Predicting Parkinson’s Disease using Machine Learning based on Biomedical Voice Measurement

# 1st Ngo Thi Thu Ha

*School of Electrical and Electronics Engineering Ha Noi University of Science and Technology* Ha Noi, Viet Nam

[ha.ntt203836@sis.hust.edu.vn](mailto:ha.ntt203836@sis.hust.edu.vn)

# 3rdHoang Quang Huy

*School of Electrical and Electronics Engineering Ha Noi University of Science and Technology* Ha Noi, Viet Nam [huy](mailto:huy.hoangquang@sis.hust.edu.vn)[.hoangquang@sis.hust.edu.vn](mailto:.hoangquang@sis.hust.edu.vn)

# 2nd Le Minh Duc

*School of Electrical and Electronics Engineering Ha Noi University of Science and Technology* Ha Noi, Viet Nam [duc.lm203833@sis.hust.edu.vn](mailto:duc.lm203833@sis.hust.edu.vn)

# 4th Nguyen Viet Dung

*School of Electrical and Electronics Engineering Ha Noi University of Science and Technology* Ha Noi, Viet Nam [dung.nguyen](mailto:dung.nguyenviet1@sis.hust.edu.vn)[viet1@sis.hust.edu.vn](mailto:viet1@sis.hust.edu.vn)

***Abstract*—Parkinson’s disease (PD) is a prevalent neurolog- ical condition that occurs with age and manifests in various motor and cognitive symptoms. Detecting and diagnosing PD early and accurately is essential for effective treatment and care. Biomedical speech analysis has emerged as a promising non-invasive approach for detecting and monitoring PD. We collected a dataset from the University of California at Irvine (UCI) machine learning repository, which comprised 195 voice recordings obtained from 31 patients during their examinations. Our research explores the use of a dataset to investigate several machine learning algorithms as potential diagnostic tools for Parkinson’s disease. The study employs classification algorithms with feature extraction , and k-fold cross-validation to ensure robust evaluation. At the end, we found that Random Forest with a ratio of 70:30 train/test split or 5-fold cross- validation gave the best results for our project. Random Forest with an overall accuracy of 96.61%, and f1-score of 0.98. In addition, selection data has highly correlated features, we obtained a higher level of accuracy compared to using the entire data set for all machine learning algorithms. Our results indicated that this algorithm has the potential for diagnosis of Parkinson’s Disease.**

***Index Terms*—machine learning, parkinson’s disease, voice measures, feature extraction**

1. Introduction

Parkinson’s disease (PD) is a common neurodegenerative disorder affecting millions of people worldwide [1]. In the United States, it is estimated that 1.04 million people were diagnosed with Parkinson’s disease in 2017, with a total economic cost of $51.9 billion.Considering that a significant number of people remain undiagnosed or receive incorrect diagnoses, it is highly probable that the actual figure is considerably larger. PD was more common in the 65 group and in men [2]. Of course, because the difficulties associated with Parkinson’s disease increase, the condition affects many more spouses, husbands, children, and other caretakers. While there is no cure for PD, early detection and timely intervention can significantly improve patient outcomes. Current clinical methods rely on subjective assessments by physicians and

patient self-reporting, which can be time-consuming and prone to errors [1]. Thus, monitoring the progression of PD can be challenging, especially in the early stages when symptoms may be mild or non-specific.

Recent studies have shown that PD patients exhibit char- acteristic vocal features, such as hoarseness, monotony, and reduced loudness, which can be objectively measured through voice recordings [3], [4]. Along with the strong development of artificial intelligence technologies, many solutions to sup- port the diagnosis of PD diseases have been researched and developed. Machine learning algorithms can analyze these vocal features and accurately predict PD diagnosis, even in the early stages of the disease [4]. In addition, there are some studies of early diagnosis of Parkinson’s disease based on handwriting [5]. The gait analysis and upper extremity motor assessment are also a method for diagnosing Parkinson’s disease using Machine Learning.

Therefore, in this study, we aim to explore the potential of machine learning algorithms in predicting PD using vocal features extracted from voice recordings. We hypothesize that machine learning can provide a non-invasive, cost-effective, and accurate diagnostic tool for PD. Our mainly contribution is summary seleted correlation features combine k- fold cross- validation to achieve high accuracy.

