International Journal of Innovation Studies

# DETECTION OF PARKINSON DISEASE USING MACHINE LEARNING

**Pinal D. Salot[a], Prof. K. H. Wandra[b], Prachi Pancholi[a]**

[a] Assistant Professor, L. D. College of Engineering, Ahmedabad, 380015

[b] Pro Vice-Chancellor, Rashtriya Raksha Shakti University, Gujarat, 380016

**Abstract**

*The symptoms of Parkinson's disease (PD), the second most prevalent neurological disorder in the elderly, include a wide variety of impairments in motor control and cognitive development. Following a stroke, this is the most common neurological disorder. Because the symptoms are comparable to those of other diseases, Parkinson's disease (PD) may be hard to diagnose. This category includes disorders such as essential tremor and aging. About the time you hit fifty-five, you'll start to notice symptoms like trouble walking and speaking more often. Medication is available to alleviate some of the symptoms of Parkinson's disease (PD), but there is yet no cure. Assuming they can manage their symptoms, everyone can carry on with their regular lives. Recognizing this disease and intervening to prevent its progression is of the utmost importance. Attempts to determine the disease kind have necessitated extensive investigation. We are primarily focused on creating and applying various deep learning and machine learning models for the aim of diagnosing Parkinson's disease (PD). The Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), Knowledge Network (KNN), and Multi-Layer Perceptron (MLP) are just a few examples of these types of models. Numerous models are at your disposal, not limited to KNN and MLP. The objective is to use the characteristics of the speech signals to differentiate between healthy individuals and those with Parkinson's disease (PD). The dataset, which included 150 audio recordings of exams given to 31 people, was retrieved from the UC Irvine machine learning repository. More than that, we improved our models' performance through training with features selection, hyperparameter tuning, and the Synthetic Minority Over-sampling Technique (SMOTE) (GridSearchCV). The most effective combination of GridSearchCV, SMOTE, MLP, and SVM, with a train/test split ratio of 70:30, yielded the greatest results for our project. To review, MLP attained a f1-score of 99%, an accuracy of 98.31%, a recall of 98%, and precision of 100%. Along with a remarkable 95% identification rate, 98% accuracy, 96% recall, and 97% f1-score, the support vector machine (SVM) also achieved 98% accuracy. It appears that the proposed method might be helpful for PD prediction and could be readily integrated into healthcare systems for diagnosis, according to the clinical trials conducted for this work.*

**Keywords**-: Parkinson Diseases, Machine Learning, Predication, Detection.

## 1. INTRODUCTION

Among the several types of neurodegeneration, PD is among the most prevalent, say Tysnes and Storstein (2017). A prevalence rate of 1% in those aged 60 and over is linked to an effect on 1-2 people per thousand. Dorsey et al. (2018) reported that the estimated number of people living with Parkinson's disease increased from 2.5 million in 1990 to 6.1 million in 2016. The standardization of age-based prevalence rate computations and the ever-increasing fraction of the population aged 65 and up are the reasons for this rise. Degenerative neurological condition Parkinson's disease (PD) can manifest with both motor and non-motor symptoms (Jankovic, 2008). All three stages of a motion are encompassed in it (Contreras-Vidal & Stelmach, 1995).

Immobility, rigidity, and tremors can be observed at any stage of the illness progression, even prior to the onset of cognitive and behavioral problems such as dementia (Opara et al., 2012). These symptoms could become more prominent as the illness progresses. Life quality, interpersonal dynamics, familial bonds, and community functions are all profoundly affected by Parkinson's disease (PD), which also has substantial societal and personal financial implications (Yang and Chen, 2017; Johnson et al., 2013; Kowal et al., 2013).

For a long time, people believed that difficulties with movement were the main sign of Parkinson's disease (PD). The bulk of techniques used to rate the severity of diseases have not been validated or researched enough (Jankovic, 2008). The conventional signs of Parkinson's disease (PD) may still not be present, however this does not change the fact that this is the case. In Parkinson's disease, non-motor symptoms often appear first, followed by motor ones. Cognitive alterations can manifest in a variety of ways, including but not limited to: irregular sleep schedules, abnormalities in scent perception, problems concentrating or making long-term plans (Jankovic, 2008; Tremblay et al., 2017). These symptoms may differ from patient to patient, are difficult to evaluate, and do not reveal much, as pointed out by Sesiewicz et al. (2006). Although non-motor symptoms have been utilized as a supplementary criterion in certain instances, there are not yet valid ways to diagnose Parkinson's disease (PD) just from them (Braak et al., 2003; Postuma et al., 2015).

