Disease Prediction ML Model

Team Members:

Yash Bawa- 21BCE1670

Rupkatha De- 21BCE6075

Khushi Agarwal- 21BCE1459

Problem Statement

Parkinson's disease (PD) is a widespread neurological condition that has a significant global impact on individuals. According to reports, PD affects an estimated 10 million people worldwide (0.3% of the total population), with those over 60 making up 1% of those affected.

- The most common neurodegenerative condition is Alzheimer's disease.
- Although the signs are first mild, they worsen over time. Among dementias, Alzheimer's disease is one of the most common.
- There is no cure for this disease, making it difficult to treat. The sickness is identified but only at an advanced stage. Therefore, if the condition is identified earlier, its course or symptoms may be slowed.
- We are proposing a system which can predict the Parkinson’s and Alzheimer’s disease based on the test records obtained from the clinic

Problem Objectives:

- The importance of this project that many people across the globe suffer with either of these diseases and getting to know about these diseases in the earlier stages might help them recover with some treatment. So, building a model which can accurately predict the condition of the patient is very crucial.

- The results of this project will be telling us if the person has Parkinson's disease or not and for the Alzheimer's case it will tell what kind of class that like Non-Demented, Very Mild Demented, Mild Demented, Moderately Demented.

- When it comes to the comparison of results with the existing method, we provided a better accuracy model for predicting the Parkinson's disease using different machine learning classifiers. We also used CNN for classifying the images for Alzheimer's which works better than SVM and Decision tree classifiers.

Dataset Description:

The **Alzheimer's Dataset** is in the form of Images , these images are classified into four different categories of Alzheimer Disease.
In the form of MRI images. The data has four classes of images both in training as well as a testing set:

- Mild Demented
- Moderate Demented
- Non Demented
- Very Mild Demented

DATASET LINK
https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images

- The **Parkinson's Disease dataset** comprises biomedical voice measurements from 31 individuals, including 23 with Parkinson's disease (PD).
- Each row represents a voice recording, with columns corresponding to specific voice measures. The "status" column indicates whether the individual is healthy (0) or has PD (1). There are approximately six recordings per patient, and the dataset aims to facilitate discrimination between healthy individuals and those with PD based on voice characteristics. The dataset's origin and purpose are detailed in a reference by Max A. Little et al. (2008), highlighting its relevance for telemonitoring Parkinson's disease.

  - DATASET LINK - https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set

Literature Review

[1] Parkinson's Disease Diagnosis Using Machine Learning and Voice

· In this research paper, they could diagnose and predict the Parkinson's disease through machine learning architectures using only non-invasive voice biomarkers as features.

Because of its underlying cognitive and neuromuscular function, biomarkers extracted from human voice can provide insight into neurological illnesses like Parkinson's disease (PD).

· Advantage of this research paper is that they used different machine learning algorithms

such as Decision tree, SVM, Random Forest etc on the given dataset so that even if one

of these models is not very good with the following dataset, the other models might

work well and predict the outcomes with a better accuracy.

· Disadvantages in this paper is that even though they used different machine learning models, they couldn't reach an accuracy score of 90 percent. The highest they gained was 86 percent and also the data used for these algorithm's performance was limited to a clinician. Few studies suggest that voice analysis is not that accurate to predict the

Parkinson's disease.

· The overcoming solution we have come up with to increase the accuracy of the model is, we have used many more other classifiers and models like KNN, Linear regression model, Naïve Bayes, which weren't used in the following results. Some of these showed a better accuracy than the rest for our dataset and also, we split our dataset into 90

training and 10 testing which gave the model more info to be trained and predict the outcomes more accurately.

Literature Review

- **[2] A Novel Method for Parkinson's Disease Diagnosis Utilizing Treatment Protocols**

- · In this research paper they collected the data from the University of Baghdad's College of Medicine's treatments Protocol Repository. They did survey on both female and male participants. In this they asked the patients to repeat the vowel 'a' thrice and collected their voice modulations and the frequency and took their previous medical conditions and also took their blood reports.
- · Advantage of this research paper is unlike other research papers the data they used is more diverse. They collected the data from patients of different genders, different ages groups and few control subjects who doesn't have Parkinson disease. They also used different machine learning algorithms to check whichever algorithm gives the better outputs.
- · Disadvantages in this paper is their data set is very small. They conducted a survey among only 253 participants and the they divided the dataset for both testing and training. The accuracy obtained from these models are low. The average accuracy they were receiving was 65 percentage.
- · The overcoming solution we have come up with to increase the accuracy of the model is, we have used the dataset which is as diverse as this dataset but also contains more data samples compared to this which helped our model to give a better accuracy.

