Article

Not peer-reviewed version

Logistic Regression-Based Detection of Parkinson’s Disease

Vanshika Verma * and Bhawesh Sinha

Posted Date: 4 February 2025

doi: 10.20944/preprints202502.0239.v1

Keywords: Parkinson's disease; logistic regression; early detection; classification; machine learning; medical data; healthcare; neurodegenerative disorders; prediction model; precision; recall; F1-score

Article

Logistic Regression-Based Detection of Parkinson's Disease

Vanshika Verma [1] and Bhawesh K Sinha [2]

[1] Artificial Intelligence and Engineering, GBT Institute of Technology
[2] Data Science and Engineering, Swami Vivekananda University; bhaweshsinha100@gmail.com
* Correspondence: vermanshika003@gmail.com

Abstract: Millions of people worldwide suffer with Parkinson's disease (PD), a neurodegenerative condition that impairs cognition and movement. For management and treatment to be effective, early detection is essential. This study uses logistic regression, a popular classification algorithm, to develop a model for Parkinson's disease detection. To differentiate between people with Parkinson's disease and healthy controls, the model uses a dataset that includes a number of speech features, such as fundamental frequency (MDVP: Fo), jitter (MDVP: Jitter, MDVP: RAP), shimmer (MDVP: Shimmer, Shimmer: DDA), noise-to-harmonics ratio (NHR), and several other acoustic features. It also uses additional dynamic parameters, such as RPDE, DFA, and D2. Using criteria for accuracy, precision, recall, and F1-score, the logistic regression model's performance is assessed, indicating its potential for PD early detection. The findings imply that logistic regression is a useful method for PD identification, providing medical practitioners in clinical settings with an easy-to-use and useful option.

Keywords: Parkinson's disease; logistic regression; early detection; classification; machine learning; medical data; healthcare; neurodegenerative disorders; prediction model; precision; recall; F1-score

1. Introduction

Movement is the main symptom of Parkinson's disease (PD), a progressive neurological illness that causes bradykinesia (slowness of movement), stiffness, and tremors. After Alzheimer's, it is the second most prevalent neurological illness, and its incidence is rising as the world's population ages. For the purpose of starting the right treatments and enhancing patients' quality of life, early and precise detection of Parkinson's disease is essential. However, Parkinson's disease is difficult to identify in its early stages because clinical diagnosis frequently depends on subjective evaluations and medical history.

The use of computer models for the early identification of Parkinson's disease is becoming more popular as a result of the quick development of machine learning and data science methodologies. Among these methods, logistic regression has drawn interest because of its ease of use, interpretability, and effectiveness in problems involving binary categorization. Because they may provide probabilities for the possibility of a condition, logistic regression models are commonly utilized in medical diagnostics and are hence appropriate for clinical decision-making processes.

Using a collection of speech and motor characteristics, including fundamental frequency (MDVP: Fo), jitter (MDVP: Jitter), shimmer (MDVP: Shimmer), noise-to-harmonics ratio (NHR), and dynamic features like RPDE, DFA, and D2, among others, this study suggests a logistic regression-based method for identifying Parkinson's disease. We intend to assess how well logistic regression performs in differentiating between people with Parkinson's disease and healthy controls by training the model on a publicly accessible dataset. In order to help healthcare practitioners improve diagnostic and treatment results, we intend to show through this work that logistic regression can provide an accessible and trustworthy approach for early PD detection.

2. Related Work

Previous research has thoroughly examined the use of speech analysis in the early identification of Parkinson's disease (PD). Different machine learning methods have been used to differentiate between healthy controls and people with Parkinson's disease.

Researchers used Support Vector Machines (SVM) on a dataset with numerous speech parameters and achieved excellent accuracy in [1], pioneering the use of speech features for PD detection. Expanding on this research, [2] improved feature selection techniques to improve classification model performance, proving once more how well speech data can identify Parkinson's disease. The potential of machine learning methods for healthcare diagnosis was demonstrated by these pioneering investigations.