1. Data set

The data set used in the current study was analyzed and pub- lished available at [6]. Data published methods for general voice disorders, include 195 vowel pronunciations . This data set contains a variety of biological voice measurements from 31 patients, 23 of whom have Parkinson’s disease (PD). The range for the diagnosis is 0 to 28 years of age. The age range of the survey participants was 46 to 85 due to the specificity of the condition. Each participant provided an average of six transcriptions, with each one lasting between one and 36 seconds.Every column in the dataset represents a specific

vocal measurement, while each row corresponds to one of the 195 recorded voices from these individuals, identified by the ”name” column. The scatter matrix some features is shown in 1.

The data set used consists of several attributes. These include the subject’s name and recording number in ASCII format. They also include measurements related to vocal fundamental frequency and amplitude. NHR and HNR,indicate the proportion of noise to tonal components present in the voice. The status attribute indicates the health status of the subject, with a value of one indicating Parkinson’s disease and zero indicating a healthy status. The attributes RPDE and D2 represent nonlinear dynamical complexity measures, while DFA represents the signal fractal scaling exponent. Finally, three nonlinear measures of fundamental frequency variation are recorded: spread1, spread2, and PPE.

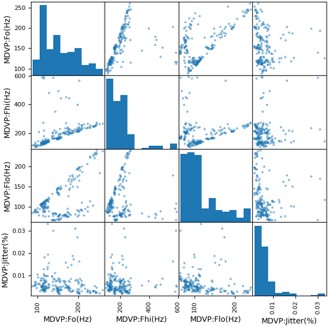


Fig. 1. Scatter matrix some features

1. Related work

Previous studies have investigated recognizing Parkinson disease using machine learning. Drotar has presented detailed reviews of previous studies et al. [7] conducted handwriting analysis from both a surface writing and a hand movement perspective when writing in the air. They identified Support Vector Machine (SVM) as the classification model and attained 80% accuracy. To detect disease based on brain tumor, Arjurkar and Ridhorkar [8] conducted an extensive in- vestigation into the effectiveness of machine learning methods in decision support systems for diagnosing Parkinson’s dis- ease. Their research revealed that boosted logistic regression outperformed all other algorithms, demonstrating remarkable accuracy in identifying patients with Parkinson’s disease. Furthermore, Gait Measures is one of the ways to detect PD.

Alex Li and Chenyu Li [9] used different machine learning methods, including logistic regression, SVM, decision tree, k-nearest neighbors (KNN) with impression accuracy.

To detect Parkinson’s disease based on voice signal, some research based on data Max A. Little at al. [6] create in their paper . They develop pitch period entropy (PPE), a new measure of dysphonia that is resistant to several uncontrollable confounding factors such as loud acoustic surroundings and typical, healthy changes in voice frequency. They use a kernel support vector machine to have high accuracy 91.4%. Max A Little also research in another paper with Athanasios Tsanas.et al about accurate tele-monitoring of PD. This groundbreaking study displays the ability to quickly and accurately assess UPDRS (Unified Parkinson’s Disease Rating Scale) remotely, with results that closely align with clinicians’ evaluations. This achievement is accomplished through uncomplicated, self- administered, and noninvasive speech tests. The significance of this research lies in its potential to contribute to tele- monitoring systems [10]. In their research, Arti Rana et al

[11] investigated the effectiveness of supervised classification algorithms, including SVM, Na¨ıve Bayes, k-nearest neighbor (K-NN), and artificial neural network (ANN), for Parkinson’s disease detection. The experimental results demonstrated that the artificial neural network (ANN) achieved the highest level of accuracy among the tested algorithms Moreover, Deep learning is also used to predict this disease. Following machine learning approach, Mehrbakhsh Nilashi et al. [12] predicted PD by employing techniques such as noise reduction, clustering, and prediction methodology. They used Principal Component Analysis (PCA) and Expectation Maximization (EM) then applied Adaptive Neuro-Fuzzy Inference System (ANFIS) and Support Vector Regression (SVR). The results also indicated that all of the approaches used in this study can significantly improve the accuracy of PD prediction and apply it to other classification and prediction problems within the medical domain. In addition to [13], the authors used Synthetic Minority Over-sampling Technique (SMOTE), hyperparameter tuning (GridSearchCV) to process data. For this study, the accuracy of Random Forest is 92%. In the same data set

, Grover et al. also applied deep learning-based techniques. In their study, which used Deep Neural Network (DNN), they were able to diagnose Parkinson’s disease symptoms in individuals with an accuracy of about 81.67% [14]. Author in [15] proposed two deep learning frameworks to classify speech data. They use two frameworks and merged features were then classified using a 10-layer CNN, comprising four consecutive convolution layers with each pair followed by a max pooling layer, a fully connected layer, and a final output layer. Two frameworks achieved accuracy of 84.5% and 86.8%, respectively.

