International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

AI-Driven Classification of Alzheimer's and Parkinson's Disease Using Phonetic Speech Patterns

Arkoneil Ghosh

Oberoi International School, Jogeshwari Vikhroli Link Road, Jogeshwari East Mumbai-400060, Maharashtra, India
Email: *ghosharkoneil[at]gmail.com*

Abstract: Neurodegenerative disorders such as Alzheimer's and Parkinson's are challenging to detect early due to their gradual onset. This study investigates the use of machine learning algorithms to identify these conditions based on phonetic features in speech. By analyzing vocal attributes, such as fluency, articulation, and acoustic variation; this research aims to establish non-invasive diagnostic models. Principal Component Analysis (PCA) was used for feature selection, while Random Forest and Support Vector Machine (SVM) classifiers were deployed for detection accuracy. Results show promising accuracy levels, particularly in the Alzheimer's model, highlighting the potential of AI in enhancing early clinical screening for cognitive decline.

Keywords: phonetic speech analysis, Alzheimer's classification, Parkinson's detection, machine learning models, AI in diagnostics

1. Introduction

Neurodegenerative conditions contribute majorly to cognitive impairment and disabilities in the aging population worldwide. Of these, Alzheimer's disease (AD) and Parkinson's disease (PD) are the most frequent and contribute substantially to age-related neurological impairment. The World Health Organization predicts that societies above 60 years of age will reach more than 2 billion by 2050, thus leading to an increase in the prevalence of such disorders. AD, the leading cause of dementia, is characterized by measured memory decline, impaired judgement, and functional deterioration. PD, while it correlates with movement-related symptoms like tremors and rigidity, can also result in cognitive impairment, particularly in the more severe forms. Both these conditions are currently irreversible and tend to be diagnosed at a later stage, when excessive damage has already occurred.

Traditional diagnosis for these illnesses are neuropsychological tests, neuroimaging technologies like MRI and PET scans, and the analysis of cerebrospinal fluid (CSF). Although effective, they are invasive, costly, or not feasible for mass screening. In recent years, researchers have started investigating non-invasive, accessible biomarkers like speech characteristics and phonic defects that could possibly provide early indicators of neurodegeneration (Lin et. al, 2020).

Phonetic speech analysis offers a promising new avenue for early diagnosis. AD and PD have been recognized to impact the brain areas involved in speech production, which results in quantitative alterations in fluency, articulation, and acoustic parameters. Identifying these alterations through machine learning provides a non-invasive, scalable framework for early screening and intervention.

This study demonstrates the creation of an AI-based model that identifies and classifies phonetic speech patterns to differentiate between AD, PD, and healthy controls. The model will use machine learning methods trained on databases publicly available with samples of speech phonetics. Vital machine learning methods such as Principal Component Analysis (PCA) a dimensionality reduction method used to select the prominent features in high-dimensional data and classifying methods such as Random Forest (RF) and Support Vector Machine (SVM) will be considered while creating models. RF is a commonly-used machine learning algorithm, trademarked by Leo Breiman and Adele Cutler, that combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fueled its adoption, as it handles both classification and regression problems. SVM is a supervised machine learning algorithm that classifies data by finding an optimal line or hyperplane that maximizes the distance between each class in an N-dimensional space.

In contrast to conventional statistical models, where assumptions are specified as a priority, machine learning models can learn these complex patterns in the data automatically. This is best suited for the analysis of sophisticated and high-dimensional input like speech. Integrating data-driven learning with real-world, speech-based clinical markers, this project hopes to provide novel inputs to making available tools for neurodegenerative disease detection at an early stage.

In Section 2, we provide a review of statistical methods and AI tools available in literature for the diagnosis of neurodegenerative disorders like Alzheimer's disease (AD) and Parkinson's disease (PD). In Section 3, we go through a list of machine learning algorithms, including linear regression, logistic regression, decision tree, random forest, and support vector machine (SVM). In Section 4, we will develop the AI-based model for Parkinson's disease by explaining the dataset, listing the observations, explain PCA and explain the plot created. In Section 5 we will develop the AI-based model for Alzheimer's disease by using the Random Forest approach, explaining the dataset, and the plot formed. In Section 6 we have a brief conclusion of the findings and the research performed. This study is significant in demonstrating a scalable, non-invasive approach to diagnosing neurodegenerative diseases, potentially contributing to early intervention strategies in public healthcare systems.

2. Literature Review

In the past decade, much work has appeared investigating the application of machine learning for the diagnosis of neurodegenerative disorders like Alzheimer's disease (AD) and Parkinson's disease (PD). The diseases usually remain undiagnosed until considerable mental or motor impairment has set in, hence the importance of early diagnosis. Researchers have turned more and more towards computational methods, particularly those employing machine learning, as substitutes for conventional means of diagnosis based on neuroimaging or cerebrospinal fluid analysis (Myszczynska et. al, 2020).

A principal area of interest within this area has involved the creation of strong multi-class classification models able to identify healthy controls, AD, PD, and other dementia disorders. A range of models have integrated feature extraction methods such as PCA with classifiers including SVM and RF for enhanced accuracy. One of the notable works used PCA for dimensionality reduction, Fisher Discriminant Ratio for feature selection, and was able to achieve 100% accuracy in differentiating between healthy controls, PD patients, and SWEDD (Scans Without Evidence of Dopaminergic Deficit) individuals. This indicates the promise of such techniques in early, non-invasive diagnosis, especially in the case of multi-class classification tasks.

Concurrently, researchers have explored the diagnostic capability of speech features. Neurodegenerative diseases frequently damage brain areas involved in the production of speech, leading to measurable alterations in fluency, articulation, prosody, and timing. These alterations can be documented as phonetic patterns and acoustic markers. Research has indicated that speech-related indicators are early signs of both AD and PD, especially at stages when other symptoms are not yet significant. Yet, most of these studies are preliminary, with small datasets and heterogeneous generalizability.

Machine learning has been particularly useful in pulling out meaningful patterns from these speech data. SVM and RF classifier-based models have been found promising in distinguishing people with cognitive impairment from normal controls. In biomarker-based classification, plasma amyloid-beta, tau proteins, and α-synuclein levels were measured in 377 individuals by a study and classified using LDA and RF classifiers to get an average accuracy of 76% for classifying AD, PD spectrum, and frontotemporal dementia (FTD). Similar attempts have been made to combine speech with biological markers, but researchers are still developing multimodal systems.