Literature Review

[3] Diagnosis of Parkinson's disease using EEG and MRI

- In this research paper, they collected the fMRI and EEG data from the patients and used that to diagnose and predict the Parkinson's disease. Here the collected data is in the form of images. fMRI scans are of the scans of their brain and the EEG produces the readings of the electrical activity that is taking place in the brain.
- Advantages of this paper is that, they used the scans of brain and like many studies suggested that brain scans give more insights about the person's Parkinson condition than the data collected from that same person's vocal modulation and frequencies.
- Disadvantages of this paper is that, collecting brain scans survey is difficult because the scans are costlier than the voice recordings. Due to this reason the data available for this paper was less compared to the other papers.
- The overcoming solution we have come up for this is that we can just go with the voice frequency since we have enough data and even if we need some extra data, we can collect it from people around us just by taking their voice frequencies.

Literature Review

- **[4] Prediction of Alzheimer's disease using Machine learning Classifiers.**
- · In this research paper, they took the MRI scans of brain and used those images as their dataset to train the models which will help in detecting the Alzheimer's disease.
- · Advantage of this paper is that for detecting Alzheimer's the only way is to use MRI scans of the brain and for machine learning algorithms they used classifiers like Support Vector Machine and Bayesian Support Vector Machines. Based on the accuracy that each model is giving, they proceeded with the one giving the most accuracy.
- · Disadvantage of this paper is that they used SVM. SVM gives good results but for SVM they first had to create some features from the image and then they have to feed those features to the classifier and also there might be human errors while creating those features and there will be chances of losing spatial interaction between pixels during this process.
- · The overcoming solution we have come up for this is that we used CNN model instead of SVM. CNN can be taught of automatic feature extraction which will reduce the possible errors that can occur when we make our own features and feed the model.

Literature Review

[5] Alzheimer Disease Prediction using Machine Learning Algorithms

· In this research paper, they took the MRI scans of brain and used those images as their dataset to train the models which will help in detecting the Alzheimer's disease. · Advantage of this paper is that for detecting Alzheimer's the only way is to use MRI scans of the brain and for machine learning algorithms they used different classifiers like Support Vector Machine and Decision tree classifier. Based on the accuracy that each model is giving, they proceeded with the one giving the most accuracy.

· Disadvantage of this paper is that they used SVM and decision tree. Both SVM and Decision tree works very good and give better results when it comes to the numerical parameters but for these models they first had to create some features Page 7 from the image and then they have to feed those features to the classifier and also there might be human errors while creating those features and there will be chances of losing spatial interaction between pixels during this process.

· The overcoming solution we have come up for this is that we used CNN model instead of SVM or decision tree classifier. CNN can be taught of automatic feature extraction which will reduce the possible errors that can occur when we make our own features and feed the model.

Literature Review

[6] Deep learning in Alzheimer's disease: Diagnostic classification and prognostic prediction using Neuroimaging data

- 16 papers selected from 192 for Alzheimer's biomarker identification using multimodal neuroimaging data.
- Deep learning applied directly to original neuroimaging data, bypassing feature selection steps.
- No comparison of accuracy with alternative models.
- All major learning algorithms employed, with preference for highest accuracy model.

Literature Review

- **[7] Alzheimer's detection based on segmentation of an MRI image**
- · They have used the system of Magnetic Resonance Imaging (MRI) images and applying the model of wavelet transform to detect Alzheimer's.
- · Their proposed method involves applying wavelet transform on the input image and then using inverse wavelet transform. Next, they proceed to reconstruct the image using 3D reconstruction and then segment it. Based on this the decision is made.
- · Since wavelet transform has been used here, some of the serious disadvantages of this model are: shift sensitivity, poor directionality and lack of phase information.
- · Our model is based on choosing the best possible model by testing it with different learning methods and choosing the one with the best accuracy and hence there are no cases of overfitting or shift sensitivity.

