Parkinson's Disease Diagnosis Based on XGBoost Algorithm

Shilong Yao[1]

[1] College of Electronic Information Engineering, Zhuhai College of Science and Technology, Zhuhai, Guangdong, 519040, China
1345827235@stu.zcst.edu.cn

Abstract. In light of China's aging population and improved living standards, there is an increasing focus on health issues. Parkinson's disease is a commonly-seen neurodegenerative disorder that greatly impacts patients' quality of life, and its prevalence is on the rise. The traditional approach to Parkinson's disease detection relies on subjective symptom assessment by physicians using a standardized rating scale. This approach is susceptible to high misdiagnosis rates, and it is time and labor-intensive. Currently existing Parkinson's prediction systems pose problems of complicated operation and suboptimal algorithms, which hinders the advancement of experiments. In light of these challenges, this study seeks to improve and enhance commonly-used algorithms to enable more accurate diagnosis of Parkinson's disease. To facilitate the automated diagnosis and symptom prediction in patients, this study utilizes a Parkinson's disease voice prediction model that is based on speech analysis and machine learning algorithms. By collecting patients' voice data and using 22 different parameters, including average vocal fundamental frequency, maximum vocal fundamental frequency, minimum vocal fundamental frequency, and jitter, the model achieves high accuracy in diagnosing and predicting patient symptoms.

Keywords: Machine Learning, XGBoost, Parkinson's Disease, Classification.

1 Introduction

Parkinson's disease is a degenerative disease of the nervous system of the brain and is one of the common brain system-specific diseases. It causes tremors in the body and hands, and makes the body stiff [1, 2]. The disease progresses to an advanced stage for which there is currently a definite cure. This is why early detection of the disease is so important. Early detection of Parkinson's disease not only reduces the cost of the disease, but may also save lives [3]. Common symptoms of Parkinson's disease include: depression, anxiety, sleep and memory related problems, loss of smell and balance problems. The causes of Parkinson's disease are not known, but studies by researchers suggest that several factors are responsible for the disease. Genetic factors, for example, studies have identified certain very rare mutated genes, genetic

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_2

variants that usually increase the risk of Parkinson's disease, and environmental factors, due to certain harmful toxins or chemicals found in the environment that can trigger the disease but have a smaller impact. Although it develops at the age of 65, 15% can be found in young people under the age of 50 [4, 5].

Early diagnosis is important for people with Parkinson's, as it can help patients get earlier treatment to manage their condition so that their quality of life is even comparable to that of a normal person. 90% of people with Parkinson's develop speech disorders early in life, which they may not notice on their own [6]. Parkinson's disease is primarily assessed subjectively by physicians based on a scale of symptoms. However, this approach can lead to a high rate of misdiagnosis. Wearable sensors have been developed to assist physicians in providing more accurate diagnoses, however, this approach requires physicians and patients to work together in specific settings, which is inconvenient for Parkinson's disease patients with limited movement based on this status [7]. This investigation resulted in the development of a diagnostic solution based on a learning algorithm.

Bayestehtashk et al. chose a support vector machine with RBF kernel function, and finally obtained a classification accuracy of 55% after analyzing the training set data and optimizing the classifier parameters [8]. In contrast to the direct application of sample features for classification, Betul et al. used data analysis methods to correlate the speech features and UPDRS scores for Parkinson's to obtain an optimal UPDRS threshold for classifying Parkinson's disease, obtaining the highest classification accuracy of 96.4% [9].

Utilizing a training dataset and a test dataset, this study investigates the potential of the XG Boost algorithm, using machine learning, for the diagnosis of Parkinson's disease.

2 Method

2.1 Dataset

The data set was created by Max Little of Oxford University in collaboration with the National Center for Speech and Voice in Denver, Colorado, which records speech signals [10]. In the initial study, a technique for extracting features related to speech problems was introduced. The dataset consisted of biological speech measurements obtained from 31 participants, 23 of whom were diagnosed with Parkinson's disease (PD). The table lists different speech metrics, identified by their corresponding "Name" column, with each row corresponding to a single recording from one of the 195 individuals. The primary objective of the data is to distinguish between healthy individuals and those afflicted with PD, with the "Status" column displaying 0 for healthy and 1 for PD.

