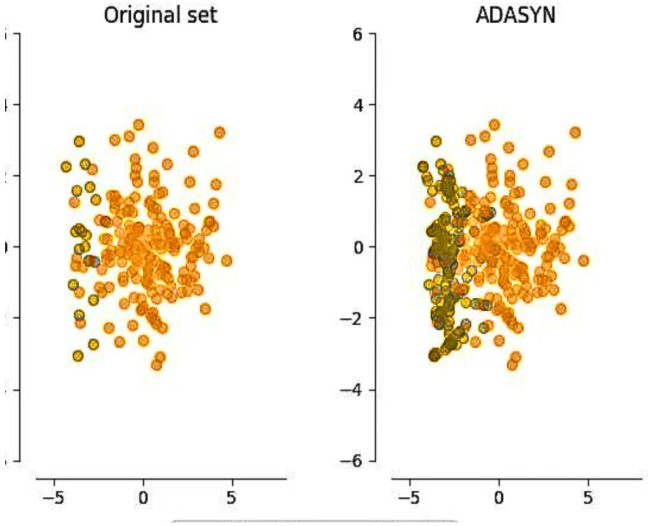

Fig. 7. Orginal dataset vs ADASYN

Considering the inherent imbalance between positive and negative classes, our selection of evaluation metrics—precision, recall, F1-score, AUC-ROC, and specificity—introduced a nuanced dimension to our assessment.

REFERENCES

[1] R. Prashanth, Sumantra Dutta Roy, Early detection of Parkinson's disease through patient questionnaire and predictive modelling, International Journal of Medical Informatics, Volume 119,2018,Pages 75-87,ISSN 1386-5056.

[2] Naranjo L, Pérez CJ, Martín J, Campos-Roca Y. A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications. Comput Methods Programs Biomed. 2017 Apr;142:147-156. doi: 10.1016/j.cmpb.2017.02.019. Epub 2017 Feb 22. PMID: 28325442.

[3] Zhang-Li Wang, Lin Yuan, Wen Li, Jia-Yi Li, Ferroptosis in Parkinson's disease: glia–neuron crosstalk, Trends in Molecular Medicine,Volume 28, Issue 4,2022,Pages 258-269,ISSN 1471-4914.

[4] Imran Ahmed, Sultan Aljahdali, Muhammad Shakeel Khan and Sanaa Kaddoura-Classification of Parkinson Disease Based on Patient's Voice Signal Using Machine Learning. Signal Process. Control, vol. 26, pp. 80–89, Apr. 2016.

[5] Illner, V., Sovka, P., Rusz, J. (2020). Validation of freely-available pitch detection algorithms across various noise levels in assessing speech captured by smartphone in Parkinson's disease. Biomed. Signal Process. Control., 58, 101831.

[6] G. Solana-Lavalle, J.-C. Galán-Hernández, and R. Rosas-Romero, "Automatic Parkinson disease detection at early stages as a pre-diagnosis tool by using classifiers and a small set of vocal features," Biocybern. Biomed. Eng., vol. 40, no. 1, pp. 505–516, Jan. 2020.

[7] I. El Maachi, G.-A. Bilodeau, and W. Bouachir, "Deep 1D-convnet for accurate Parkinson disease detection and severity prediction from gait," Expert Syst. Appl., vol. 143, Apr. 2020, Art. no. 113075

[8] Sunny Kusawa "Feature Scaling Techniques in Data Science: A Comprehensive Guide with Formulas and Python Implementations"

[9] Desai, R. (2019). Top 10 Python Libraries for Data Science. Available online at: https://towardsdatascience.com/top-10-python-libraries-for-data-sciencecd82294ec266 (accessed July 3, 2022)

[10] SAUTHOR: CHARLES DURFEE,SMOTE Oversampling for Imbalanced Classification with Python

[11] Khachnaoui, H.; Khlifa, N.; Mabrouk, R. Machine Learning for Early Parkinson's Disease Identification within SWEDD Group Using Clinical and DaTSCAN SPECT Imaging Features. J. Imaging 2022, 8, 97

[12] Wroge, T.J.; Özkanca, Y.; Demiroglu, C.; Si, D.; Atkins, D.C.; Ghomi, R.H. Parkinson's disease diagnosis using machine learning and voice. In Proceedings of the 2018 IEEE Signal Processing in Medicine and Biology Symposium, IEEE, Philadelphia, PA, USA, 1 December 2018; pp. 1–7

[13] Borzì, L.; Mazzetta, I.; Zampogna, A.; Suppa, A.; Olmo, G.; Irrera, F. Prediction of Freezing of Gait in Parkinson's Disease Using Wearables and Machine Learning. Sensors 2021, 21, 614.

[14] Haibo He,Yang Bai,Edwardo A. Garcia,Shutao Li-ADASYN: Adaptive synthetic sampling approach for imbalanced learning.

[15] Khalid, A.; Senan, E.M.; Al-Wagih, K.; Ali Al-Azzam, M.M.; Alkhraisha, Z.M. Hybrid Techniques of X-ray Analysis to Predict Knee Osteoarthritis Grades Based on Fusion Features of CNN and Handcrafted. Diagnostics 2023, 13, 1609.

[16] Mostafa, S.A.; Mustapha, A.; Mohammed, M.A.; Hamed, R.I.; Arunkumar, N.; Abd Ghani, M.K.; Khaleefah, S.H. Examining multiple feature evaluation and classification methods for improving the diagnosis of Parkinson's disease. Cogn. Syst. Res. 2019, 54, 90–99.

[17] Bind, S., Tiwari, A. K., Sahani, A. K., Koulibaly, P., Nobili, F., Pagani, M., et al. (2015). A survey of machine learning based approaches for parkinson disease prediction. Int. J. Comput. Sci. Inf. Technol. 6, 1648–1655.

[18] Parisi, L.; RaviChandran, N.; Manaog, M.L. Feature-driven machine learning to improve early diagnosis of Parkinson's disease. Expert Syst. Appl. 2018, 110, 182–190.

[19] Harel, B., Cannizzaro, M., and Snyder, P. J. (2004). Variability in fundamental frequency during speech in prodromal and incipient parkinson's disease: a longitudinal case study. Brain Cognit. 56, 24–29. doi: 10.1016/j.bandc.2004.05.002

[20] Little, M. A., McSharry, P. E., Hunter, E. J., Spielman, J., and Ramig, L. O. (2009). Suitability of dysphonia measurements for telemonitoring of parkinson's disease. IEEE Tran. Biomed. Eng. 56, 1015–1022. doi: 10.1109/TBME.2008.2005954

[21] Kadam, V. J., and Jadhav, S. M. (2019). "Feature ensemble learning based on sparse autoencoders for diagnosis of parkinson's disease," in Computing, Communication and Signal Processing. Advances in Intelligent Systems and Computing, Vol. 810, eds B. Iyer, S. Nalbalwar, N. Pathak (Singapore: Springer), 567–581. doi: 10.1007/978-981-13-1513-8-58

[22] Gupta, I., Sharma, V., Kaur, S., Singh, A. K. (2022). "PCA-RF: An Efficient Parkinson's Disease Prediction Model based on Random Forest Classification". arXiv preprint arXiv:2203.11287.

[23] Y. Guan (2021), "Application of logistic regression algorithm in the diagnosis of expression disorder in Parkinson's disease," 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), 2021, pp. 1117-1120, doi: 10.1109/ICIBA52610.2021.9688135