EXAMINING ML ALGORITHMS FOR PARKINSON'S DETECTION THROUGH SPEECH DATASETS: A COMPARATIVE ANALYSIS

A dissertation
submitted in partial fulfilment of the requirement for the degree of

MASTER OF SCIENCE

in

BIOTECHNOLOGY

by

SHIKHA KADYAN
2K22/MSCBIO/46

Under the Supervision of
Prof. PRAVIR KUMAR

Department of Biotechnology

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

June, 2024

ACKNOWLEDGEMENT

At the time of submission of my M.Sc. Dissertation, I am grateful to the almighty God who has bestowed upon the wisdom, strength and patience to take up this endeavour. Apart from the effort, the success of this project depends largely on the encouragement and guidelines of many others. I, therefore, take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

My initial thank is addressed to my mentor Prof. Pravir Kumar, Department of Biotechnology, Delhi Technological University, who gave me this opportunity to work in a project under him. It was his enigmatic supervision, constant encouragement and expert guidance which enable me to complete this work. I humbly seize this opportunity to express my gratitude to him. Words are inadequate in offering thanks to Ms. Mehar Sahu and Mrs. Neetu Rani who like a family, shown confidence in me and helped me in my project.

I extend my thanks to technical staff Mr. Jitendra Singh and Mr. C.B Singh who have been an aid whenever required. Lastly, I wish to extend my thanks to my family and friends who have supported me through the entire process.

Shikha Kadyan
2K22/MSCBIO/46

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I, Shikha Kadyan, 2K22/MSCBIO/46 student of M.Sc. Biotechnology, hereby certify that the thesis entitled **"Examining ML algorithms for Parkinson's detection through speech datasets: A comparative analysis"** in partial fulfilment of the requirement for the award of the Degree of Master of Science submitted in the Department of Biotechnology, Delhi Technological University is an authentic record of my own work carried out during the period from May 2023 to May 2024 under the supervision of Prof. Pravir Kumar.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other institute.

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiner in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CERTIFICATE BY THE SUPERVISOR

Certified that **Ms. Shikha Kadyan**, 2K22/MSCBIO/46 has carried out their search work presented in this thesis entitled **"Examining ML algorithms for Parkinson's detection through speech datasets: A comparative analysis"** from Department of Biotechnology, Delhi Technological University, Delhi under my supervision. The thesis embodies result of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

05/06/2024
Prof. Pravir Kumar
Supervisor, Dean IA
Department of Biotechnology
Delhi Technological University

05.06.24
Prof. Yasha Hasija
Head of Department
Department of Biotechnology
Delhi Technological University

Date:

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi- 110042

PLAGIARISM VERIFICATION

Title of the thesis: Examining ML algorithms for Parkinson's detection through speech datasets: A comparative analysis.

Total Pages: 23

Name of the Scholar: Shikha Kadyan

Supervisor: Prof. Pravir Kumar

Department of Biotechnology

This is to report that the above thesis was scanned for similarity detection. The process and outcome are given below:

Software used: Turnitin **Similarity Index:** 4%

Total Word Count: 7855

Date: 05-06-2024

05/06/2024

Candidate's Signature **Signature of Supervisor**

Examining ML algorithms for Parkinson's detection through speech datasets: A comparative analysis

Shikha Kadyan

ABSTRACT

Parkinson's disease or PD, the most well-known neurological condition impacting the human neurological system, causes dopamine-producing neurons in the midbrain to degenerate. It is a primary concern to detect PD in its early stages to slow down its progress by engaging patients in early medical therapies and foster a better quality of life for them. Although new research appears to indicate that majority of the PD patients experience speech impairments in the early stages of the disease, the primary impacts of PD are on motor and cognitive function. Within the framework of this study, a number of machine learning (ML) models, including Principal Component Analysis (PCA), Random Forest (RF), Gaussian Naïve Bayes (GNB), K-Nearest Neighbours (KNN), Decision Tree (DT), Logistic Regression (LR), Extreme Gradient Boosting (XGB), and Support Vector Machine (SVM), have been comparatively analysed on two different speech datasets consisting multiple attributes, using three different approaches for classification. The models were assessed and evaluated, for their efficiency in PD classification, using different scoring metrics such as accuracy, precision, recall, and F1-score. Here, we discovered that the XGB and SVM models of the second approach—where the data was oversampled—were the most efficient models. XGB demonstrated 98.30% accuracy and 96.67% precision with Dataset 1 while SVM achieved 97.8% accuracy and 99.1% precision with Dataset 2. They also depicted maximum area under the curve for ROC curve, highlighting their capability to discriminate between true positives and true negatives. The highest degree of accuracy and precision in the early detection of PD has been rendered attainable by ML algorithms. When trained on an extensive set of data, these additionally possess the potential to offer 100% accuracy, or clinical-grade accuracy, through hyper-parameter optimisation.