Literature Review

- **[8] Alzheimer detection using Group Grey Wolf Optimization based features with convolution classifier**
- · This paper employed the text information for features' extraction in addition to the segmentation of brain images.
- · The model they have implemented enhanced the binary and multi0 class classification execution. The multi-class classification process remains the commitment of this work which is enabled with the chosen ideal features from Group Grey Wolf Optimization (GGWO) - motivated enhancement model. Page 8
- · Some of the critical flaws of this model are that it has a lack of accuracy, low population diversity, premature convergence and imbalance between the exploitation and exploration.
- · Our employed model overcomes flaws such as population diversity as our input data is voice recordings of patients and thus applying the model with the highest accuracy on this would yield the best possible results.

Literature Review

- **[9] <u>Detecting Parkinson's disease with sustained phonation and speech signals using machine learning techniques</u>**

- · The main contribution of this approach is evaluating the performance with the use of other classifiers and processing audio without fusing of feature sets.
- · Their proposed approach is based on motor symptoms explicitly related to voice. The voice examination begins with two vocal tasks namely - phonation and speech respectively. Based on the data they selected the best configuration for each classifier.
- · This model employed by them is expensive in terms of time since they have used varied parameters for their dataset which will not be a time efficient solution.
- · Our model provides a more time efficient solution since it focuses on the important parameters in the voice recordings that are necessary for the learning models in order to detect the disease.

Literature Review

- **[10] A deep learning approach for Parkinson's disease diagnosis from EEG signals**
- · Electroencephalogram (EEG) signals have been used as input data in this model since this is a disease related to brain abnormality. PD is characterized by gradual degradation of the motor function in the brain.
- · In this work, they have used EEG signals of twenty PD and twenty normal subjects. This model has achieved a promising performance of 88.25% accuracy, 84.71% sensitivity, and 91.77% specificity.
- · Since this model involves processing EEG signals as input data it is extremely expensive in terms of processing speeds and power. This model is allso not ready to be installed for clinical usage.
- · Our input data is a collection of thousands of MRI scans of the brain which is used by the test data later on to compare with. We have employed an automated detection system for PD using the convolutional neural network (CNN).

Literature Review

- **[11] Parkinson's progression prediction using machine learning and serum cytokines.**
- • This paper talks about how the model uses Serum Cytokines samples from patients with Parkinson's Disorder with and without the common leucine-rich repeat kinase 2 (LRRK2) G2019S mutation and determines the progression rate.
- • An advantage of this model is that it produces high accuracy results because it uses the serum samples. • But one drawback is collecting these samples is very difficult and sometimes expensive to collect.
- • In our project, we simply make use of the voice frequency values to determine the Parkinson's disorder and it is more easily accessible.

Literature Review

- **[12] Early-Stage Alzheimer's Disease Prediction Using Machine Learning Models.**

- • This paper discusses about how it identifies Alzheimer’s disease in early stages. Predictions of Alzheimer's disease are based on Open Access Series of Imaging Studies (OASIS) data, and performance is measured with parameters like Precision, Recall, Accuracy, and F1-score for ML models. These datasets are available on Kaggle. The proposed classification scheme can be used by clinicians to make diagnoses of these diseases.
- • An advantage of this is that it uses high accuracy Machine Learning models and is very efficient.
- • Our project uses MRI images dataset to determine the Alzheimer’s disease.

Literature Review

- **[13] Parkinson's disease prognostic scores for progression of cognitive decline.**
- • This model uses analysis of 19 baseline clinical, pathological and demographic variables.
- • An advantage of this model is that they employ a varied set of parameters for this data.
- • A drawback, since it has so many parameters, it is computationally expensive and leads to overfitting.
- • Our model is more efficient because it uses MRI images which are classified and easy to train.