2.2 XGBoost Algorithm

Extreme gradient boosting, also known as XG Boost, is a popular high-performance machine learning algorithm that excels in classification and regression issues. It is an

integrated learning approach that enhances a model's ability to predict outcomes by repeatedly training different decision trees. The advanced features of XG Boost, including as regularization, parallel processing, and missing value processing, can improve the model's robustness and generality. XG Boost, one of the best machine learning algorithms currently in use, has won numerous machine learning competitions.

XG Boost is a machine learning algorithm that combines distributed gradient boosting with regularized loss functions and second-order Taylor expansions for improved generalization performance. For this study, a categorical regression tree is utilized as the base learner, with the aim of achieving optimal results with n instances consisting of m-dimensional features.

$$D = \{(x_i, y_i)\}(|D| = n, x_i \in R^m, y_i \in R) \tag{1}$$

$$y_i = \sum_{k=1}^{k} f_k(x_i), f_k \in E \tag{2}$$

x_i is the speech feature of the model input; y_i is the label of whether one has Parkinson's; f_k denotes the k_{th} decision tree; E is the set space composed of all categorical regression trees.

The objective function of XG Boost is composed as follows:

$$L = \sum_{i=1}^{m} l(y_i, y_i) + \sum_{k=1}^{k} N(f_k) \tag{3}$$

$l(y_i, y_i)$ is the cross-entropy $(y_i, \log y_i)$ loss function. To prevent overfitting and enhance the generalization performance of the proposed approach, XG Boost incorporates a regularization term that controls the complexity of the tree. This regularization term is added in order to achieve the following:

$$N(f_k) = \lambda\Gamma + \frac{1}{2} y \parallel \omega\parallel^2 = \lambda\Gamma + \frac{1}{2} \sum_{i=1}^{\Gamma} \omega_i^2 \tag{4}$$

where y and λ are the model weights and the penalty parameters of the leaves, respectively; ω_1 is the weight of leaf node l; and Γ is the number of leaf nodes. In the training process, forward stepwise addition is used, i.e., after each iteration of training, the incremental function $f_t(x_i)$ is added to the model to optimize the objective function.

$$L^t = \sum_{i=1}^{m} l(y_i, y_i^{t-1} + f_t(x_i)) + N(f_t) + \Omega(f_t) \tag{5}$$

Taylor expansion of the loss term at y(t-1) is:

$$l^{(t)} \approx \sum_{i=1}^{n} \left[l\left(y_i, y_i^{t-1}\right) + g_i f_t^2(x_i)\right] + \Omega(f_t) \tag{6}$$

g_i and h_i are the first- and second-order partial derivatives (gradients) of L, respectively, removing the constant terms：

$$L^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{7}$$

The sample is decomposed into samples at each node, so the objective function is reduced to a uniform sum over the leaf nodes:

$$\tau(t) = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)\right] + yT + \frac{1}{2}\lambda \sum_{j=1}^{T}\omega_j^2 \tag{8}$$

$$\omega_j^2 = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}$$

The optimal solution is obtained by bringing in the objective function:

$$L(t)(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + yT \tag{9}$$

A greedy approach is used to discover the best splitting point because, in most cases, it is not practical to enumerate all tree structures to choose the best outcome. Iterative splits are used to add nodes to the tree starting from a single leaf node, and the loss function following the cut point slice is as follows:

$$L_{split} = \frac{1}{2}\left[\frac{G_L^2}{H_L + y} + \frac{G_R^2}{H_R + y} - \frac{\left(G + G^R\right)^2}{H_L + H_R + y}\right] - \lambda \tag{10}$$

where R and L stand for the respective branches of the right and left subtrees. performs the function of threshold splitting and managing the tree's complexity, and the split is only carried out to do prepruning when the gain after splitting is larger than.

2.3 SVM Algorithm

Cortes and Vapnik introduced the Support Vector Machine (SVM) in 1995 to address problems related to pattern recognition such as small sample sizes, nonlinearity, and high dimensionality. Since then, SVM has shown many significant advantages in solving these issues. It can be applied to further machine learning tasks as well, including function fitting. It has a number of traits. The SVM algorithm requires a relatively small amount of samples, therefore it can learn from small samples. Second, SVM's nonlinearity, or ability to effectively handle the situation where the sample data is linearly inseparable, is primarily accomplished through relaxation variables and kernel function approaches. Thirdly, high-dimensional pattern recognition refers to the situation where the sample dimension is very high, such as tens of thousands of dimensions. Other algorithms are essentially unable to handle this, but SVM can. This is mainly due to the classifier generated by SVM being extremely concise and using very little sample information, which makes even if the sample dimension is very high, it does not bring about much trouble to storage and computation.