LIST OF PUBLICATIONS

Title of the Paper: Analyzing speech patterns for Parkinson's diagnosis: Insights from ML models

Authors Name: Shikha Kadyan and Pravir Kumar

Name of the Conference: 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)-2024 organised by Galgotias College of Engineering & Technology (GCET)

Date of Conference: 14th and 15th May, 2024

Indexing: IEEE

Status of Paper: Accepted

Date of Acceptance: 23rd April, 2024

Date of Registration: 24th April, 2024

Date of Camera Ready Submission: 27th April, 2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

PD	Parkinson’s Disease
NDD	Neurodegenerative Disease
ML	Machine Learning
DL	Deep Learning
SVM	Support Vector Machine
LR	Logistic Regression
KNN	K-nearest Neighbors
DT	Decision Tree
RF	Random Forest
XGB	Extreme Gradient Boosting
GNB	Gaussian Naïve Bayes
PCA	Principal Component Analysis
ROC	Receiver Operating Characteristics
AUC	Area Under the Curve
AD	Alzheimer's Disease
ALS	Amyotrophic Lateral Sclerosis
HD	Huntington's Disease
SN	Substantia Nigra
PWP	Patient with Parkinson’s
SuStaIn	Subtype and Stage Inference
IoT	Internet of Things
AIoMT	Artificial Intelligence of Medical Things

CHAPTER 1

INTRODUCTION

A neurological condition that progresses over time, PD is typified by a broad spectrum of motor and non-motor symptoms. Tremors, rigidity of the muscles, bradykinesia, and unsteady posture constitute typical motor symptoms [1]. The principal molecular mechanisms of PD include misfolding and clumping of α-synuclein proteins; malfunctions in the energy-producing mitochondria of the cell; challenges in eliminating unwanted proteins owing to troubles with the ubiquitin-proteasome and autophagy-lysosomal systems; nervous system inflammation; and oxidative stress [2]. Interruptions in the pathways of neurotransmitters such as dopamine, adrenaline, adenosine, serotonin, and glutamate, further complicate the symptoms of PD [3], [4]. Early diagnosis is very crucial to hinder its progression and significantly improve PWP's lives. It additionally allows patients to engage in specialized and early treatment plans for enhanced results.

There is an urgent need to introduce novel technologies to ensure early and accurate diagnosis of PD. Keeping this in mind, AI emerges as a promising technology, with the capability of generating machines akin to human intelligence for detecting biological changes. This is achieved by collecting significant data from the patient and then comparing it with the already available large datasets for analysis, which allows healthcare professionals to make informed and accurate decisions. The intersection of healthcare and advanced technological applications has ushered in a new era in diagnosing and managing diseases, particularly in neurodegenerative disorders [5]. Patient classification as either healthy or Parkinson's can be done using ML models, an inexpensive, streamlined, reliable, and efficient approach.

Studies have shown that assessing voice abnormalities can act as a marker for early PD identification [6], [7]. Reduced intensity, pitch, and monotonous loudness, as well as decreased tension, tense silence, rapid speech bursts, erratic tempo, ambiguous consonant enunciation, and dysphonia, which is characterised by hoarse and whispering voices, are common speech impairments associated with PD [8], [9]. It has been estimated that in the earliest phases of the disease, voice and difficulties with speech impact 90% of Parkinson's patients [10]. Given that vocal cord abnormalities are very easy to quantify and can be evaluated remotely, it can be beneficial to identify and track these impairments early in PD.

Therefore, this study aims to investigate numerous ML models for the early detection of PD utilising different approaches and two speech datasets, Dataset 1 comprising 195 voice recordings and 22 variables, while Dataset 2 comprising 756 voice recordings and 754 variables . The findings show that the XGB and SVM models outperforms all other models in terms of performance accuracy, post-training on 22 characteristics using over-sampled data i.e., the second approach used in the methodology. XGB displayed an impressive accuracy of 98.30% and precision of 96.67% with Dataset 1, while SVM achieved 97.8% accuracy and 99.1% precision with Dataset 2.