Literature Review

- **[14] A Comparative Analysis of Machine Learning Algorithms to Predict Alzheimer's Disease**
- • This paper uses the OASIS dataset to determine the Alzheimer's disease.
- • An advantage of this is that it uses different attributes and gets more accurate results.
- • Our project uses MRI images to determine Alzheimer's disease

Literature Review

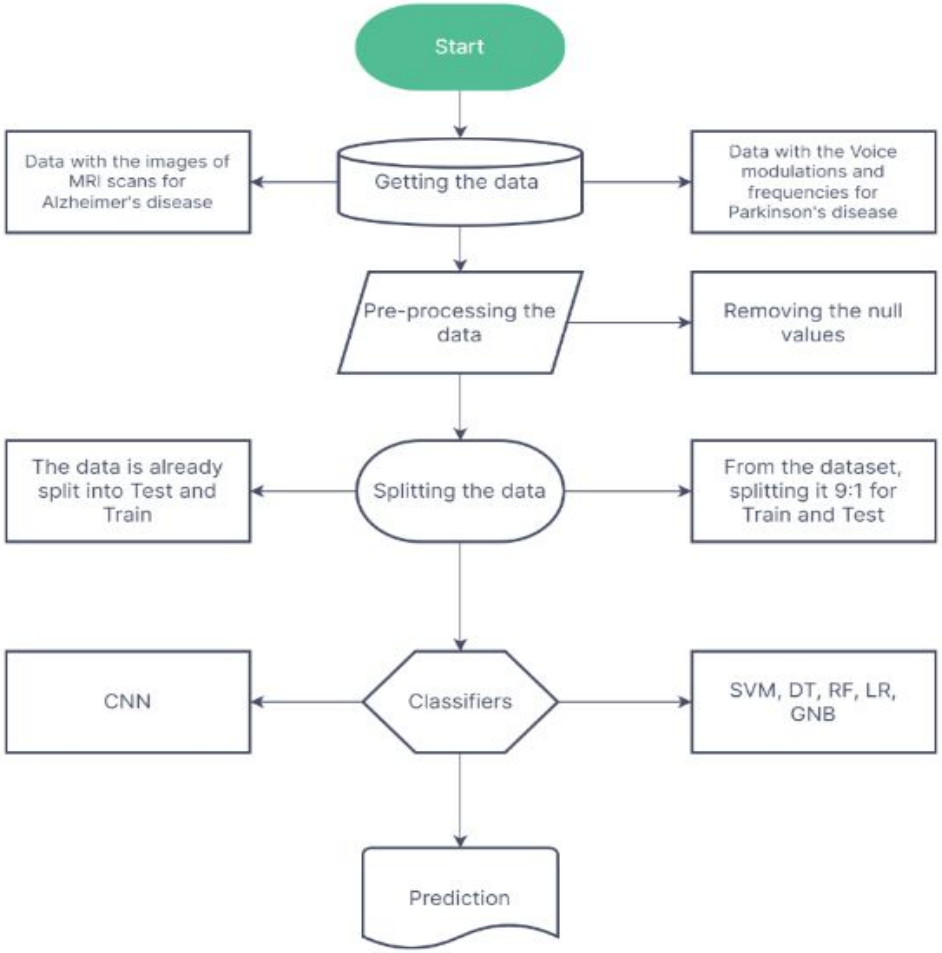
- **[15] Optimizing Machine Learning Methods to Improve Predictive Models of Alzheimer's Disease.**
- This paper discusses about investigating the accuracy of different ML methods and different features to classify cognitively normal (CN) individuals from Alzheimer's disease (AD) and to predict longitudinal outcome in participants with mild cognitive impairment (MCI).

Proposed Method

- For Parkinson's disease prediction, various ML classifiers were tested on voice modulation data from Kaggle, with KNN achieving 98% accuracy (K=3) and Decision Tree close behind at 97%, prompting the use of Decision Tree for predictions.