More recent work has utilized deep learning models, especially within large epidemiological studies. For example, a longitudinal study in the English Longitudinal Study of Ageing used deep neural networks such as TabTransformer to predict risk of neurodegenerative disease from a broad array of clinical and behavioral characteristics. TabTransformer significantly surpassed conventional Cox regression models as to discriminative ability and time-dependent accuracy, demonstrating the benefits of utilizing flexible neural architectures in survival analysis for population-level predictions.

Even with recent progress, using speech phonetics as a key feature in machine learning studies has not been explored much. Most models focus on what is being said (linguistic content) or how it sounds (acoustic signals), but they do not build systems that classify based only on phonetic patterns. Also, most research looks only at either Alzheimer's Disease (AD) or Parkinson's Disease (PD), not both. And many studies don't use advanced machine learning techniques, like PCA-based feature selection.

This void offers an opportunity. By developing a machine learning model that is based on phonetic speech data to differentiate between AD, PD, and healthy controls, this research aims to span theoretical understanding and clinical utility. Compared to earlier work that tends to rely on costly neuroimaging or invasive biomarker acquisition, the suggested model provides a speech-based, accessible, and scalable diagnostic method. By employing PCA as the dimensionality reduction method and RF and SVM as classifiers, one seeks to create a system with interpretability that enhances early-stage diagnostic performance.

3. Machine Learning

Machine learning (ML), a subsection of artificial intelligence, is the development of computer models that learn on their own from data without rule-based programming. Opposing traditional algorithms, which depend on predetermined logic, ML models make inferences of patterns, correlations, and representations from training data and apply them to predict or decide in novel, unseen situations. This adaptability of ML renders it especially well-adapted to biomedical high-dimensional and complex data, for instance, speech signals, that might harbor subtle, nonlinear patterns that prove difficult for standard statistical methods to capture (Mahesh B, 2020).

In the past decade, the use of ML in medicine has expanded considerably, particularly in the study of neurodegenerative diseases. These have been extensively used to understand medical images (such as MRI and PET scans), genomics, and even health data taken from social media. Their advantage is their capacity to process a variety of multimodal sources of data and derive useful perspectives in an efficient and scalable manner. For example, ML algorithms have achieved plausible results when estimating Alzheimer's and Parkinson's disease by using longitudinal data, clinical features, and sensor-based signals.

One of the most important features of ML in disease classification is its resistance to high-dimensionality. Biomedical datasets tend to have hundreds of features, such as speech phonetics to genetic markers, and conventional models have been prone to overfitting or bad generalization. ML techniques like PCA to reduce dimensionality, and classifiers like RF and SVM, have proved better performance by focusing on the most informative features while eliminating noise and redundancy (Alvarez et. al, 2019).
In addition, deep learning models (a subcategory of ML), such as deep neural networks (DNNs) and expert structures

like TabTransformer or DenseNet, have surpassed traditional models in some clinical prediction problems by capturing spatial, temporal, or contextual relationships in data. These models can handle organized data such as in the form of tables, and unstructured, which include audios, images, videos and other non-conventional inputs, providing flexibility.

Overall, machine learning's flexibility, scalability, and predictability make it a perfect tool for constructing automated models to identify precursors of neurodegenerative diseases. Its combination with speech-based features is a promising step towards affordable, non-invasive, and low-cost diagnostic devices (Aguayo et. al, 2023).

3.1 Linear Regression

Linear regression is perhaps one of the simplest methods in statistical learning and predictive modeling. It is a type of supervised learning that tries to learn the association between a dependent response variable and one or more independent predictor variables. The idea is to fit a linear equation that best explains how the input variables contribute to the output, hence its value not just for prediction but also for understanding underlying patterns in the data.

Linear regression uses an equation of the form:

$$\boldsymbol{y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\boldsymbol{x}_1 + \boldsymbol{\beta}_2\boldsymbol{x}_2 + \ldots + \boldsymbol{\beta}_n\boldsymbol{x}_n + \boldsymbol{\varepsilon},$$

where y is the response or dependent variable (which must be numerical variable), x's are the input variables, β values are the coefficients, and ϵ is the error. For the model to work well, the errors must follow the normal distribution.

One of the key reasons linear regression is widely used is its interpretability. Unlike many complex models, linear regression provides direct insights into how each predictor variable affects the response. This makes it particularly valuable in early stages of analysis when researchers aim to understand which factors contribute most significantly to the outcome of interest. For instance, in medical data, it can be employed to determine which patient features or symptoms are most correlated with disease progression or onset.

Typically, linear regression is applied when input-output relationship is presumed to be linear and when data satisfies some assumptions based on statistics, for example, homoscedasticity (errors having constant variance) and normality of residuals. It works particularly well on fairly small datasets or where the number of features is small and under control. In more advanced or high-dimensional situations, linear regression can still be used as a comparison baseline or as a part of a larger pipeline that includes feature selection.

Practically, the approach is to estimate the coefficients that specify the linear relationship from training data. These coefficients are then interpreted to determine the strength and direction of each predictor's influence on the outcome. After constructing the model, its performance is assessed with standard measures like residual error, explained variance, or tests of statistical significance for the coefficients. If the model passes validation criteria, it can be used on new data for prediction or monitoring.

Although linear regression is never going to be the most useful method in high-dimensional or non-linear situations, it is a tried and trusted and easy-to-interpret method which forms a good building block for more sophisticated machine learning techniques. Applied to neurodegenerative disease diagnosis, it has been employed both as a sole tool and in conjunction with other tools for predicting disease risk or progression from tabular clinical and epidemiological information.