1. Methodology

This flow chart in figure 2 illustrates the process of devel- oping our machine learning algorithm. The first stage involves.

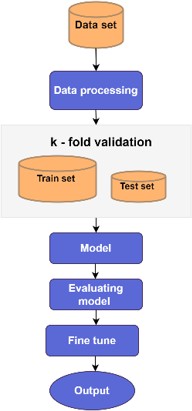


Fig. 2. Flow chart

obtaining and importing the data set, which serves as the foundation for the entire process. Subsequently, the data goes through data exploring and pre-processing, where it is cleaned, extracted features, and prepared for analysis. We split data in two ways. Firstly, data is divided into two subsets: the test set and the training set with the ratio is 7 - 3. The test set acts as a benchmark to evaluate the model’s performance, while the train set is used to train the algorithm. Secondly, we use cross-validation to split data. The goal of our project is to classify based on their ”status” value. A ”status” value of 0 corresponds to healthy individuals, 48 samples. On the other hand, a ”status” value of 1 indicates patients with Parkinson’s disease.Then, we train with classification model. After the model has been trained, it moves on to the evaluating model phase. The model’s accuracy, F1-score are measured. We train various models using the data set to figure out the most desirable algorithm.

1. *Feature extraction*

In our research, we calculate Pearson’s correlation coeffi- cient in equation (1)

cov(*X, Y* )

Using correlation between multiple variables to remove features with low correlation. The removal of highly correlated features helps the linear model work better, avoiding bias between features. The extracted features that were utilized for model training consisted in the Table I

TABLE I

Description and Statistic of Extracted Features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No. | Feature | Mean | Max | SD |
| 1 | MDVP:Fo(Hz) | 154.2 | 260.1 | 41.39 |
| 2 | MDVP:Fhi(Hz) | 197.1 | 592.0 | 91.49 |
| 3 | MDVP:Flo(Hz) | 116.3 | 239.2 | 43.52 |
| 4 | MDVP:Jitter(%) | 0.006 | 0.033 | 0.005 |
| 5 | MDVP:Jitter(Abs) | 0.00004 | 0.0003 | 0.00003 |
| 6 | MDVP:RAP | 0.003 | 0.021 | 0.003 |
| 7 | MDVP:PPQ | 0.003 | 0.019 | 0.003 |
| 8 | Jitter:DDP | 0.010 | 0.064 | 0.009 |
| 9 | MDVP:Shimmer | 0.030 | 0.119 | 0.019 |
| 10 | MDVP:Shimmer(dB) | 0.282 | 1.302 | 0.194 |
| 11 | Shimmer:APQ3 | 0.016 | 0.056 | 0.010 |
| 12 | Shimmer:APQ5 | 0.018 | 0.079 | 0.012 |
| 13 | MDVP:APQ | 0.024 | 0.138 | 0.017 |
| 14 | Shimmer:DDA | 0.047 | 0.169 | 0.030 |
| 15 | NHR | 0.025 | 0.315 | 0.040 |
| 16 | spread1 | -5.684 | -2.434 | 1.090 |
| 17 | status | 0.754 | 1 | 0.432 |
| 18 | PPE | 0.207 | 0.5273 | 0.090 |

1. *Cross-validation*

Cross-validation is a statistical method used when optimiz- ing model hyperparameters to estimate the skill of machine learning models. This method also helps to detect overfitting and the data sets utilized are split into multiple subsets of equal size through the process of cross-validation. In our research, we use k-fold cross-validation and *k* = 5. This ap- proach is widely used because of its simplicity and its ability to yield less biased evaluations compared to other methods like a traditional train and test split. This approach utilizes a technique known as k-fold cross-validation, which involves randomly partitioning the observations into k folds. The first fold serves as the validation set, while the model is trained on the remaining *k* = 1 folds.

1. *Classification based on machine learning.*
   1. *Decision Tree:* A decision tree is a supervised learning algorithm employed for performing classification and regres- sion tasks in machine learning. It is a predictive modeling technique that builds a flowchart-like structure, resembling a tree, by recursively partitioning the data based on a set of decision rules.