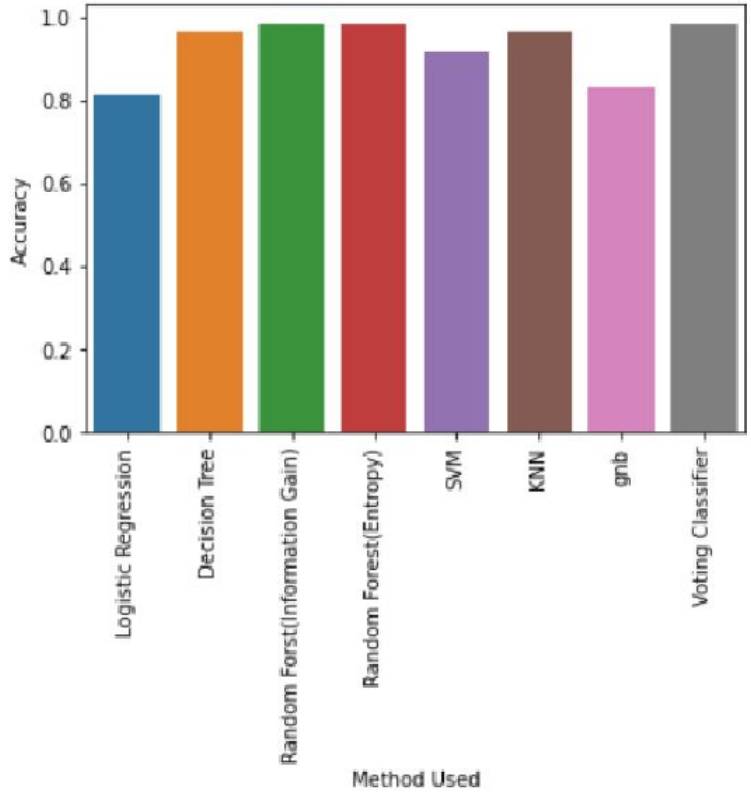
- Alzheimer's prediction utilized MRI scans from Kaggle, employing Convolutional Neural Networks due to their automated feature extraction capabilities, contrasting with SVM and Decision Trees requiring manual feature creation and risking loss of spatial interactions between pixels.

Proposed Method

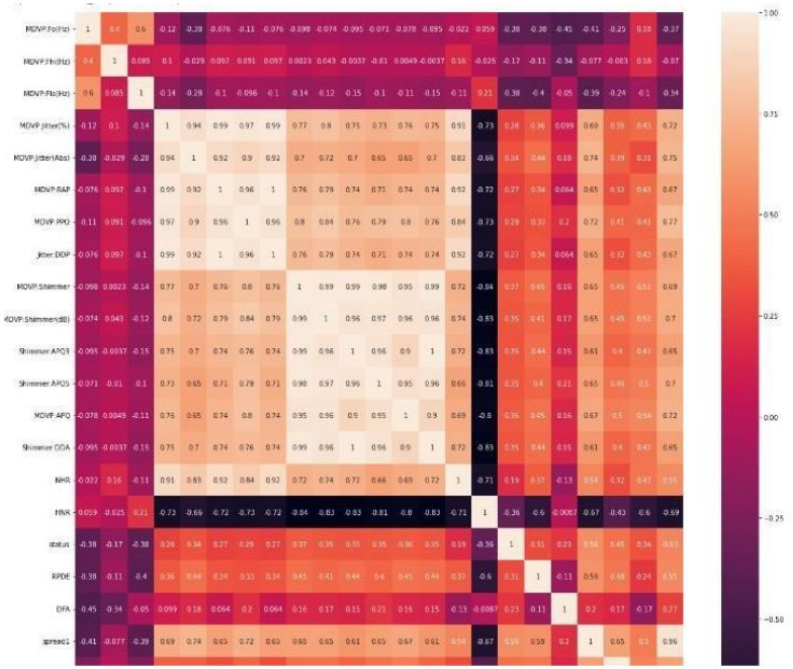


Comparative analysis

Our model



Corelation



Comparison with other Models

(a)

Data set / Performance criteria	Baseline data group			MFCC data group		
	Classifier algorithms					
	Decision tree (DT)	kNN	SVM	Decision tree (DT)	kNN	SVM
Accuracy rate (%)	71.35	68.75	66.67	69.79	75	73.44
Sensitivity	0.75	0.7	0.72	0.6	0.72	0.7
Specificity	0.68	0.68	0.61	0.79	0.79	00.78
F-measurement	0.72	0.69	0.66	0.69	0.75	0.73
Kappa	0.43	0.38	0.33	0.4	0.5	0.47
Area under the ROC curve	0.71	0.71	0.68	072	0.72	0.72

(b)