Machine learning is being used by more and more healthcare companies. As its name suggests, machine learning is predicated on the idea that computers can learn and, semi-automatically, build meaningful representations from data. Furthermore, machine learning enables the integration of data from many modalities, including SPECT and MRI, to diagnose Parkinson's disease (PD), as demonstrated by Cherubini et al. (2014b) and Wang et al. (2017). Therefore, important yet undervalued features in clinical PD diagnosis can be uncovered by applying machine learning techniques. Using these additional methods, we may be able to identify rare or preclinical Parkinson's disease.

Research on the application of machine learning to the diagnosis of Parkinson's disease has skyrocketed in the past few years. The use of machine learning in the evaluation and diagnosis of Parkinson's disease (PD) has been the subject of several investigations. Sadly, motor symptoms, kinematics, and data obtained from wearable sensors have been the exclusive foci of the studies carried out by Ahlrichs and Lawo (2013), Ramdhani et al. (2018), and Belić et al. (2019). A few of these assessments also exclude publications published in 2015 or 2016, according to Pereira et al. (2019). This research aims to accomplish the following: (a) to gather all the different types of data, their sources, and the outcomes of studies that used ML models to diagnose PD. (b) to compare and evaluate various machine learning approaches according to their efficacy and feasibility in diagnosing Parkinson's disease. (c) for machine learning experts interested in PD diagnosis, to explain the many models and data types used, the results of these studies, and the best standards for publishing experimental techniques and outcomes so that they may be replicated next time. Consequently, machine learning has oftentimes generated quite accurate diagnoses in human patients when trained on data from a variety of non-clinical and clinical methodologies. New biomarkers and machine learning algorithms may soon find greater usage in clinical practice as a result of this. It is possible that these instruments could improve the accuracy and reliability of medical decisions.

## 2. MATERIAL AND METHODS

### 2.1.Dataset

Our deepest appreciation goes out to Little et al. (2007, 2009) for their work in creating the National Center for Voice. Their generosity made it possible for us to get the dataset needed for this study from the UO repository. The dataset is also available at the UCI Machine Learning Repository, as stated by Little (2008). The first research outlined techniques for feature extraction and discussed several voice issues in general. A total of thirty-one individuals had their voices recorded. There were a total of 23 people diagnosed with Parkinson's disease (PD), with 16 men and 7 women affected. The remaining eight people, three of whom were male and five of whom were female, were called Healthy Controls (HC). Among the 195 entries and 24 columns comprising the dataset are biological voice measures. In Table 1 you can see these measurements. Table 1's "name" column displays individual voice recordings, while separate columns show each voice metric. Each patient had an average of six recordings obtained; nine individuals had seven and twenty-two had six. From 46 to 85 years old, the patients' ages varied from an average of 65.8 with a standard deviation of 9.8. From zero to twenty-eight years elapsed between the time of diagnosis and the relevant period in question. A 36-second voiceover is included in every row. The soundproof recording booth was provided by an industrial acoustic company. Eight centimeters away from the mouth was where the microphone was placed (2009). By assigning a value of 0 to the "status" column of the dataset for those without PD and a value of 1 for those with PD, we can distinguish between the two groups.