There are two types of decision tree. The first type is regression trees which predict continuous values based on previous data or information sources. The second type and the one we used to be classification trees determine whether an

*ρ* =

*σxσy*

(1)

event happened or did not happen which involves a “yes” or “no” outcome.

* 1. *Random Forest:* A random forest can be described as a

classification model that comprises a set of tree-based classi- fiers *{h*(*x, k*)*, k* = 1*, ...}* [16] where the *{k}* are independent distribute each tree casts a unit vote for the most popular class

at input x. The random forest algorithm is recognized for its robustness, scalability, and its capability to effectively handle high-dimensional data sets. Leo Breiman and Adele Cutler trademark Random Forest.

The hyperparameters of the random forest algorithm consist.

TABLE II

Confusion Matrix

Actual Value

|  |  |  |
| --- | --- | --- |
|  | Positive | Negative |
| Positive | *T P* | *FP* |
| Negative | *T N* | *FN* |

Predicted Value

and recall are high, a model will achieve a high F1 score, and conversely. In confusion matix, it calculated by (3)

of the node size, which determines the minimum number of

2 *∗ Precision ∗ Recall*

2 *∗ TP*

samples required to form a node, the number of trees in the

*F* 1 =

= (3)

*Precision* + *Recall* 2 *∗ TP* + *FP* + *FN*

forest, and the number of features randomly selected for each tree during training. Random Forest offers several advantages over individual decision trees. The algorithm is less prone to overfitting due to the ensemble nature of combining multiple.

Accuracy: Accuracy is a commonly used metric to evaluate classification models. It is the percentage of correct predictions made by the model. Accuracy is defined as follows in (4) :

trees while also capable of handling high-dimensional and large data sets effectively.

* 1. *Logistic Regression:* Logistic Regression is another su-

*B. Result*

*TP* + *TN*

*Accuracy* =

*TP* + *TN* + *FP* + *FN*

(4)

pervised machine learning algorithm used for binary classifi- cation tasks. Despite its name, logistic regression is primarily used for classification, not regression tasks.

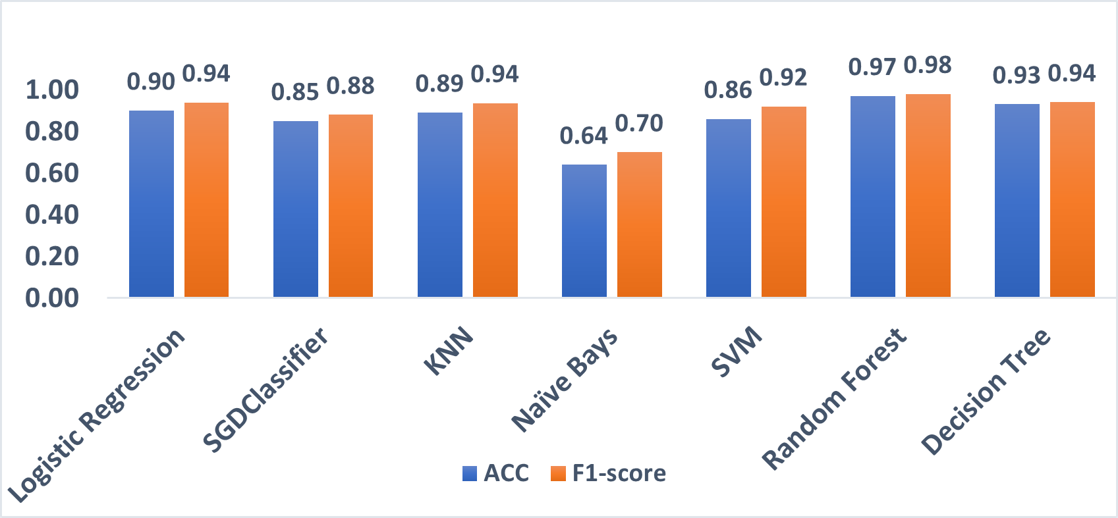
The algorithm models the relationship between the input features and the probability of the binary outcome using the logistic function, also known as the sigmoid function. The sigmoid function maps any real-valued number to a value between 0 and 1, which can be interpreted as the probability of belonging to the positive class.