3.2 Logistic Regression

Logistic regression is a statistical method that is commonly used in medical studies to understand situations in which there are 2 possible outcomes (binary), as in the presence or absence of a disease. Previously discussed, linear regression is typically used for continuous outcomes, logistic regression is reserved for situations in which one wants to estimate the probability of a categorical outcome. To clarify, this does not imply that linear regression is used when predictors are continuous, instead, it is used when the outcome is continuous. An application of logistic regression, lies in understanding neurodegenerative diseases such as Alzheimer's and Parkinson's. Here, the likelihood of the disease occurring is estimated through understanding the impact of clinical, demographic or biological variables. This technique involves developing a model to see the impact of different clinical, demographic, and biological variables on the probability of disease occurrence.

The technique includes estimating the likelihood of an outcome based on the independent variables, which are either continuous, categorical, or both. Instead of modeling the probabilities themselves, logistic regression models the logarithm of the odds, commonly called the logit. Doing so enables the model to treat bounded probabilities of zero and one automatically, making predictions valid in all cases without respect to predictor values. The logit transformation allows logistic regression to model changes in the odds of an outcome linearly with respect to the predictors and so is especially well-suited to medical work where relative risks are frequently reported as odds ratios.

Logistic regression does not predict the probability of something happening directly. Instead, it predicts the log of the odds, written as:

$$log(p\ /\ (1\ -\ p)),$$

where, p is the probability of the outcome. This helps the model work with values that can go from very small to very large, which makes it easier to fit a straight-line equation. After making the prediction, the model can convert it back into a probability between 0 and 1.

One of the main strengths of logistic regression is that it can control for multiple covariates at once, allowing one to determine each predictor's independent contribution while holding others constant. Logistic regression is also applied in epidemiological research where more than one independent

variable may interact with or confound another. For instance, logistic regression can be applied to understand how features of speech, age, and scores on clinical tests collectively affect the likelihood of getting Alzheimer's disease.

The fitting of a logistic regression model would often start from choosing informative predictor variables by prior knowledge or data exploration. After selecting the variables, the model is estimated by calculating the coefficients that will maximize the likelihood of seeing the observed data. The coefficients are then shown in terms of odds ratios, which are the multiplicative change in the odds of the outcome for each one-unit change in the predictor, controlling for other variables. Practically, the coefficients' interpretation gives useful information on how a change in speech characteristics or cognitive scores would raise or lower the chances of a neurodegenerative diagnosis.

To evaluate a logistic regression model's performance, one must assess its goodness of fit and prediction ability. Methods like likelihood-ratio tests, classification tables, and measures of the area under the receiver operating characteristic curve (AUC) are often employed to determine how well the model can differentiate individuals with and without the disease. In addition, logistic regression models need to be thoroughly screened for problems like multicollinearity between predictors or for violations of assumptions about linearity of the logit.

Logistic regression has been popular not just for its interpretability but also because it is relatively easy and widely available in standard statistical packages. It also acts as a simple model in most medical studies and is a starting point to understand complicated relations in data. For academic research which deals with small datasets or sparse event rates, however, logistic regression needs to be used carefully in order to avoid over-fitting or unreasonable estimates. In such environments, penalized forms of logistic regression or more complex techniques will be required. For example, in small sample sizes or sparse data, techniques like Lasso and Ridge regression are sometimes used.

In summary, logistic regression is a cornerstone statistical technique for the modeling of binary responses in clinical research. Its strengths in quantifying the effect of predictors using odds ratios, in handling confounders, and generating results that are interpretable make it well-tuned to examining the multivariate etiology of Alzheimer's and Parkinson's diseases. Its application in this research offers a straightforward, statistically sound way of examining the interrelation of phonetic speech measures with neurodegenerative disease classification likelihood (LaValley 2008).

3.3 Decision Trees

Decision trees is a form of predictive model with applications varying across statistics, machine learning, computational biology amongst more. This method is used in problems of classification and regression. This method is made to address problems which require categorization of data into classes. Essentially, decision trees operate by posing a series of queries regarding the features of an item which partitions the data into successively homogenous subsets. Each of the tree's internal nodes corresponds to an individual decision made on the basis of a feature, and the tree continues by recursively dividing the dataset until it arrives at the terminal nodes, or leaves, which label the data points that fall within each with a predicted class label.

Perhaps the greatest strength of decision trees is their transparency and interpretability. They are able to model sophisticated decision-making processes in simple, hierarchical terms, and the resulting trees are easy to visualize and comprehend. Unlike models like neural networks or support vector machines, which are frequently considered "black boxes," decision trees enable researchers to follow exactly how a decision was reached. This makes them particularly well-suited to domains where model explainability is crucial (Kingsford et. al, 2008).

Decision tree construction is based on examining a collection of training data whose class labels are already determined. It starts at the root node by considering potential splits along every feature, selecting the one that most effectively divides the data based on certain standards. Typical measures applied in evaluating the quality of a split are the entropy and the Gini index, which are measures of the purity of a collection of data points. A split that results in high-purity subsets where all or most examples are from the same class is desirable. The process recurses, with each tree branch further splitting the data until a condition for stopping, like finding perfect classification or reaching a minimum number of samples per node.

As a tree is being built, overfitting, where the model becomes excessively specialized in the training data and does not generalize to new examples, should be avoided. To this end, techniques like pruning are employed to reduce the tree by discarding nodes that make little contribution to classification accuracy on novel data. Pruning may be done by testing subtrees' performance on an independent validation set and removing non-beneficial branches that do not enhance predictiveness. Some algorithms use minimum description length principles to trade model complexity against classification performance.

Decision trees are very flexible because they can accommodate datasets with both categorical and continuous attributes, as well as datasets that have missing data. They can also address multi-class classification problems and can be adapted to regression tasks by predicting continuous values rather than discrete classes. Decision trees, having been trained, are computationally cheap and can quickly classify new instances by merely traversing the tree based on the responses to the questions in each node.

New advancements in decision tree techniques have seen the development of ensemble methods, where predictions from several trees are aggregated to enhance accuracy and stability. Random forests, for instance, create an ensemble of decision trees through training each tree on a random portion of the data and attributes and then averaging their outputs. This helps in decreasing variance and enhancing generalization through the averaging of many dissimilar models' outputs. The other method, boosting, trains trees in succession, with

each successive tree weighing most heavily those examples misclassified by the previous tree. Blended boosted trees will frequently produce models of very high accuracy.