**Table 1: Utilization of the UCI Dataset in the research**

| | |
|---|---|
| Name | In addition to the subject's ASCII name and recording number, categorical variables are also included.. |
| MDVP: Fo(Hz) | There are a number of quantitative parameters that influence the fundamental frequency of the average male voice.. |
| MDVP: Fhi(Hz) | There are numerical variables that determine the fundamental frequency of the human voice that is the highest. |

| | |
|---|---|
| MDVP: Flo(Hz) | Calculations that determine the minimum fundamental frequency of human voice are based on numerical considerations.. |
| MDVP: RAP | Various numerical variables that quantify the variance in fundamental frequency. |
| NHR | Quantifications of the noise-to-tone ratio in |
| HNR | the voice (Numerical variables). |
| status | 0 for HC and 1 for PD (Numerical variables). |
| RPDE | Mathematical variables for the investigation of nonlinear dynamical complexity. |
| DFA | Signal fractal scaling exponent (Numerical variables). |

### 2.2.Methods

There is a method that has been developed to detect whether or not the individual in question is suffering from Parkinson's disease (PD). This method makes use of the Google Colab environment and the Python programming language. Taking the whole thing into consideration, this model suggests a strategy that is made up of six different procedures. Data preparation, feature selection, SMOTE, GridSearchCV, hyperparameter tuning, machine learning/deep learning classification models, and performance assessment are all components of this approach. Additional components include performance evaluation. The aforementioned procedures are incorporated into this process. It is possible to view a visual depiction of these proposed model stages in Figure 1, which you may consult if you are interested in receiving further information about them.