* 1. *Na¨ıve Bayes Classifier:* Naive Bayes is a probabilistic machine learning algorithm commonly used for classification tasks. It is based on Bayes’ Theorem, which calculates the probability of an event given prior knowledge or evidence. The Bayes’ Theorem is described mathematically in form of the following Equation (2)

An overview of the results for some models above, and KNN, SGD Classifier, SVM. It can be seen in 3. A chart comparing the results presents the results for algorithms trained on normal train/test splited data. For these metrics , Logistic Regression, Random Forests (RF) and Decision Tree have an accuracy larger than 90%. This evaluation is to show that RF tells us the best percentage of correctly classified, while Na¨ıve Bayes has 64% accuracy. The experimental outcomes show that the F1- score of SVM is 0.92, RF is 0.98, KNN is 0.94, and Logistic Regression is 0.94. It is concluded that the RF is the most accurate one with the highest both accuracy and F1-score. Additionally, Table III shows results of different models on

*P* (*Y |X*) =

*P* (*X|Y* ) *∗ P* (*Y* ) *P* (*X*)

(2)

Through the application of this theorem, the probabilities *P* (*X|Y* ) and *P* (*Y* ) can be calculated from the training data. However, as the number of feature increases, this process

becomes much more difficult and requires more processing power. This leads to the naive variation of the theorem extending and simplifies the original Bayes theorem, which provides a mechanism for determining the probability of a target occurrence.

1. Result and Discussion

*A. Evaluation Metrics*

In order to detect PD, we employ various classifier metrics including accuracy (ACC), F1-score, and other metrics found in the confusion matrix. Each of these evaluation criteria has its own calculation formula, which aids in determining the most suitable classifier for the analysis in terms of quality.The confusion matrix is displayed in Table II.

F1-Score: The F1 score is a measurement that represents the harmonic mean of precision and recall. When both precision

Fig. 3. The different machine learning-based algorithms

the training and test datasets. One interesting finding is that for 5-fold cross- validation, the accuracy performs worse. As future work, we plan to perform an extensive analysis of all methodologies on this data set.

In Table III, the results of the models displayed include 70:30 train/test split, 5-fold cross-validation for four model. The table shows that the maximum accuracy of 94.92% for all features and 96.61% for extracted features is achieved by the Random Forest algorithm. This performance is closely followed by Decision Tree and Logistic Regression algorithms. Finally, with the worse performance, Na¨ıve Bayes achieved the lowest accuracy of 67.8% and 64.41% for all features and extracted features, respectively.

*C. Discussion*

TABLE III

Training Model Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No. | Model | Split | data | All fea-  ture | Extracted  feature |
| 1 | Decision Tree | 70 -  k-fold | 30  (n=5) | 91.53  84.10 | 93.22  88.72 |
| 2 | Random Forest | 70 -  k-fold | 30  (n=5) | 94.92  90.77 | 96.61  91.28 |
| 4 | Na¨ıve Bayes | 70 -  k-fold | 30  (n=5) | 67.8  69.74 | 64.41  68.21 |
| 5 | Logistic Regression | 70 -  k-fold | 30  (n=5) | 89.83  85.13 | 89.83  84.1 |

TABLE IV

Compare result previous works

|  |  |  |
| --- | --- | --- |
| References | Method | Accuracy |
| [11] | SVM  na¨ıve Bayes KNN | 87.17%  74.11%  87.17% |
| [17] | ANN  RF  Na¨ıve Bayes | 94.87%  87.17%  71.79% |
| [18] | Decision tree | 85.08% |
| [19] | Random Forest | 90.26% |
| Our Research | Logistic Regression  RF  Decision Tree | 89.83%  96.61%  93.22% |

machine learning algorithms based on biomedical voice mea- surement.

Overall, our findings revealed that Random Forest algorithm achieved the highest prediction accuracy among the evaluated models, and thus was chosen as the output model. Additionally, we employed the K-fold cross-validation technique to mitigate the risk of overfitting and ensure the robustness of our models. However, improving the models is still likely to produce improved performance and greater accuracy. There are other metrics that can be implemented in the evaluation model.

References

1. J. I. S. Lawrence I. Golbe, Margery H. Mark, *Parkinson’s Disease handbook*. American Parkinson Disease Association, Inc., 2014.
2. W. Yang, J. Hamilton, C. Kopil, J. Beck, C. Tanner, R. Albin, E. Dorsey,

N. Dahodwala, I. Cintina, P. Hogan, and T. Thompson, “Current and projected future economic burden of parkinson’s disease in the u.s,” *npj Parkinson’s Disease*, vol. 6, 12 2020.