Aside from their real-world applications in medicine and bioinformatics, decision trees have a long history. Their origin lies in biological taxonomic systems, used to classify organisms, however decision trees have been subject to advancements in artificial, statistics, and other fields. Previously used models such as Automatic Interaction Detection (AID) and later on such as ID3, C4.5 formed the foundation of modern decision trees methods. Today, decision trees and their versions are widely used, and address tasks which require this (De Ville, 2013).

3.4 Random Forest

Random Forest (RF) is a form of learning method. Random Forest's usage lies in problems concerning classification and regression. RF is used by compiling many decision trees during training and making predictions. The key concept behind Random Forests is the process of assembling the predictive power of many decision trees to enhance generalization and accuracy, especially on difficult and high-dimensional data.

The process operates by creating many decision trees, each trained on a bootstrap sample drawn from the original dataset. In addition, at every decision node, RF evaluates a random subset of input variables instead of all the variables, which introduces extra randomness into the model and minimizes the correlation between trees. The outputs of all the trees that have been grown are then combined through majority voting or averaging to generate the final prediction. This method provides robustness against overfitting, especially in big datasets having lots of variables.

One of the major advantages of RF is that it can manage datasets in which the number of variables is greater than the number of observations. RF is easily scalable with large datasets and provides flexibility for a large variety of learning tasks. Furthermore, Random Forests have the built-in capability of providing feature importance measures by gauging the impact of each variable on the predictive ability. Such features make RF, as a method, an effective one, when there are high dimensions, i.e., a large number of variables.

The theoretical basis of Random Forests shows that the larger the number of trees, the generalization error of the model stabilizes, and overfitting risk decreases. The generalization performance is controlled by two aspects, which are the individual classifiers' strength and their correlation among themselves. Smaller correlations and stronger classifiers in general provide better ensemble performance.

Although RF has robust predictive power, it is also somewhat opaque and has often been called a "black-box" model. Some developments, like variable importance scores and partial dependence plots, have enhanced interpretability. Out-of-bag error estimates where every observation is predicted by trees that did not use it in their bootstrap sample, offer an internal and unbiased measure of model performance without the need for a holdout validation set (Cutler et. al 2001).

3.5 Support Vector Machines (SVMs)

Support Vector Machines (SVM) are learning algorithms that are mainly used for classification, though they can also be applied in regression problems. The main goal of an SVM is to form a decision boundary, called a hyperplane, which best differentiates points belonging to various classes. This hyperplane is chosen on the basis of the concept of maximum margin, i.e., the distance between the hyperplane and the closest data points of each class, i.e., support vectors. A higher margin typically reflects improved generalization performance on novel data.

The logic behind SVM is that it can transform a problem of classification into an optimization one. By increasing the margin under the condition that training data points are correctly classified, SVM gets strong class separation. The parameters of the model are tuned to determine the hyperplane that meets this condition of optimality. When it comes to linearly separable datasets, SVM identifies the hyperplane that maximally separates the classes. Real-world data, however, is not always differentiable. In these situations, SVM uses slack variables, which permit some misclassifications so that the model can trade between maximizing the margin and minimizing classification mistakes.

SVM's flexibility is further achieved through employing kernel functions, which are generally known as the "kernel trick." Through this approach, it is possible for the algorithm to be used in a mapped feature space without knowing the coordinates of the data in the space. By substituting dot products with kernel functions, SVM is able to effectively deal with complicated, non-linear relationships. Linear, polynomial, and radial basis function (RBF) kernels are common ones and each of them can capture different kinds of data patterns. With this mechanism, SVM can have high accuracy even when the decision boundary in the original feature space is non-linear.

SVM is trained to solve a convex quadratic optimization problem. This ensures that the discovered solution is globally optimal, as compared to other machine learning algorithms that are likely to be stuck, and unable to do this. For large data sets, there are specific algorithms like Sequential Minimal Optimization (SMO) that are utilized to provide an efficient solution to the optimization problem by dividing it into smaller subproblems that are simple to handle.

Another benefit of SVM is that it is interpretable. The decision made by the model is dictated by a comparatively small number of the training points, the support vectors, so the decision boundary is more easily comprehensible. Furthermore, SVM has an easy geometric interpretation of classification, with it highlighting the margins and the support vectors as playing key roles in its predictive model.

SVM has performed well in different areas, especially in high-accuracy and high-robust tasks like image recognition, text classification, and bioinformatics. Its ability to deal with both linear and non-linearly separable data, as well as its strong mathematical base, makes it a trusted instrument in

machine learning studies, and their applications (Hearst et. al 2014).

4. AI model for Parkinson's disease

In this section we are interested in building an AI model for Parkinson's disease. To do that, we used a dataset that contains information of the voice quality of people that are Parkinson's patients and also people that are healthy.

The data set consists of 195 voice recordings, 147 belonging to Parkinson's disease patients and 48 belonging to healthy people. Each recording has 22 features that measure different aspects of speech production, e.g., pitch, jitter, shimmer, noise ratios, and nonlinear dynamics. These readings correspond to the stability, clarity, and variability of the speaker's voice, all qualities which are generally affected in Parkinson's disease. The data were free of missing values

Pitch-related parameters include MDVP.Fo.Hz. (mean fundamental frequency), MDVP.Fhi.Hz. (upper pitch), and MDVP.Flo.Hz. (lower pitch). These have mean values of about 154.2 Hz, 197.1 Hz, and 104.3 Hz, respectively. The voice frequency ranges from a low of 65.48 Hz to a high of 592.03 Hz, with great inter-individual variation. The large frequency range indicates how Parkinson's can affect control of the voice.

Jitter-related characteristics such as MDVP.Jitter(%), MDVP.Jitter.Abs., MDVP.RAP, MDVP.PPQ, and Jitter.DDP quantify the pitch variability. For example, MDVP.Jitter(%) had an average of 0.0062%, which indicates slight but quantifiable frequency deviations. Such minor irregularities in vocal fold vibration are typical of Parkinson's patients and reflect tremor-like symptoms.

Equivalently, shimmer measures like MDVP.Shimmer, MDVP.Shimmer(dB), Shimmer.APQ3, Shimmer.APQ5, Shimmer.DDA, and MDVP.APQ record fluctuations in vocal amplitude. Their means are between 0.0279 and 0.0374, with MDVP.Shimmer's mean at 0.0274. Large shimmer values indicate lower vocal steadiness, another hallmark of Parkinsonian speech.