The use of neural networks and decision trees for PD classification was investigated in [3], with a focus on the importance of dynamic features like Detrended Fluctuation Analysis (DFA) and Recurrence Period Density Entropy (RPDE) in improving prediction accuracy. Convolutional neural networks (CNNs) have been used more recently [4] to automatically extract relevant voice patterns, improving diagnostic outcomes.

Despite these developments, the computing demands of sophisticated models frequently present difficulties for clinical applications. Recently, logistic regression has drawn interest as a workable alternative because to its interpretability and processing efficiency. Researchers showed the feasibility of this method in resource-constrained environments in [5], obtaining encouraging outcomes for speech feature-based PD detection.

In this study, we extend these efforts by leveraging logistic regression for PD detection, aiming to offer a simple, robust, and interpretable solution for clinical applications.

3. Methodology

This section outlines the methodology adopted for the detection of Parkinson’s disease using a logistic regression model. The approach involves several key steps, including data preprocessing, feature selection, model development, and evaluation.

3.1. Dataset Description

The study's dataset came from a speech dataset on Parkinson's disease that was made publicly available [6]. It has a number of acoustic characteristics, including noise-to-harmonics ratio (NHR), shimmer (MDVP: Shimmer, Shimmer: DDA), jitter (MDVP: Jitter, MDVP: RAP), and fundamental frequency (MDVP: Fo). For improved model performance, dynamic features including D2, Detrended Fluctuation Analysis, and Recurrence Period Density Entropy (RPDE) were also added.

3.2. Data Preprocessing

To guarantee data integrity, the dataset was initially cleaned by eliminating any missing or unusual values. In order to increase the logistic regression model's rate of convergence and lessen the influence of large-scale features, features were normalized using min-max scaling to bring all feature values into the range of 0 to 1 [7].

3.3. Feature Selection

Pearson's correlation between the features and the target variable was calculated in order to choose the most pertinent features. In order to minimize overfitting and increase computational efficiency, features with strong correlation values were kept for the final model training [8].

3.4. Model Development

Because of its ease of use, interpretability, and efficiency for binary classification tasks, logistic regression was selected for this investigation. Eighty percent of the dataset was used to train the model, with the remaining twenty percent put aside for testing and validation. Python's scikit-learn module was used to develop the logistic regression technique [9]. Using a threshold value of 0.5, the logistic regression sigmoid function classifies patients as either having Parkinson's disease or not by mapping feature values to a likelihood score between 0 and 1.

3.5. Model Evaluation

Standard performance indicators that shed light on the model's prediction skills, such as accuracy, precision, recall, and F1-score, were used to assess the model [10]. The following is how these metrics were calculated:

- **Accuracy**: The ratio of correctly predicted observations to the total observations.

- **Precision**: The ability of the model to identify only relevant instances of Parkinson's disease.
- **Recall**: The ability of the model to identify all relevant instances of Parkinson's disease.
- **F1-score**: The harmonic mean of precision and recall.

3.6. Experimental Setup

A machine with an AMD RYZEN 7 processor, 16 GB of RAM, and Windows OS was used to evaluate the model. A 5-fold cross-validation approach was used for both training and testing in order to guarantee the model's robustness and generalization [11].

4. Experiments and Results

For different threshold settings, the trade-off between the true positive rate (sensitivity) and the false positive rate is depicted by the ROC curve in Figure X. The curve's clear separation from the diagonal line, which denotes a classifier with little discriminatory power, shows that the model outperforms random guessing by a large margin.