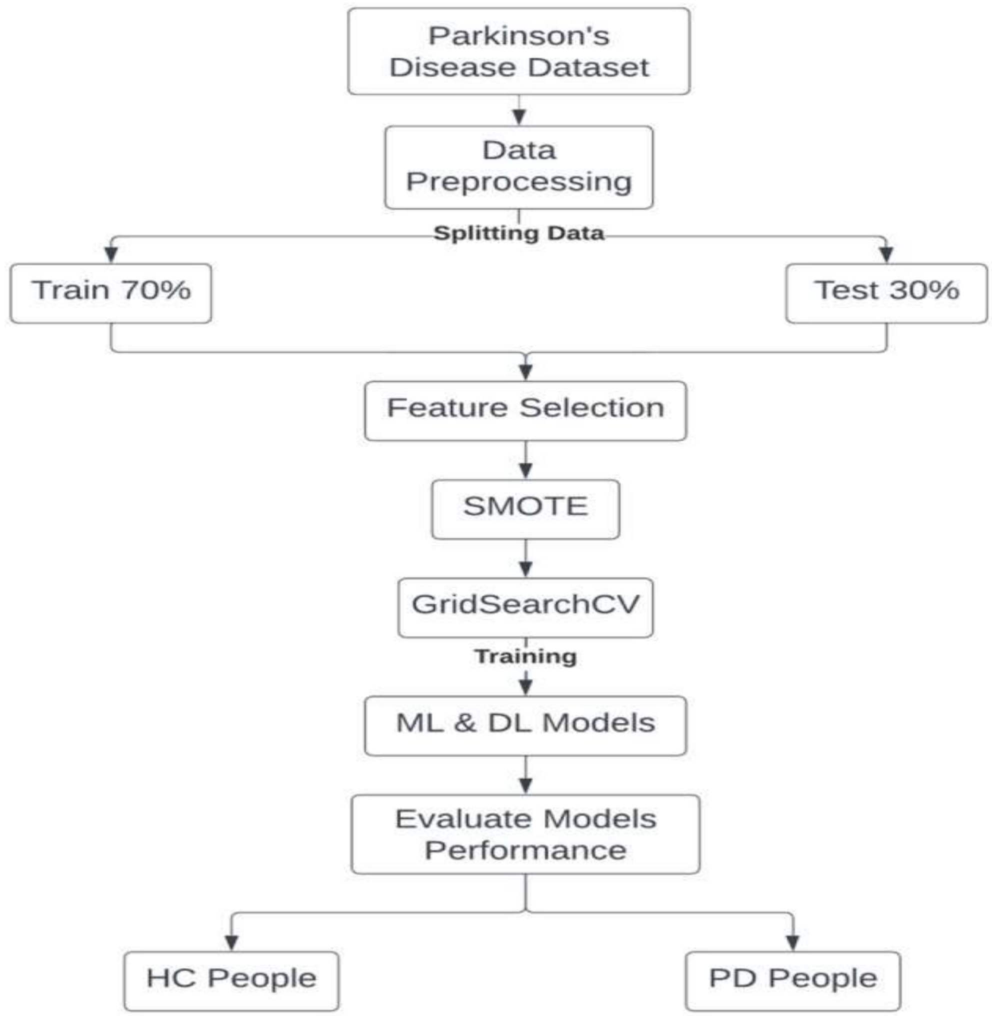



**Figure 1: Steps of the proposed classification models.**

2.2.1 The first stage of processing the data The most crucial part of data processing is pretreatment, says Singh (2020), as it allows the model to learn the input characteristics effectively and gets rid of irrelevant information. Therefore, the most important part of processing data is pretreatment. The dataset was imported into the Google Colab platform using a CSV file. The import of the dataset was made possible by means of the Pandas program. The imbalance in the dataset was detected by looking at the "status" column, which had 147 PD items and 48 HC entries separately. This disparity is around 25% for HC and about 75% for PD. After filtering for duplicates or missing data, we concluded that the dataset was imbalanced. We divided our dataset into a 70:30 train/test ratio to prevent under-fitting and over-fitting. We did this to make sure these two issues never happened. It is possible to transfer the model's training set knowledge to new data sets because its outputs are already known. This is due to the fact that the model has already produced results. It is feasible to scale each feature independently based on its unique properties by computing the appropriate statistics on the samples that make up the training set. After that, we use StandardScaler to get the average and standard deviation. We can then use that information to apply the transform function to future data (Teo, 2021). The mathematical statement that represents the StandardScaler normalization format is Equation (1). We utilized many libraries to achieve the objectives of our investigation. These included, but were not limited to, NumPy, Pandas, Matplotlib, Seaborn, and Sickit-learn (Sklearn). The Numpy module is the backbone of Python's scientific computing package. With the aid of this tool, any type of mathematical operation may be included into the code. The ability to include massive multidimensional arrays and matrices into your code is only one of many advantages of using this tool. The Pandas library is a great tool for data manipulation and analysis, but it is also widely used for importing and organizing datasets. Two Python packages, Matplotlib and Seaborn, lay the groundwork for data visualization in Python. To create many different kinds of two-dimensional graphs, you may use the Matplotlib library in Python. With the help of other libraries like Numpy and Pandas, this may be achieved. With Matplotlib, Pandas, and Numpy, the tool used for graph charting is Seaborn. The last one is Sklearn, the most trustworthy and user-friendly Python machine learning library. It offers a consistent interface built on Python and contains tools for classification, regression, clustering, and dimensionality reduction.

$$Standard\ Scaler = \frac{x_i - mean(x)}{stdev(x)} \quad (1)$$

3. In order to choose the eight characteristics of the dataset that were deemed to be the most significant, the Select K Best algorithm was applied at this step. SelectKBest has been recognized as the second most often used strategy for dimensionality reduction, accounting for 29.1% of the overall utilization, according to the outcomes of the study investigation that was conducted by Bilgen et al. in the year 2020. Using this approach, features are chosen based on the highest k score. This assists in the deletion of data that is less significant and minimizes the amount of time that is required for an individual to undergo training. It was determined to use these eight characteristics: There are several different types of MDVP, including PP, HNR, spred1 spread2, MDVP: Fo(Hz), MDVP: Flo(Hz), MDVP: Shimmer, MDVP:APQ, and HNR.I.e.
4. SMOTE, which stands for Synthetic Minority Over-sampling, is the method that is being discussed here. The number of HC samples in our dataset is lower than the number of PD samples. The practice of oversampling samples from the minority class is one approach that may be taken to remedy uneven classes. The utilization of duplicate instances from the minority class inside the training dataset is one method for accomplishing this goal. It is possible that this will bring about a more even distribution of classes; nonetheless, this does not disclose anything new. The Synthetic Minority Oversampling Technique, often known as SMOTE, is an additional method that may be utilized to improve minority data because it is based on previous samples. A linear link is established by the utilization of close characteristics using the SMOTE approach. Following that, it selects a new sample from the minority class that falls along that class line.