2.4 KNN Algorithm

The closest neighbor algorithm, or K-Nearest Neighbor (KNN) classification

algorithm, is one of the most elementary classification techniques used in data mining. Its basic idea is to represent a sample by its k nearest neighbors. This simple and efficient approach has made it a popular choice in machine learning. KNN is classified as a lazy-learning algorithm, which means that the classifier can learn without utilizing the training data. The computational complexity of KNN classification is determined by the total number of documents in the training set, resulting in a time complexity of O(n) when n documents are present. Although the KNN approach relies heavily on local neighborhood samples rather than a formal method of differentiating class domains, it is particularly suited for classifying sets of samples with significant intersection or overlap among class domains compared to other approaches.

2.5 Random Forest Algorithm

The random forest algorithm is composed of multiple decision trees, which act as its building blocks. The trees are integrated through integrated learning, resulting in an ensemble approach that is fundamental to machine learning. "Random" and "forest" are critical components of its name. Intuitively, each tree functions as a classifier, thus N trees generate N classification outcomes for a given input sample. The final output is determined by selecting the category with the highest vote after incorporating all classification results from the random forest. This algorithm ' s distinguishing characteristics include its higher accuracy when compared to other algorithms, its ability to operate efficiently on large datasets, and its ability to process input samples without dimensionality reduction. Additionally, it can judge the significance of individual features on classification problems and provide an unbiased estimate of the internal generation error during the generation process.

3 Results and Discussion

3.1 Feature Distribution

In order to provide a clearer description of the distribution of each feature ' s characteristics within the dataset, a bar chart was created and displayed in Figure 1. Notably, some of the key features included trend fluctuation analysis (DFA), harmonic signal-to-noise ratio (HRN), fundamental frequency perturbation (Jitter), and amplitude perturbation (Shimmer). On the other hand, fundamental frequency (F0) and noise harmonic ratio (NHR) did not show significant differences when compared to normal individuals. The graph indicates that Parkinson ' s disease patients have a significantly smaller HNR compared to healthy individuals, due to notable speech symptoms such as reduced speech loudness, increased vocal fold tremors, and dyspnea (noise) in the speech samples of Parkinson ' s disease patients. HNR can effectively capture these symptoms, enabling differentiation between healthy and speech-disordered individuals. The technique known as trend fluctuation analysis (DFA) is utilized to study the long-range correlation of a signal ' s time series, primarily to determine the extent to which noise in speech signals is self-resembling. The airflow along the vocal tract, where sound is produced, is the primary cause of

noise within speech signals. Vocal cord lesions from Parkinson's disease patients might affect this noise in a random way. As a result, the variation in random noise can aid in determining the condition of the subject.