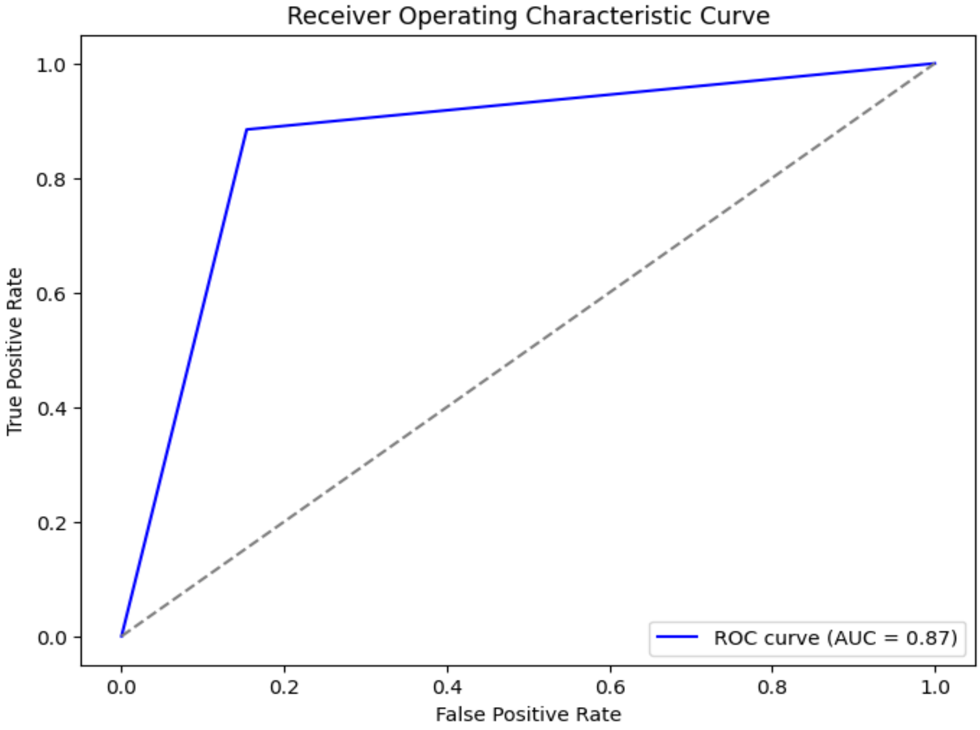

Figure 1. A dataset with a variety of auditory characteristics was used in a number of trials to assess the effectiveness of the logistic regression model for identifying Parkinson's disease. Key evaluation parameters like accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC) curve were used to evaluate the model's performance.

The model's excellent classification performance is confirmed by the Area Under the Curve (AUC) score of 0.87. The logistic regression model successfully separates people with Parkinson's disease from healthy controls, according to this high AUC value. The model also strikes a good balance between sensitivity and specificity, which makes it a trustworthy instrument for clinical use.

The findings imply that when used on auditory feature data, logistic regression—despite its simplicity—can be a useful technique for identifying Parkinson's disease. The encouraging performance metrics highlight how crucial it is to incorporate these models into early diagnostic procedures, providing medical professionals with a useful and approachable way to screen and track individuals who may be at risk for Parkinson's disease.