## DATA ANALYSIS

With the use of speech samples, this project aims to construct a reliable PD detection system. Our goals were accomplished by utilizing the UCI dataset, which comprises 195 recordings of speech signal characteristics derived from 48 HC and 147 PD. We assessed the effects of several deep learning and classical machine learning approaches on our models. These included k-nearest neighbors (KNN), decision trees (DT), random forests (RF), and multilayer perceptrons (MLP), among others. By dividing the sample into two separate data sets using the dataset, we were able to demonstrate that the suggested method was effective. About 70% of the samples in the first group made it; this group contains the training samples. The remaining samples will be used for testing, validation, and evaluation in order to ascertain the system's correctness. Our sample contains much more PD than HC, suggesting an imbalance. The Synthetic Minority Over-sampling Technique (SMOTE) was used to build a balanced dataset in order to address this issue. In order to determine the optimal features and hyperparameters for our models, we utilized GridSearchCV and SelectKBest. The following figures display the results both with and without the use of SMOTE, GridSearchCV, and FeatureSelection. We also display further findings. Combining SMOTE with GridSearch yielded the best results, as seen in Figures 2 and 33. The best possible outcomes were attained, as shown by these results. Unfortunately, when we ran FeatureSelection on our dataset, we got disappointing results. The reason behind this is that every single characteristic is crucial to the training process. Because of this, we choose to use every single feature.

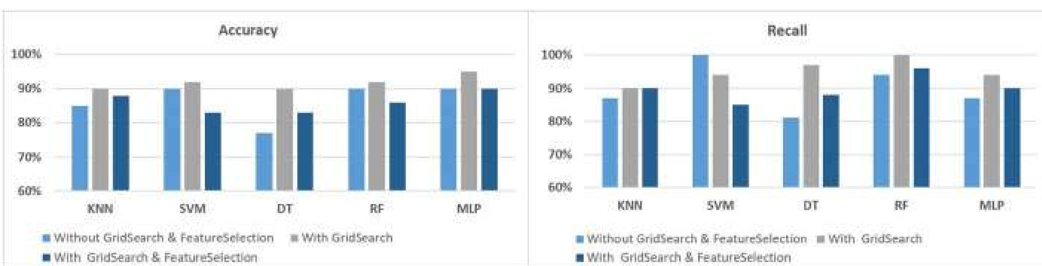



**Figure 2: Without SMOTE, classification results for PD detection based on speech signal properties.**

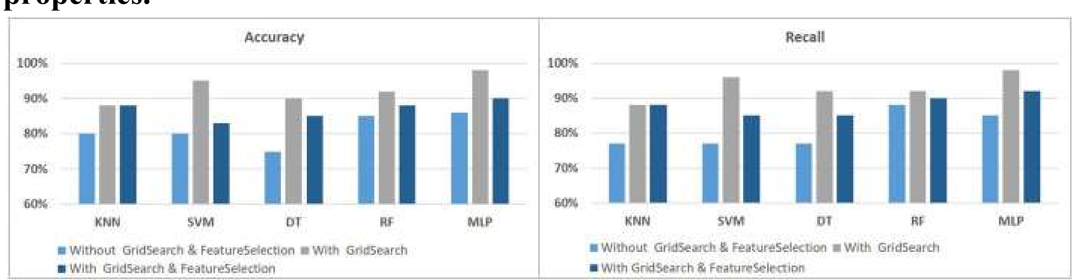



**Figure 3: The classification performance using SMOTE to identify PD using voice signal features.**

According to the findings of this study, reduced voice performance might be the first indicator of motor impairment in Parkinson's disease (Ma et al., 2020). The complexity and accuracy needed for vocalization might cause issues to arise in the limbs here first. Patients with Parkinson's disease have been shown to exhibit audible and visual changes in vocal pitch. Therefore, we know for sure

that voice might be a significant biomarker for PD. While conventional diagnostic markers such as DaT scans and the Unified Parkinson's Disease Rating Scale (UPDRS) clinician-scored supervised movement tests are used for clinical diagnosis, our technique is based entirely on voice data. We hypothesize that a more rapid and precise diagnosis might be achieved by utilizing voice analysis, which is one of the earliest detectable symptoms. Magnetic resonance imaging (MRI) and handwriting analysis are examples of traditional methods that are considered risky. This is in contrast to them. The future of healthcare also lies in voice diagnostics because to its simplicity, low cost, and ease of integration with existing systems. By comparing several models, the authors of this study want to identify the most effective one for Parkinson's disease (PD) categorization.
Table 2 shows the results of the project. A combination of SMOTE and hyperparameter tweaking (GridSearchCV) was employed to attain a precise outcome. Classification accuracy rates of 95% for conventional machine learning (SVM) and 98.31% for deep learning (MLP) were the highest, respectively.