Clarion of voice is measured in terms of NHR (mean = 0.0227) and HNR (mean = 21.89). Lower HNR and increased NHR refer to noisier, less harmonious voice signals, which are typically seen in Parkinson's.

Higher-order nonlinear dynamics including RPDE (average = 0.509), DFA (average = 0.720), spread1, spread2, D2, and PPE also define the complexity and randomness of speech. These features capture patterns undetectable by human perception. D2 is of particular interest with a median value of 2.36, which reflects the dimensionality of the chaotic vocal system for PD.

From the full data in hand, 70% of the data have been used for training, and the other 30% for testing. A logistic regression has been used to build a good predicting model, and a confusion matrix was created. In Table 1, we provide a summary of the coefficients from the logistic regression model.

Table 1: Logistic regression model summary for the Parkinson's disease analysis.

Term	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-80900	38900	-2.079	0.038
MDVP:Fo(Hz)	34.98	17.29	2.023	0.043
MDVP:Fhi(Hz)	2.388	1.288	1.854	0.064
MDVP:Flo(Hz)	22.45	11.58	1.939	0.053
MDVP:Jitter(%)	1838	932.7	1.97	0.049
MDVP:Jitter(Abs)	-1774000	894900	-1.982	0.047
MDVP:RAP	-34380	17290	-1.989	0.047
MDVP:PPQ	-13390	6832	-1.961	0.05
Jitter:DDP	5514	2786	1.979	0.048
MDVP:Shimmer	25.75	13.01	1.98	0.048
MDVP:Shimmer(dB)	2.008	1.012	1.984	0.047
Shimmer:APQ3	1069	539.1	1.983	0.047
Shimmer:APQ5	-25.72	13.17	-1.953	0.051
MDVP:APQ	35.76	18.32	1.952	0.051
Shimmer:DDA	-3206	1611	-1.989	0.047
NHR	107.6	53.64	2.005	0.045
HNR	0.1911	0.09373	2.038	0.042
RPDE	11.55	5.803	1.99	0.047
DFA	-14.19	7.366	-1.927	0.054

The aim is for the model to accurately predict individuals on the basis of voice features. Although none of the features proved statistically significant at the common 5% level of significance, MDVP.APQ, a shimmer measure, showed a p-value of about 0.09, suggesting its potential marginal significance. This accords with clinical findings that vocal fold instability and tremor, expressed as shimmer, are central biomarkers of Parkinson's.

The following data suggests how the model performed

1) Accuracy: 81% of overall accuracy of the predictions.
2) Sensitivity: 69.2% of healthy people correctly identified.
3) Specificity: 84.4% of Parkinson's cases were correctly diagnosed.
4) Kappa Score: 0.50 moderate agreement over chance.
5) Balanced Accuracy: 76.8%, the average efficiency over both classes, useful for cases when the classes are imbalanced.

Additionally, Principal Component Analysis (PCA) was applied to visualize the 2 components as a scatter plot. PCA is a dimensionality reduction technique based on statistics that keeps as much variance as possible from the original data. When applied in machine learning and data analysis, particularly with neurodegenerative disorders, PCA is very essential in extracting important features from complicated and high-dimensional data such as audio recordings or biomarkers. PCA identifies directions of maximum variance, called principal components. These components are linear combinations of the original variables and are orthogonal to one another, with no redundancy in the lower-dimensional representation.

The primary motivation for the use of PCA is data visualization simplification, increased computational efficiency, and elimination of noise or redundancy which can interrupt learning algorithms. In the case of speech or biomarker data in which thousands of variables may be collected per patient but the number of samples is not large, PCA can be used to find a smaller number of variables that retain the underlying structure of the data. This is particularly

useful for medical data, where missing data, correlations, and variance differences are prevalent.

PCA finds most application when data are high-dimensional and multicollinear, which is common in biological and medical environments. PCA allows machine learning algorithms to concentrate on the most descriptive features by projecting data into a lower-dimensional subspace, enhancing classification or prediction accuracy. In this project, PCA is applied as a feature selection and dimension reduction method prior to classification with methodologies such as Random Forests or Support Vector Machines. The technique provides a means of extracting the most useful phonetic features that differentiate between speech patterns of Alzheimer's, Parkinson's, or healthy controls.

PCA includes standardizing the dataset to ensure that all the features are equally vital. This is done by calculating the covariance matrix to ensure the features interact. Then the eigenvectors and eigenvalues are obtained from the matrix to identify the principal components and their significance. This data is transformed into a coordinate system through these components. Hence PCA is able to provide a better, concise, comprehensive representation of the data. It is an important processing technique in AI-powered diagnostic systems.

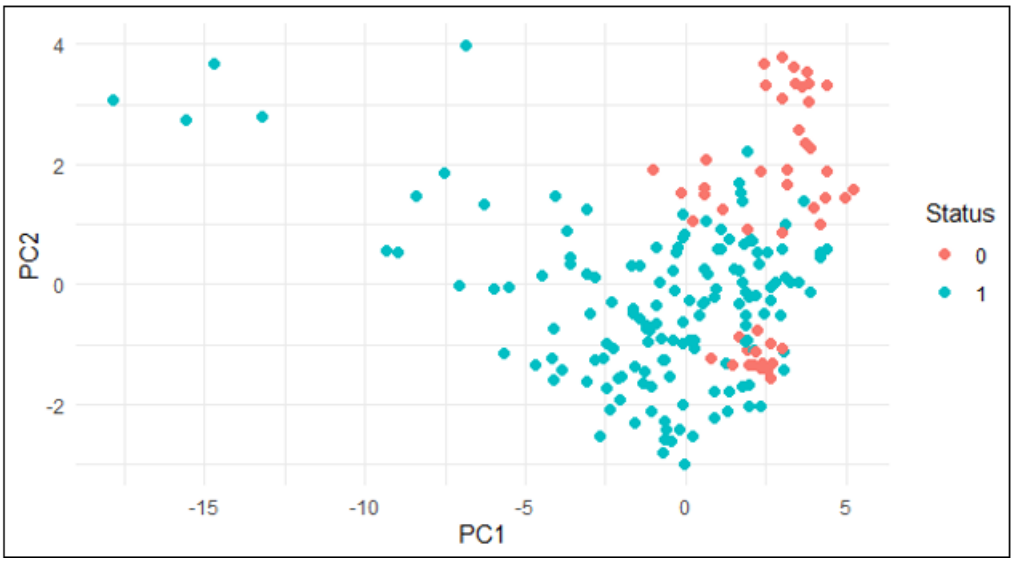

Figure 1: Principal Components of the PD dataset showing the 195 samples.