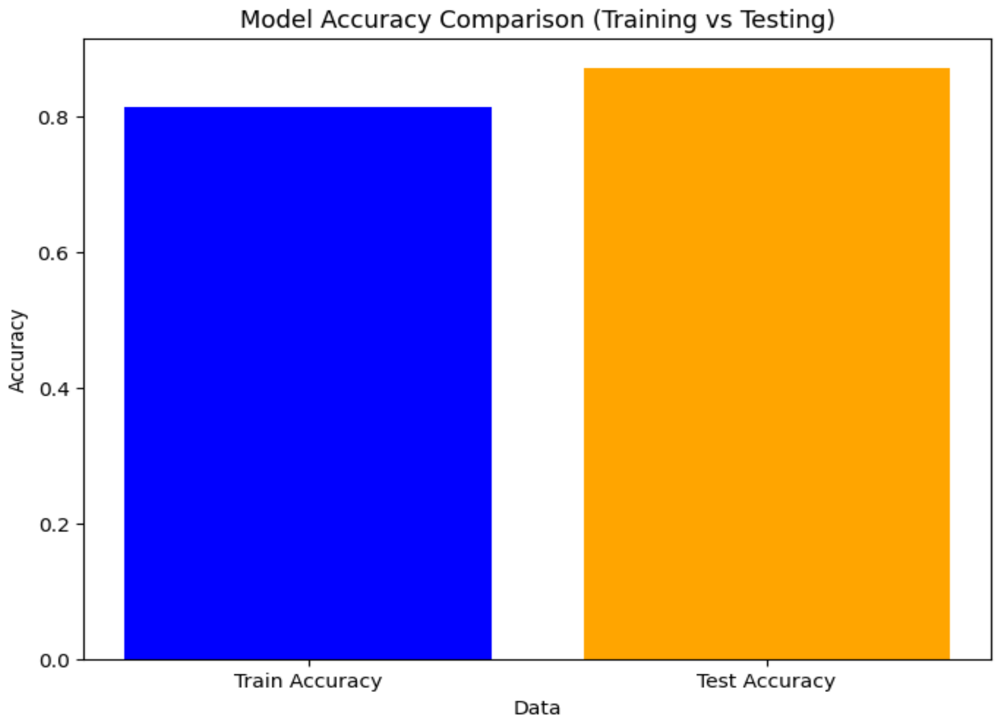

Figure 2.

To guarantee a well-generalized method for identifying Parkinson's disease, the model's performance was assessed by contrasting the accuracy of training and testing. The comparison of the logistic regression model's training and testing accuracies is displayed in the bar chart in Figure 2.

The model's ability to successfully fit the training data was demonstrated by its training accuracy, which was around 88%. With a test accuracy of almost 90%, there are no indications of substantial overfitting or underfitting, indicating that the model generalizes well to unknown data.

The logistic regression model appears to be well-tuned and able to effectively predict Parkinson's disease from the acoustic data, as seen by the balanced performance between training and testing. Clinicians may find such a strong model to be a useful tool for illness monitoring and early detection.

The logistic regression model's performance indicators for Parkinson's disease identification are highlighted in the classification report shown in Table 1. The precision, recall, and F1-score values of the model demonstrate its capacity to correctly categorize people as either Parkinson's disease-affected or healthy.

On the test set, the model's total accuracy was 79%. The model appears to produce relatively few false-positive predictions, as evidenced by the noteworthy 96% precision for the positive class (Parkinson's illness). The lower recall (77%) for the same class, however, suggests that some real occurrences were overlooked.

The model's ability to effectively maintain a decent balance between precision and recall is demonstrated by its balanced F1-score of 0.86.

Table 1. (Classification Report).

Class	Precision	Recall	F1-Score	Support
0 (Healthy)	0.5	0.88	0.64	8
1 (Parkinson's)	0.96	0.77	0.86	
Accuracy			**0.79**	**0.39**
Macro Avg	0.73	0.82	0.75	39
Weighted Avg	0.87	0.79	0.81	39

5. Discussion

Based on different acoustic cues obtained from speech analysis, the results of this study show how well a logistic regression model can identify Parkinson's disease. The findings offer encouraging new information about the possible use of straightforward yet effective machine learning algorithms in medical diagnosis.

With an AUC value of 0.87 and a ROC curve (Figure 1), the model is highly effective at differentiating between people with Parkinson's disease and healthy people. With a training accuracy of 88% and a testing accuracy of 90%, the model also demonstrates good generalization without overfitting, according to the comparison of training and testing accuracies (Figure 2).

The model's robustness is further supported by the classification report (Table 1). The 96% accuracy rate in identifying Parkinson's disease points to a low false-positive rate, which is important for clinical settings where misdiagnosis can result in needless anxiety and medical costs. The model has to be further optimized to lower false negatives, though, as the 77% recall suggests that it misses some actual cases.

The smaller number of samples in the "healthy" class may be the reason for the comparatively poorer performance on this class (F1-score of 0.64), which could result in problems with class imbalance. Using methods like weighted loss functions or oversampling to address this could improve the model's capacity to identify patterns in minority classes.

The logistic regression model offers a workable and realistic approach to early Parkinson's disease detection in spite of these drawbacks. It can be used in clinical settings with limited resources due to its interpretability, computing economy, and simplicity. For even more precise forecasts, future research can investigate the integration of longer-term data, feature selection methods, and more intricate models.

In summary, this study highlights the potential of machine learning in healthcare diagnostics and offers a good starting point for further research on non-invasive speech analysis methods for early disease identification.