**Table 2: The efficiency of the model.**

| | Without grid search and feature selection | With Grid Search | With Grid Search and feature selection |
|---|---|---|---|
| KNN | Accuracy = 80% | Accuracy = 88% | Accuracy = 88% |
| | Recall = 77% | Recall = 88% | Recall = 88% |
| | Precision = 97% | Precision = 98% | Precision = 98% |
| | F1-Score = 86% | F1-Score = 92% | F1-Score = 92% |
| SVM | Accuracy = 80% | Accuracy = 95% | Accuracy = 83% |
| | Recall = 77% | Recall = 96% | Recall = 85% |

| | | | |
|---|---|---|---|
| | Precision = 97% | Precision = 98% | Precision = 93% |
| | F1-Score = 86% | F1-Score = 97% | F1-Score = 89% |
| DT | Accuracy = 75% | Accuracy = 90% | Accuracy = 85% |
| | Recall = 77% | Recall = 92% | Recall = 85% |
| | Precision = 90% | Precision = 96% | Precision = 95% |
| | F1-Score = 83% | F1-Score = 94% | F1-Score = 90% |
| RF | Accuracy = 85% | Accuracy = 92% | Accuracy = 88% |
| | Recall = 88% | Recall = 92% | Recall = 90% |
| | Precision = 93% | Precision = 98% | Precision = 96% |
| | F1-Score = 90% | F1-Score = 95% | F1-Score = 92% |
| MLP | Accuracy = 86% | Accuracy = 98% | Accuracy = 90% |
| | Recall = 85% | Recall = 98% | Recall = 92% |
| | Precision = 98% | Precision = 100% | Precision = 96% |
| | F1-Score = 91% | F1-Score = 99% | F1-Score = 94% |