The scatter plot is created through PCA performed on the dataset of the 195 samples explained above. PCA transformed the original high-dimensional data into a new coordinate frame with principal components. In Figure 1, one point corresponds to one voice sample with red points for healthy people (status = 0) and teal points for people who have Parkinson's disease (status = 1).

In the x-axis (PC1) we have the significant principal component that takes into account the largest variance in the data. In the y-axis (PC2) is the second most significant principal component, accounting for the second highest variance following PC1. As we graph, we are able to identify various patterns and groups emerge.

The points are distributed throughout the plot created. There are visible groups, patterns and overlapping points. The areas towards the right have a larger density of blue points. A few of the remaining blue points are distributed to the left side. There are more red points towards the upper-right corner of the plot, with the other few in the lower-right corner. Despite the pattern emerging only with 2 dimensions, the distribution of color gives an idea of how the original features work together. There is some separation between the two classes, especially along PC1, but still there is some overlap, especially for the PC2 points, i.e., the red points. This suggests voice features carry some useful information for separating Parkinson's patients against the healthy controls, but PCA alone cannot fully separate the two.

5. AI model for Alzheimer's Disease

In order to build an AI model for Alzheimer's disease, we used a dataset that contains information of the voice quality of Alzheimer's patients and also people that are healthy. As in Section 4, 70% of the data has been used for training, and the remaining 30% of the data has been used for testing.

This Alzheimer's dataset has health and cognitive data for a total of 2149 patients. The AI model we are building in this section of the paper aims to accurately predict Alzheimer's disease. The participants are between the ages of 60 and 90, with a mean age of approximately 75 years, representing an older population in which the risk of Alzheimer's is most prominent. Gender split seems fairly even, with educational attainment differing across four intervals (0 to 3), having a median of 1, indicating that the majority of participants had basic to moderate formal schooling, a well-documented predictor of cognitive reserve and resistance to dementia.

Physical health markers consist of body mass index (BMI), which ranges from 15.0 to almost 40.0, and averages close to 27.7, putting most in the overweight range. Together with lifestyle measures, such as diet quality (mean: 7.1), physical activity (mean: 7.0), and sleep quality (mean: 6.4), these measures represent moderately healthy behavioral tendencies, albeit with considerable variation that can impact cognitive aging trajectories.

The data includes comprehensive cholesterol and blood pressure measures. The level of total cholesterol varies from 90 to 339.9 mg/dL, with an average close to 202, along with a mean HDL of 52.9 and LDL of approximately 120. Triglyceride levels also fluctuate considerably (mean: 121), with some having increased cardiovascular risk. The cholesterol-to-HDL ratio, a significant indicator of metabolic well-being, has a mean of around 4.1. Blood pressure readings further support this pattern: systolic pressure averages around 134, while diastolic centers near 85, values that may hint at early-stage hypertension in a significant subgroup.

Cognitive function and performance are captured through scores on various screening instruments. Measures such as CognitiveTest1 (mean: 0.28), CognitiveTest3 (~0.51), and MMSE (Mini-Mental State Examination, mean: 24.4) indicate mild to moderate cognitive impairment in the sample. Behavioral warning signs like memory complaints, confusion, disorientation, and personality changes are also observed in the form of binary flags, as well as functionality ratings related to activities of daily living (ADL) and general functional assessments.

The medical history variables are highly heterogeneous: hypertension, diabetes, depression, and cardiovascular disease are all coded as binary fields with high prevalence throughout the population. Family history of Alzheimer's is also monitored as a genetic risk proxy. Lifestyle risk indicators such as alcohol consumption, smoking status, and sensory deficits (e.g., hearing or vision impairment) add to the predictive picture.

Random Forest has been used as a method to build a strong predictive model. In total, 500 trees were used in the Random Forest model. Decision trees in particular were discussed in detail in Section 3.4.

According to the confusion matrix generated in the R software, the Random Forest classifier model built to predict Alzheimer's disease outcomes has a very high performance. From a test set, the model identified 407 as not having Alzheimer's (true positives) and 192 as having Alzheimer's (true negatives), having misclassified 45 cases in total. Specifically, 9 false positives and 36 false negatives. The model is 92.39% accurate overall, meaning it was correct on 92 out of every 100 occasions. This degree of accuracy is particularly impressive for a clinical prediction task, where getting a diagnosis wrong can result in potentially harmful consequences.

Notable here is the 97.83% sensitivity, which indicates the performance of the model in accurately classifying healthy subjects. Its high value indicates that the classifier is very good at recognizing non-Alzheimer's subjects, perhaps due to features such as Functional Assessment and ADL scores that measure independent functioning being more distinct in the healthy subjects. Contrarily, the specificity is a little less at 82.46%, which implies approximately 17.54% of the true Alzheimer's patients were overlooked. This deficiency appears in the 36 false negatives, patients who in reality had Alzheimer's but were forecasted to be healthy. Though in ratio the number is small, such mistakes can be life-altering in medical settings and must be kept to a minimum.

The Kappa metric of 0.828 reaffirms that there is high agreement between predicted and actual results and is not coincidental. Indeed, the P-value for accuracy > No Information Rate (NIR) is < 2.2e-16, which clearly suggests that this model performs much better than random guess.

Furthermore, the balanced accuracy of 90.15% indicates that the model performs well even on an imbalanced dataset, where 64.6% of the people did not have Alzheimer's (as indicated by the prevalence statistic). Balanced accuracy is particularly valuable here since the 'healthy' class is dominant, and a naïve model might have simply predicted all entries as healthy and still had 64.6% accuracy.