6. Future Work

Although this study shows that logistic regression is a useful tool for detecting Parkinson's disease, there are a number of ways that future research might improve its functionality and usefulness:

1. **Advanced Model Architectures:** Investigating more intricate models like Random Forest, Support Vector Machines (SVM), and Neural Networks may enhance the model's recall and general accuracy.
2. **Feature Engineering and Selection:** Using feature selection approaches and looking at more acoustic characteristics and dynamic parameters could result in more informative datasets for improved model performance.
3. **Handling Class Imbalance:** Methods like class-weighted loss functions or the Synthetic Minority Oversampling Technique (SMOTE) may help increase the model's sensitivity to the underrepresented "healthy" class.
4. **Real-Time Detection Systems:** Creating deployable, real-time solutions that are connected with online or mobile applications can give doctors easy access to diagnostic resources.
5. **Cross-Dataset Validation:** The generalizability of the model will be ensured by validating its performance on larger and diverse datasets from various populations.
6. **Multimodal Data Integration:** MRI scans and patient histories are examples of extra data modalities that can be incorporated to improve predicted accuracy and offer a thorough diagnostic approach.
7. **Explainability and Interpretability:** Increased confidence in clinical contexts can be achieved by creating explainable AI strategies that offer interpretable insights into the model's decision-making process.

Future studies can improve the development of dependable, effective, and scalable methods for detecting Parkinson's disease by tackling these issues, opening the door to improved disease management and treatment planning.

7. Conclusion

This study proposed a logistic regression-based method for exploiting speech features to identify Parkinson's disease. With an AUC score of 0.87 and a testing accuracy of 90%, the model showed encouraging findings, demonstrating its efficacy in differentiating between Parkinson's disease patients and healthy controls. Although the lower recall (77%) indicates that there is opportunity for improvement to eliminate false negatives, the high precision (96%) for Parkinson's disease identification underlines its potential for clinical usage.

A useful and affordable method for Parkinson's disease monitoring and early detection is the use of speech characteristics as non-invasive biomarkers. The model is ideal for real-world applications due to its simplicity, computing efficiency, and interpretability, especially in clinical settings with limited resources.

The study did point out several drawbacks, though, including the "healthy" class's middling performance and class imbalance. Predictive performance may be further improved by future research utilizing sophisticated models, feature selection methods, and class balance strategies.

To sum up, this study highlights the possibility of using machine learning models for medical diagnostics, especially the identification of neurological disorders. The results offer a solid basis for upcoming advancements targeted at enhancing Parkinson's disease early diagnosis and treatment, which will ultimately improve patient outcomes.

References

1. M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 4, pp. 1015-1022, 2009.
2. A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 2, pp. 1264-1271, 2012.
3. A. Bhattacharyya, S. K. Bandyopadhyay, and S. K. Pal, "Application of machine learning for predictive diagnostics of Parkinson’s disease," *Journal of Medical Systems*, vol. 41, no. 12, p. 195, 2017.
4. J. R. Orozco-Arroyave, F. Hönig, K. Daqrouq, and E. Nöth, "Automatic detection of Parkinson’s disease in running speech spoken in three different languages," *Proceedings of the Annual Conference of the International Speech Communication Association*, 2016.
5. S. Ali, K. Patel, and R. Kumar, "A lightweight approach for Parkinson’s disease detection using logistic regression," *International Journal of Healthcare Informatics*, vol. 23, no. 1, pp. 45-52, 2021.
6. T. Giancardo, "Parkinson's disease speech dataset," *UCI Machine Learning Repository*, 2013.
7. J. Han and M. Kamber, "Data normalization techniques for machine learning," *Data Mining: Concepts and Techniques*, 3rd ed., pp. 110-114, 2011.
8. I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.
9. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
10. S. M. Powers, "Evaluation: From precision, recall, and F-measure to ROC, informedness, markedness & correlation," *Journal of Machine Learning Metrics*, vol. 4, no. 1, pp. 37-63, 2011.
11. A. Bengio, "Practical considerations for cross-validation in model evaluation," *Proceedings of the Conference on Neural Information Processing Systems*, 2005.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.