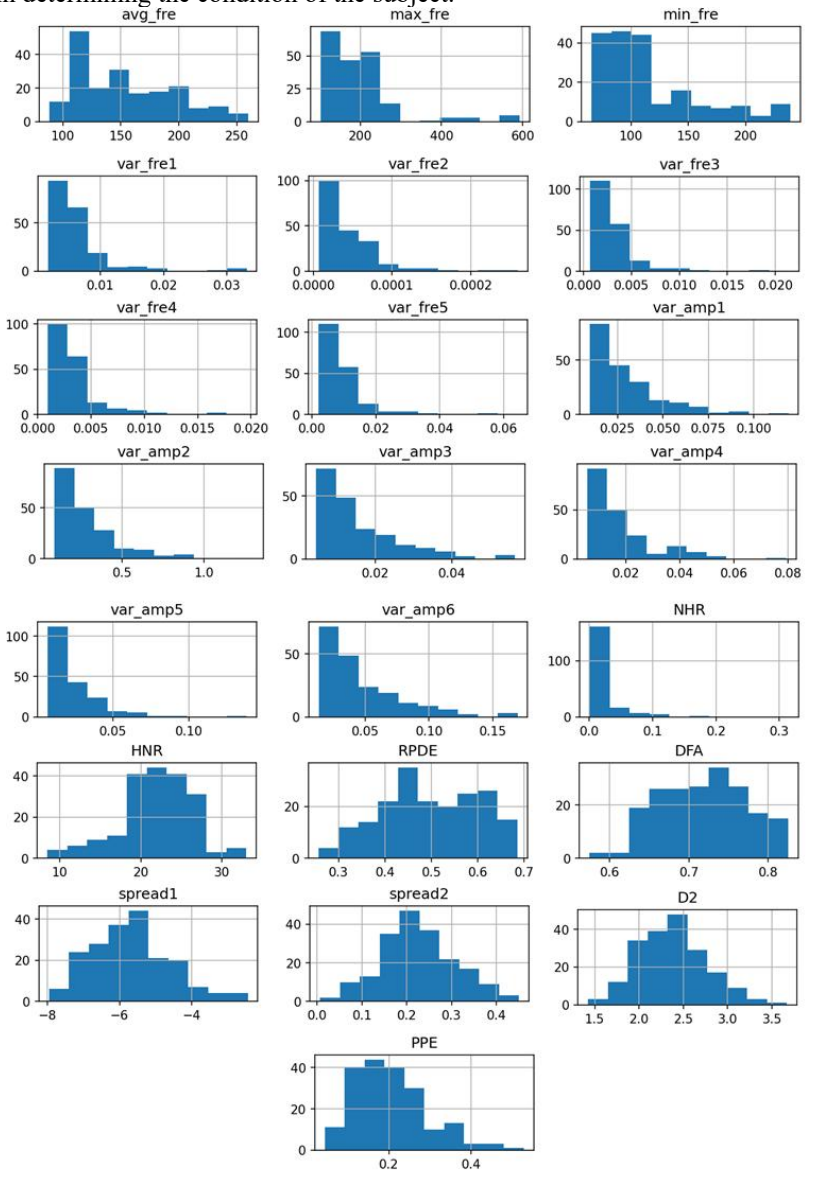

Fig. 1. Feature Distribution (Picture credit: Original).

DFA comprises two primary components: analyzing trends in changes in speech signals, and analyzing fluctuations in speech signals based on these trends. Parkinson's disease increases speech signal noise in patients, making trend fluctuation

analysis useful for observing disease progression. Therefore, Parkinson's disease diagnosis depends primarily on features such as HRN and DFA. Although features such as Jitter and Shimmer also affect patient conditions, their impact is considerably less than that of HRN and DFA.

3.2 NHR Density Distribution

Figure 2 illustrates this feature, which measures the proportion of noise to tone components in the voice. It is clear from Figures 1-2 that this feature is centrally distributed between 0.00 and 0.02 and is not the primary indicator of Parkinson's disease.

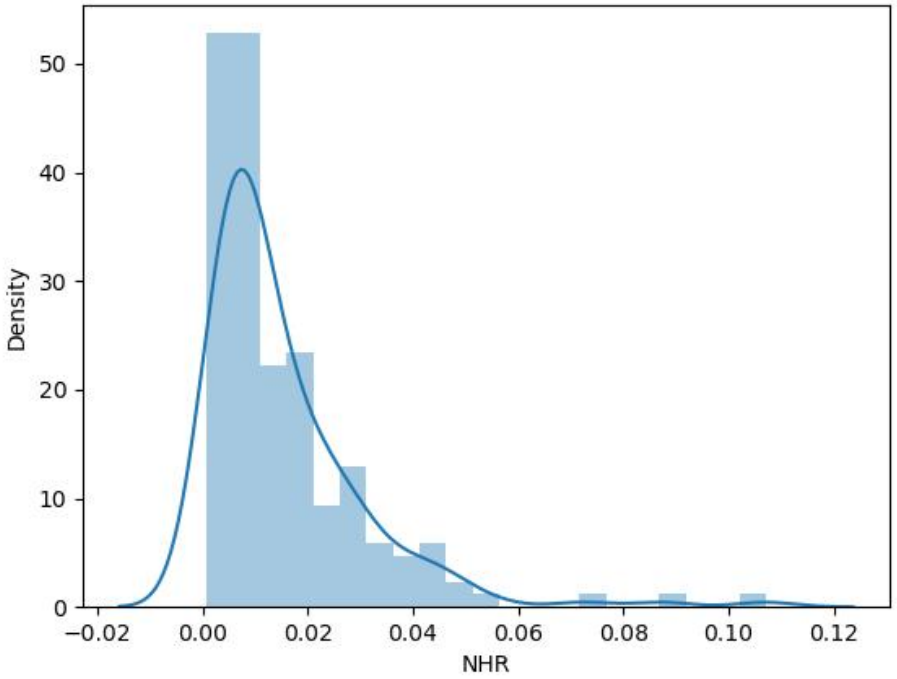

Fig. 2. Distribution of NHR characteristics (Picture credit: Original).