The confusion matrix also validates previous findings in the summary. Dimensions such as MMSE scores (mean ~14.8), FunctionalAssessment (~5.1), and BehavioralProblems (~0.16) probably assisted the model in distinguishing well. For instance, participants who were high on MMSE and low on behavioral issues were placed more consistently in the healthy class. The Detection Rate (63.2%) and Detection Prevalence (69.4%) also indicate that the model slightly overestimates the number of healthy people, but not by a disturbing margin.

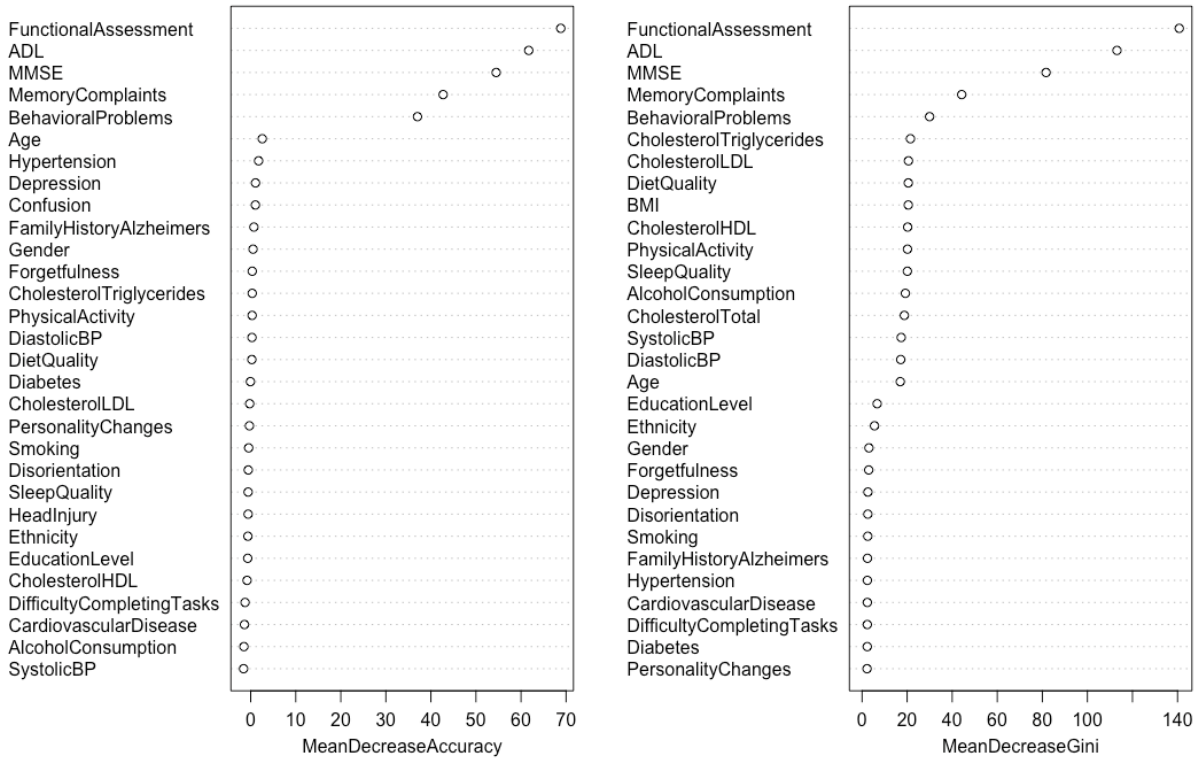

Figure 2: Important Features of Alzheimer's Dataset using Random Forest

Figure 2 shows the relative ranking of each feature utilized in the Random Forest model to classify Alzheimer's disease. Two measures are plotted: Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG), both of which measure how much each feature contributes to the model's accuracy. Out of all the features, FunctionalAssessment is the most impactful with the highest values in both MDA (close to 70) and MDG (over 140). Not unexpectedly, as it measures one's capacity to carry out daily activities, something drastically hits patients with Alzheimer's. It is closely trailed by the MMSE (Mini-Mental State Examination), a formal test designed for cognitive impairment screening. The model sees this as an effective diagnostic tool in how short-term memory, orientation, and logical reasoning, areas assessed by MMSE, can be employed to distinguish between patients effectively.

MemoryComplaints and BehavioralProblems come out high in importance as well. These are subjective experiences and overt behaviors respectively, corroborating that both patient self-report of symptoms and behavioral indicators are crucial hints for machine learning algorithms. Remarkably, PatientID comes out extremely high on both scales of importance. Although this is frightening, it probably indicates data leakage, maybe the ID contains the sequence or grouping that suggests diagnosis. Such a variable needs to be omitted from the final models in order not to overfit. Among the following level of features, biomedical measurements such as Triglycerides, HDL, LDL, and Total Cholesterol indicate that cardiovascular health has a quantifiable impact on Alzheimer's risk or symptoms. The findings are consistent with medical studies associating vascular injury with cognitive impairment.

Lifestyle factors such as PhysicalActivity, SleepQuality, and DietQuality are moderately significant, reflecting that habits of daily life are helpful supportive variables, but not necessarily first-order predictors. Demographic variables such as Age, Gender, and EducationLevel are marginal in importance here, even though these have been traditionally important in studies of population levels. Depression, Confusion, and Disorientation rank lower, which could be because these duplicate effects with higher-ranked cognitive measures.

6. Conclusion

This academic research set out to explore how to use machine learning methods, to classify 2 neurodegenerative diseases: Alzheimer's (AD) and Parkinson's (PD) based on their speech patterns. Due to the increase in prevalence of these diseases, in the aging population, the research paper began by describing their profiles, and previous attempted methods to use diagnostic approaches which were shortcomings, and the increasing trend of using speech as a biomarker. Specifically, it highlighted the impact of cognitive and motor decline in these diseases on speech production with resulting recognizable patterns of phonetics. It followed this by undertaking a thorough literature review of the current diagnostic methods using machine learning. This covered the emphasis on dimensionality reduction methods such as Principal Component Analysis (PCA) and classification algorithms like Random Forest (RF) and Support Vector Machine (SVM) that have proven promising to handle high-dimensional speech data. The next section delves into the theory; foundations of machine learning approaches. These were explained in detail, and were focused more on the mathematical aspects, how they can be used, and the

relevance in this field. Together these methods can be used for further development and understanding of AI-based models to classify neurodegenerative diseases based on speech patterns.