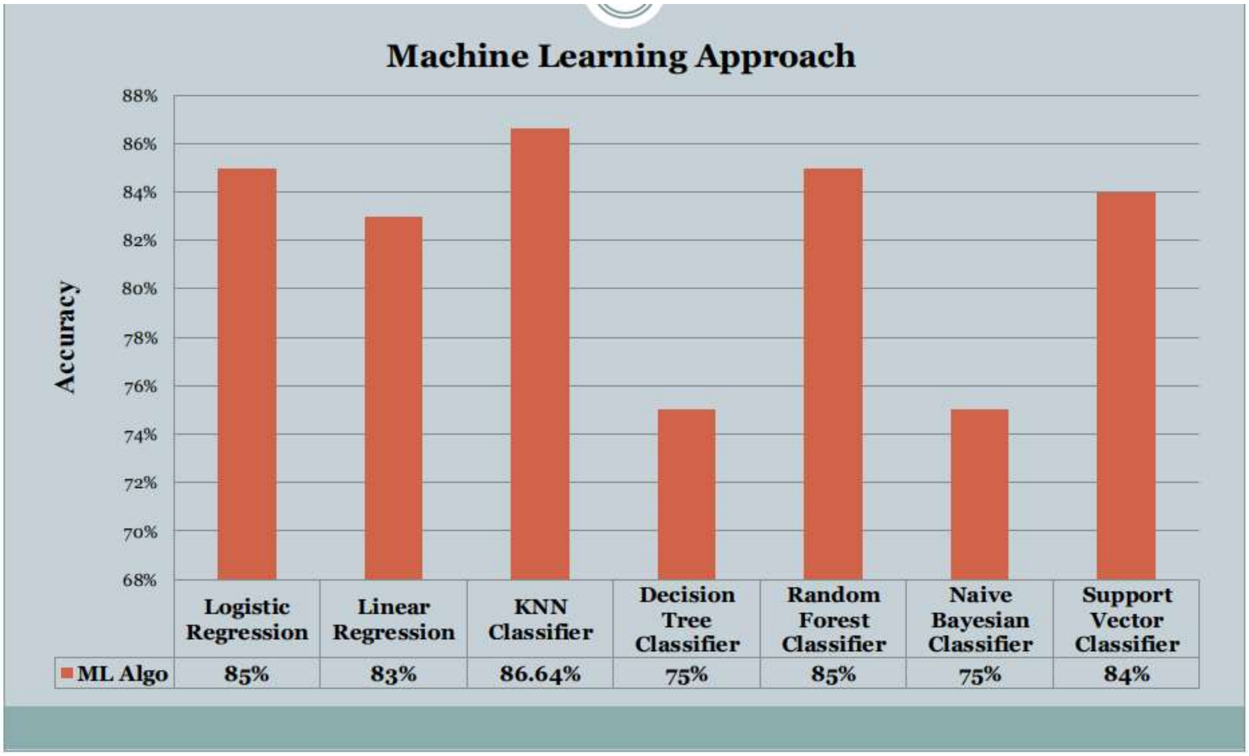
Many scientists analyzed the information using statistical tools like regression and classification analysis to learn more about Parkinson's disease. by comparing and contrasting this study's results with those of others that utilized the identical dataset. In terms of efficiency and precision, this inquiry produces the best possible outcomes. Three separate feature selection approaches were employed by Senturk (2020): CART utilized seven features, SVM and ANN thirteen features, etc. Not even a 93.84% can be considered adequate for that particular paper. It is interesting that Das (2010) utilized four distinct methods for data categorization; this is true regardless of whether the algorithms' performance was below 92.9%. Nevertheless, in Table 3, researchers discussed how to establish dataset class balance using SMOTE. Not a single other outlet did. Reducing the size of the minority class to match that of the majority class and eliminating many outliers from both sets of data allowed them to boost the classifiers' accuracy, they asserted. Given that the MLP method attained the highest accuracy of 98.31%, it can be inferred that the use of hyperparameter tweaking (GridSearchCV) and SMOTE processes was the most significant component in our investigation's conclusions. The overall method of designing the MLP architecture is outlined by the four phases below:

1) The input vector, which is directly associated to Layer 1 (the input layer), contains all of the parameter fields that are contained inside the patient's record.

2) Level 2 (the layer that is concealed) Determining the number of neurons that are concealed in this layer is the most challenging aspect of the process of developing the network. Through the utilization of hyper parameter tweaking, we were able to cut down on the amount of time spent

experimenting with the appropriate number of neurons, network activation function, and algorithm to improve the network weights (solver). The best value was found to be one hidden layer with sixteen neurons, a relu activation function, and a lbfgs solver. This was discovered through the utilization of this approach.

3) The prediction layer, Layer 3 (output layer), determines whether the patient outcome is HC or PD. We used the classification report to assess the model's accuracy.




## 5. CONCLUSION

We are aware of no other study that has compiled the findings of all the research that has utilized machine learning algorithms to diagnose Parkinson's disease; this particular study is the first one that we are aware of. This part of the article contains a summary of the study that was included, as well as links to the conclusions that they obtained on the following topics: (a) Machine learning techniques that are utilized in the diagnosis of Parkinson's disease and its consequent effects; (b) biometric, clinical, and behavioral data that have the potential to enhance the precision of diagnoses; (c) potential biomarkers that can assist in the process of clinical decision making; and (d) additional relevant data, such as databases, that can be utilized to supplement and broaden the scope of existing datasets. There is a possibility that the process of identifying Parkinson's disease might be sped up by the creation of novel biomarkers. Last but not least, the utilization of machine learning-assisted Parkinson's disease diagnosis has the potential to bring about a more structured framework for clinical decision-making. Since this is the case, the utilization of machine learning by medical experts may result in an increase in the resources available for screening, detecting, or diagnosing Parkinson's disease.