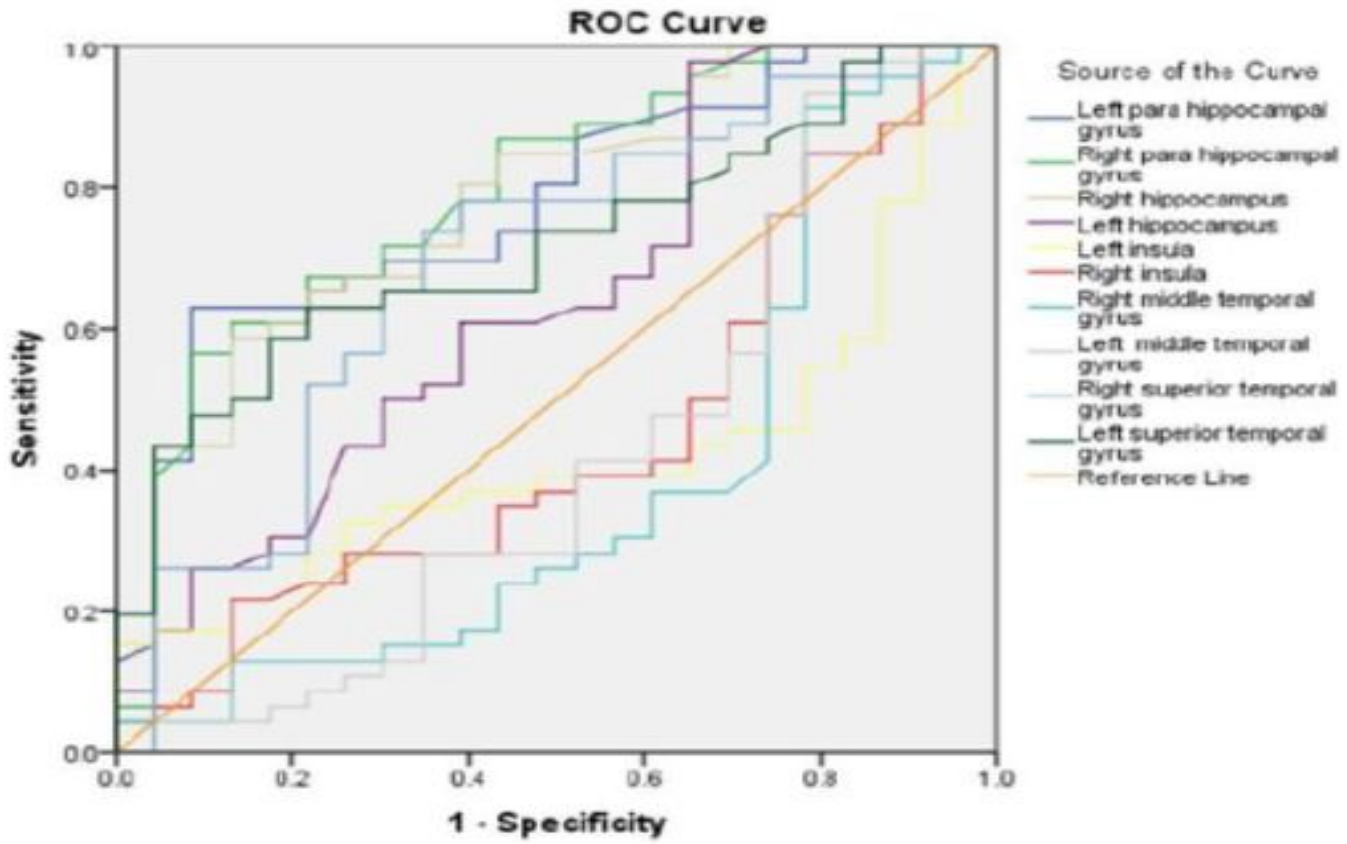
Data set / Performance criteria	Time data group			Vocal data group		
	Classifier algorithms					
	Decision tree (DT)	kNN	SVM	Decision tree (DT)	kNN	SVM
Accuracy rate (%)	66.15	67.71	72.4	61.98	64.06	63.02
Sensitivity	0.73	0.82	0.8	0.64	0.58	0.64
Specificity	0.59	0.53	0.65	0.6	0.7	0.63
F-measurement	0.65	0.65	0.72	0.62	0.64	0.63
Kappa	0.32	0.35	0.45	0.24	0.28	0.26
Area under the ROC curve	0.65	0.69	0.70	0.61	0.63	0.64

(c)