For creating an AI-based model to detect Parkinson's, the logistic regression model was used. The dataset had 195 voice samples (147 were Parkinson's patients, and the other 48 were control). The model was trained with 70% of the data, and the other 30% to test it. In order to have more accuracy, PCA was used to reduce dimensionality of other variables, which could have impacted the performance. The model was 81% accurate, had 84.4% specificity, and 76.8% balanced accuracy. No feature was found to be 5% significant, but variables relating to shimmer were found to have potential predictive impact.

In order to detect Alzheimer's disease through the AI-based model, Random Forest was used. It had 500 trees, which were trained through 70% of the dataset, and the other 30% to test it. The database had 2,149 participants, and considered other variables such as health, cognitive and lifestyle characteristics. The model had an accuracy of 92.4%, 97.8% sensitivity, and 82.5% specificity. The balanced accuracy was 90.2%. The kappa score was 0.828, which implies it was a strong agreement. Key predictive variables involved Functional Assessment, MMSE scores, and Behavioral Problems, whereas the existence of Patient ID among the top features suggested data leakage.

Out of the given 2 models, the Alzheimer's one is evidently better than the Parkinson's one. This is due to various reasons, and through methodological and data-driven improvements. Firstly, the Alzheimer's dataset is much larger than the Parkinson's database. The former has 2,149 participants, compared to a meager 197 in the latter. That means the larger dataset can provide a larger variability to learn from and a lesser chance of overfitting. The Alzheimer's model also had a larger range of characteristics which didn't just consider speech, rather also clinical, lifestyle and cognitive variables such as cholesterol level, eating habits, exercise amongst more. This is an example of multi-dimensional information which allows the Random Forest to identify better patterns, and improve its performance.

On the other hand, the Parkinson's model simply depended on speech samples, which are important, but reduce the extent to which it can predict, and can make it difficult for the algorithm to distinguish in a better manner. This is seen in the performance also. The Alzheimer's model had an accuracy of 92.4%, compared to 83.3% (Parkinson's). The kappa score was 0.828 for the former, compared to 0.662 for the latter. This shows a reliable agreement and prediction. Similarly, sensitivity and specificity were also larger for the Alzheimer's model, which allowed it to verify its ability to categorize the individuals. Thus due to a larger dataset, a better feature space, makes the Alzheimer's model much better due to the machine learning method used (Random Forest vs logistic regression) and also the quality of the database.

This academic research is intended to be extended in the future. Through the model developed, it can be integrated into simple video games. In various palliative care centers available, through setting up a device, which enables the inhabitants to play the video game, information received such as their performance, time spent, and speech pattern, this model can produce real-time data and information. The data can be reviewed by the caregivers and other important authorities.

References

[1] Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, *2*(4), 433-459.
[2] Aguayo, G. A., Zhang, L., Vaillant, M., Ngari, M., Perquin, M., Moran, V., ... & Fagherazzi, G. (2023). Machine learning for predicting neurodegenerative diseases in the general older population: a cohort study. *BMC medical research methodology*, *23*(1), 8.
[3] Álvarez, J. D., Matias-Guiu, J. A., Cabrera-Martín, M. N., Risco-Martín, J. L., & Ayala, J. L. (2019). An application of machine learning with feature selection to improve diagnosis and classification of neurodegenerative disorders. *BMC bioinformatics*, *20*, 1-12.
[4] Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. *Ensemble machine learning: Methods and applications*, 157-175.
[5] De Ville, B. (2013). Decision trees. *Wiley Interdisciplinary Reviews: Computational Statistics*, *5*(6), 448-455.
[6] Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their applications*, *13*(4), 18-28.
[7] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). Linear regression. In *An introduction to statistical learning: With applications in python* (pp. 69-134). Cham: Springer international publishing.
[8] Kingsford, C., & Salzberg, S. L. (2008). What are decision trees?. *Nature biotechnology*, *26*(9), 1011-1013.
[9] Lin, C. H., Chiu, S. I., Chen, T. F., Jang, J. S. R., & Chiu, M. J. (2020). Classifications of neurodegenerative disorders using a multiplex blood biomarkers-based machine learning model. *International Journal of Molecular Sciences*, *21*(18), 6914.
[10] Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*(1), 381-386.
[11] Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M., Neil, D., Saffari, A., Mead, R., ... & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature reviews neurology*, *16*(8), 440-456.
[12] IBM. (n.d.). *Random forest*. https://www.ibm.com/think/topics/random-forest
[13] IBM. (n.d.). *Support vector machine*. https://www.ibm.com/think/topics/support-vector-machine
[14] LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395-2399.
[15] Nick, T. G., & Campbell, K. M. (2007). Logistic regression. *Topics in biostatistics*, 273-301.
[16] Singh, G., Vadera, M., Samavedham, L., & Lim, E. C. H. (2016). Machine learning-based framework for

multi-class diagnosis of neurodegenerative diseases: a study on Parkinson's disease. *IFAC-PapersOnLine*, *49*(7), 990-995.

[17] Tagaris, A., Kollias, D., Stafylopatis, A., Tagaris, G., & Kollias, S. (2018). Machine learning for neurodegenerative disorder diagnosis,survey of practices and launch of benchmark dataset. *International Journal on Artificial Intelligence Tools*, *27*(03), 1850011.

[18] Tăuţan, A. M., Ionescu, B., & Santarnecchi, E. (2021). Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. *Artificial intelligence in medicine*, *117*, 102081.

Author Profile

Arkoneil Ghosh is a 15-year-old student, based in Mumbai, India. He is interested in studying computational neuroscience, and understanding how neurodegenerative diseases can be diagnosed early on with the help of artificial intelligence and machine learning. Arkoneil intends to major in Computer Science, and open his own gaming company someday which creates educational video games. These video games will be integrated with such AI-models to collect cognitive data from patients and assist in early detection and diagnosis of neurodegenerative diseases.