## REFERENCES

1. Divya, J., Radhakrishnan, P., Pavithra, G., Gopatoti, A., Baburao, D., & Krishnamoorthy, R. (2023, April). Detection of Parkinson Disease using Machine Learning. In 2023 International Conference on Inventive Computation Technologies (ICICT) (pp. 53-57). IEEE.
2. Isenkul, M.E.; Sakar, B.E.; Kursun, O. . 'Improved spiral test using digitized graphics tablet for monitoring Parkinson's disease.' The 2nd International Conference on e-Health and Telemedicine (ICEHTM-2014), pp. 171-175, 2014.
3. Mani, I., and Zhang, I. (2003). “KNN approach to unbalanced data distributions: a case study involving information extraction,” in Proceedings of Workshop on Learning From Imbalanced Datasets, Vol. 126 (ICML), 1-7.
4. Parziale, A., Della Cioppa, A., Senatore, R., and Marcelli, A. (2019). “A decision tree for automatic diagnosis of Parkinson's disease from offline drawing samples: experiments and findings,” in International Conference on Image Analysis and Processing (Cham: Springer), 196– 206.
5. Polat, K. (2019). “A hybrid approach to Parkinson disease classification using speech signal: the combination of smote and random forests,” in 2019 Scientific Meeting on Electrical-Electronics and Biomedical Engineering and Computer Science (EBBT) (Istanbul:IEEE), 1–3.
6. Wang, W., Lee, J., Harrou, F., & Sun, Y. (2020). Early detection of Parkinson’s disease using deep learning and machine learning. IEEE Access, 8, 147635-147646.
7. Senturk, Z. K. (2020). Early diagnosis of Parkinson’s disease using machine learning algorithms. Medical hypotheses, 138, 109603.
8. Sriram, T. V., Rao, M. V., Narayana, G. S., Kaladhar, D. S. V. G. K., & Vital, T. P. R. (2013). Intelligent Parkinson disease prediction using machine learning algorithms. Int. J. Eng. Innov. Technol, 3(3), 1568-1572.
9. Govindu, A., & Palwe, S. (2023). Early detection of Parkinson's disease using machine learning. Procedia Computer Science, 218, 249-261.
10. Pahuja, G., & Nagabhushan, T. N. (2021). A comparative study of existing machine learning approaches for Parkinson's disease detection. IETE Journal of Research, 67(1), 4-14.
11. Challa, K. N. R., Pagolu, V. S., Panda, G., & Majhi, B. (2016, October). An improved approach for prediction of Parkinson's disease using machine learning techniques. In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1446-1451). IEEE.
12. Prashanth, R., Roy, S. D., Mandal, P. K., & Ghosh, S. (2016). High-accuracy detection of early Parkinson's disease through multimodal features and machine learning. International journal of medical informatics, 90, 13-21.
13. Almeida, J. S., Rebouças Filho, P. P., Carneiro, T., Wei, W., Damaševičius, R., Maskeliūnas, R., & de Albuquerque, V. H. C. (2019). Detecting Parkinson’s disease with

sustained phonation and speech signals using machine learning techniques. Pattern Recognition Letters, 125, 55-62.
14. Sahu, L., Sharma, R., Sahu, I., Das, M., Sahu, B., & Kumar, R. (2022). Efficient detection of Parkinson's disease using deep learning techniques over medical data. Expert Systems, 39(3), e12787.
15. Abdulhay, E., Arunkumar, N., Narasimhan, K., Vellaiappan, E., & Venkatraman, V. (2018). Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease. Future Generation Computer Systems, 83, 366-373.
16. Raval, S., Balar, R., & Patel, V. (2020, June). A Comparative Study of Early Detection of Parkinson's Disease using Machine Learning Techniques. In 2020 4th international conference on trends in electronics and informatics (ICOEI)(48184) (pp. 509-516). IEEE.
17. Younis Thanoun, M., & T. YASEEN, M. O. H. A. M. M. A. D. (2020, October). A comparative study of Parkinson disease diagnosis in machine learning. In Proceedings of the 4th International Conference on Advances in Artificial Intelligence (pp. 23-28).
18. Mittal, V., & Sharma, R. K. (2021). Machine learning approach for classification of Parkinson disease using acoustic features. Journal of Reliable Intelligent Environments, 7(3), 233-239.
19. Johri, A., & Tripathi, A. (2019, August). Parkinson disease detection using deep neural networks. In 2019 Twelfth international conference on contemporary computing (IC3) (pp. 1-4). IEEE.