Data set / Performance criteria	Wavelet data group			Whole data group		
	Classifier algorithms					
	Decision tree (DT)	kNN	SVM	Decision tree (DT)	kNN	SVM
Accuracy rate (%)	68.23	73.44	79.69	69.79	76.56	85.42
Sensitivity	0.81	0.76	0.97	0.6	0.81	0.94
Specificity	0.55	0.72	0.63	0.79	0.72	00.78
F-measurement	0.66	0.73	0.76	0.69	0.76	0.86
Kappa	0.36	0.47	0.59	0.4	0.53	0.72
Area under the ROC curve	0.69	0.74	0.82	0.70	0.78	0.89

Comparison for Alzheimer's

Model from Literature survey



Summary of Key Findings and Issues

- **Potential applications**: Early disease detection enabling preventative measures, improved risk stratification for targeted interventions, personalized treatment plans, and potentially reduced healthcare costs.
- **Relevant research**: Machine learning shows promise in disease prediction, with studies showcasing successful applications in areas like memory loss and movement restriction risk assessment, early sepsis detection, and hospital readmission prediction.
- **Data privacy and security:** Ensuring patient data anonymization and adherence to regulations.
- **Model bias and fairness:** Addressing potential biases in algorithms and datasets that could lead to discriminatory outcomes.
- **Explainability and interpretability:** Making the model's predictions understandable to medical professionals and patients.

Highlighting the Gaps:

- **Limited access to high-quality healthcare data:** Need for broader data availability and collaboration with healthcare institutions.
- **Integration with existing healthcare systems:** Developing seamless integration for clinical adoption and workflow efficiency.
- **Long-term impact and cost-effectiveness:** Assessing the model's impact on patient outcomes and healthcare costs over time.
- **Addressing complexities:** Refining models for diseases with multiple contributing factors and intricate biological mechanisms.

Conclusion

1. Utilizing non-invasive vocal biomarkers, automated machine learning systems effectively diagnose and forecast Parkinson's disease, showcasing the potential of different classifiers in noisy, high-dimensional data for clinical-level precision through careful feature selection. This suggests promising avenues for leveraging voice data in assessing patient health and neurological conditions, with future prospects for denser feature sets including spoken word or video modalities.

2. Employing Convolutional Neural Networks on brain MRI data, our project achieves a 95% accuracy rate in distinguishing between Alzheimer's and healthy controls, offering practical applications for clinical use. This underscores the effectiveness of CNN classifiers in disease differentiation, highlighting their potential in advancing diagnostic accuracy.

Role of Each Team Member

- **Yash Bawa**
  - Data Collection
  - Model Training
- **Rupkatha De**
  - Model Development
  - Model Training
- **Khushi Agarwal**
  - Model Evaluation
  - Model Analysis

References

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8950225/
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8896926/
- https://www.kaggle.com/datasets/vikasukani/parkinsons-disease-data-set
- https://www.kaggle.com/datasets/tourist55/alzheimers-dataset-4-class-of-images

  https://www.researchgate.net/publication/328726100_Parkinson%27s_Disease_Diagnosis_Using_Machine_Learning_and_Voice

  https://www.researchgate.net/publication/362433153_A_Novel_Method_for_Parkinson's_Disease_Diagnosis_Utilizing_Treatment_Protocols

  https://www.researchgate.net/publication/364338000_Diagnosis_of_Parkinson's_disease_using_EEG_and_fMRI

  https://www.researchgate.net/publication/344766443_Prediction_of_Alzheimer%27s_Disease_Using_Machine_Learning_Classifiers
- https://ieeexplore.ieee.org/document/9074248

- "Deep learning in Alzheimer's disease: Diagnostic classification and prognostic prediction using Neuroimaging data" (Published: 2019) by Taeho Jo, Kwangsik Nho, Andrew J. Saykin
- "Alzheimer's detection based on segmentation of an MRI image" (Published: 2017) by K.S. Biju, S.S. Alfa, Kavya Lal, Alvia Antony, M. Kurup Akhil
- "Alzheimer detection using Group Grey Wolf Optimization based features with convolution classifier" (Published: 2019) by Shankar K, [illegible]hmanaprabu S.K., Ashis Khanna, Sudeep Tanwar, Joel J.P.C. Rodrigues, Nihar Ranjan Roy