1. K. K. L. A. Ma and D. Thyagarajan, “Voice changes in parkinson’s disease: What are they telling us?” *Journal of Clinical Neuroscience*, 2019.
2. M. C. et al, “Characterisation of voice quality of parkinson’s disease using differential phonological posterior features,” *Computer Speech & Language*, 2017.
3. “Handwriting dynamics assessment using deep neural network for early identification of parkinson’s disease,” *Future Generation Computer Systems*, vol. 117, pp. 234–244, 2021.
4. 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*,

In our research, we employed two different data splitting ap- proaches, and both of them yielded favorable outcomes. When we extracted data and specifically selected highly correlated features, we obtained a higher level of accuracy compared to using the entire data set.The Random Forest model achieved the best performance with an accuracy rate of 96.61% and an F1-score of 0.98, representing the highest outcome for our research. Random forests are renowned for their accuracy and effectiveness due to the utilization of multiple decision trees. Although each decision tree in the Random Forest algorithm utilizes a subset of the training data, which may result in potential underfitting, the aggregation of predictions from numerous decision trees compensates for this limitation. Numerous researchers used UCI Parkinson’s dataset in classification processes. The result is compared with those of previous works that made use of the same data shown in Table IV. It shows that [18] a Decision Tree accuracy of only 85.08%, our study achieved a significantly higher accuracy index of 93.22%. This outcome further strengthens the validity.

and the effectiveness of our data processing techniques.

1. Conclusion and Future work

Parkinson’s Disease is a widespread neurodegenerative dis- order that can severely affect the patient’s living condition and is currently incurable. While current assessment methods proved to be highly problematic in detecting PD, research and studies show that vocal features can allow for more efficient and accurate diagnostic methods. Due to this reason, our research focused on predicting Parkinson’s Disease using.

vol. 56, no. 4, pp. 1015–1022, 2009.

1. P. Drota´r, J. Mekyska, I. Rektorova´, L. Masarova´, Z. Sme´kal, and

M. Faundez-Zanuy, “A new modality for quantitative evaluation of parkinson’s disease: In-air movement,” pp. 1–4, 2013.

1. S. Warjurkar and S. Ridhorkar, “A study on brain tumor and parkinson’s disease diagnosis and detection using deep learning,” pp. 356–364, 2021.
2. A. Li and C. Li, “Detecting parkinson’s disease through gait measures using machine learning,” *Diagnostics*, vol. 12, p. 2404, 10 2022.
3. A. Tsanas, M. Little, P. Mcsharry, and L. Ramig, “Accurate telemoni- toring of parkinson’s disease progression by noninvasive speech tests,” *IEEE transactions on bio-medical engineering*, vol. 57, pp. 884–93, 11 2009.
4. A. Rana, A. Dumka, R. Singh, M. Rashid, N. Ahmad, and M. Panda, “An efficient machine learning approach for diagnosing parkinson’s disease by utilizing voice features,” *Electronics*, vol. 11, p. 3782, 11 2022.
5. M. Nilashi, A. P. D. O. Ibrahim, and A. Ahani, “Accuracy improvement for predicting parkinson’s disease progression,” *Scientific reports*, vol. 6,

p. 34181, 09 2016.

1. R. Alshammri, G. Alharbi, E. Alharbi, and I. Almubark, “Machine learning approaches to identify parkinson’s disease using voice signal features,” *Frontiers in Artificial Intelligence*, vol. 6, 2023.
2. S. Grover, S. Bhartia, Akshama, A. Yadav, and S. K.R., “Predicting severity of parkinson’s disease using deep learning,” *Procedia Computer Science*, vol. 132, pp. 1788–1794, 2018. International Conference on Computational Intelligence and Data Science.
3. H. Gu¨ndu¨z, “Deep learning-based parkinson’s disease classification using vocal feature sets,” *IEEE Access*, vol. PP, pp. 1–1, 08 2019.
4. L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 10

2001.

1. S. Marar, D. Swain, V. Hiwarkar, N. Motwani, and A. Awari, “Predict- ing the occurrence of parkinson’s disease using various classification models,” 12 2018.
2. C. Azad, S. Jain, and V. Jha, “Design and analysis of data mining-based prediction model for parkinson’s disease,” *International Journal of Computer Science Engineering (IJCSE)*, vol. 3, pp. 181–189, 05 2014.
3. T. Sriram, M. Rao, G. Narayana, T. Vital, and K. S. Dowluru, “Intelligent parkinson disease prediction using machine learning algorithms,” *IJEIT*, vol. 3, pp. 212–215, 07 2013.