3.3 Result Comparison

As shown in Table 1, the prediction accuracy of XG Boost is 94.87%, stronger than the other three algorithms. The reason for this result is XG Boost's own characteristics. Firstly, efficiency. By optimizing parallel processing and improving the cache access mode, XG Boost achieves faster execution than other gradient boosting tree based algorithms. Secondly, accuracy. By adaptive learning rate, regularization and parallelization techniques, more accurate prediction results can be obtained on various data types. Thirdly, scalability. Support parallel processing and distributed computing, can handle large data sets, and can run on multiple platforms. Fourthly, flexibility.

Provide a variety of parameter adjustment methods and customizable functions, which can be flexibly adjusted and optimized according to the different characteristics of the data.

Table 1. Prediction accuracy.

Models	Accuracy
SVM	87.17%
XG Boost	94.87%
KNN	89.74%
Random Forest	93.87%

4 Conclusion

Additional research can be conducted to evaluate the feasibility of clinical diagnosis and the model's accuracy. For example, the model's sensitivity, specificity, and false positive rate can be studied, and appropriate parameters can be adjusted or corrected to improve the model's overall performance.To improve early diagnosis of Parkinson's disease, it is crucial to take into account the practicality of the model's application, its appropriateness for widespread use in clinical practice, and the particular scenarios in which it may be utilized. Parkinson's disease is a degenerative disorder that mainly affects individuals over the age of 60. The early stages of Parkinson's disease are frequently misidentified as age-related physical decline due to the unclear pathogenesis, absence of standardized diagnostic criteria, and the possibility of missed or incorrect diagnoses with some forms of medication. With China's aging population, an increasing number of individuals are diagnosed with Parkinson's disease, leading to significant research implications for the development of innovative diagnostic techniques. This investigation begins with speech feature-based Parkinson's disease diagnosis, an area of study that has gained popularity in recent years, utilizing the XG Boost algorithm to train the data set and accomplish its objectives.

References

1. Bloem, B. R., Okun, M. S., & Klein, C.: Parkinson's disease. The Lancet, 397(10291), 2284-2303 (2021).
2. Blauwendraat, C., Nalls, M. A., & Singleton, A. B.: The genetic architecture of Parkinson's disease. The Lancet Neurology, 19(2), 170-178 (2020).
3. Tolosa, E., Garrido, A., Scholz, S. W., & Poewe, W.: Challenges in the diagnosis of Parkinson's disease. The Lancet Neurology, 20(5), 385-397 (2021).
4. Vázquez-Vélez, G. E., & Zoghbi, H. Y.: Parkinson's disease genetics and pathophysiology. Annual review of neuroscience, 44, 87-108 (2021).
5. Yang, W., Hamilton, J. L., Kopil, C., Beck, J. C., Tanner, C. M., et al.: Current and projected future economic burden of Parkinson’s disease in the US. npj Parkinson's Disease, 6(1), 15 (2020).

6. Senturk, Z. K.: Early diagnosis of Parkinson's disease using machine learning algorithms. Medical hypotheses, 138, 109603 (2020).
7. Rovini, E., Maremmani, C., & Cavallo, F.: How wearable sensors can support Parkinson's disease diagnosis and treatment: a systematic review. Frontiers in neuroscience, 11, 555 (2017).
8. Bayestehtashk, A., Asgari, M., Shafran, I., & McNames, J.: Fully automated assessment of the severity of Parkinson's disease from speech. Computer speech & language, 29(1), 172-185 (2015).
9. Betul, O., Ibrahim, K. A., Berna, D., Temel, T., & Handan, A.: Clinical significance of serum lncRNA H19, GAS5, HAR1B and linc01783 levels in Parkinson's disease. Ideggyogyaszati szemle, 76(5-6), 189-196 (2023).
10. Little, M., Mcsharry, P., Roberts, S., Costello, D., & Moroz, I.: Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection. Nature Precedings, 1-1 (2007).