CHAPTER 2

LITERATURE REVIEW

2.1 Neurodegenerative Disease

The term "neurodegenerative diseases (NDD)" encompasses an umbrella of diseases such as PD, Alzheimer's disease (AD), amyotrophic lateral sclerosis (ALS), Huntington's disease (HD), among others, wherein the nervous system's structure and function gradually deteriorate, impairing movement, cognitive function, and other neurological operations. Based upon the location in the brain where the loss of neurons is taking place, specific disease can be identified by their specific manifestations. Usually, there is a clear correlation between the degree of neuronal death and the onset and development of clinical manifestations. In AD, neuronal loss occurs early in the hippocampus, impacting memory formation, while in PD, symptoms like tremor, movement sluggishness, and imbalanced body posture typically appear after substantial loss of dopamine generating neurons in the substantia nigra (SN). Protein aggregates, such as α -synuclein in PD, huntingtin protein in HD, TAR DNA-binding protein 43 in ALS and amyloid-beta plaques in AD, serve as a distinctive characteristic for NDD [11-13]. These aggregates cause disruptions of regular cellular functions and are implicated in the malfunction and demise of neurons. Other pathological mechanisms such as inflammation in the brain, excitotoxicity of glutamate receptors, oxidative stress, and mitochondrial dysfunction that leads to impaired calcium homeostasis and ATP production, further aggravate the neuronal and brain damage, hastening the advancement of NDDs [14], [15]. The predominant risk-enhancer contributor for NDD is age, individuals aged 45 years and above are more vulnerable to NDD. It is revealed that an individual's genetic constitution, way of living, and environment also contribute to risk enhancement [16-18]. Genetic abnormalities may occur via various mutations or by altering the regulation of some key genes. Environmental factors such as pollutants, exposure to toxins, brain injuries, and the way of living further modulate the disease onset and advancement [19]. Diagnosing NDDs in the preliminary stages can be very challenging, as these are marked by only a few common symptoms such as mobility problems, behavioural changes, and cognitive impairments. Therefore, it is necessary to comprehend all these factors thoroughly to diagnose NDD and develop novel treatment plans. Making noteworthy changes in the lifestyle, such as a healthy diet, regular yoga and exercising, and regular cognitive activities, could significantly manage NDD symptoms and enhance the well-being of the patients. R&D advancements have paved the way for researchers to delve more into novel approaches such as gene therapy, stem cell therapy, protein targeting therapy, brain simulation techniques and discovering more neuroprotective agents. [20], [21].

2.2 Parkinson's Disease

The 2nd most prevalent NDD, Parkinson's disease (PD) is epitomised by a gradual diminishment of dopamine- generating neuronal cells in the SN pars compacta of the midbrain. Dopaminergic neurons loss in PD brain results in a decline in dopamine levels, a neurotransmitter that regulates motor functions, enthusiasm, cognitive functions, memory, and other processes. The reduction in dopamine levels in the PD brain is the root cause for deficiency in motor functions and may also be the cause for decline in cognitive functions that some PWP perceive. PD involves multiple neuronal networks and various organs, including the adrenal glands, heart, skin, and retina. It is characterized by the widespread presence of abnormal protein aggregates called α-synuclein-containing Lewy bodies and Lewy neurites within cells and initial signs of pathology often appear in the olfactory bulb and gastrointestinal tract [22]. The motor symptoms associated with PD such as rigidity in the muscles, bradykinesia, imbalanced posture, and resting tremor are collectively referred to as parkinsonism [23]. PD advances in stages, beginning with moderate symptoms emerging only on one side of the human body and progressing to bilateral symptoms, significant issues related to balance and coordination, severe symptoms seeking aid, end-stage mobility impairment, and cognitive decline [24]. There is a substantial gap in the emergence of clinical symptoms in PD from the time that cells in vulnerable nervous system nuclei are first damaged. Typically, PD symptoms and signs do not appear until 70–80% of dopaminergic neurons have been lost [25]. It is crucial to stumble upon reliable molecular biomarkers in order to diagnose PD, gauge its progression, and assess the efficacy of current treatments. The four main categories of these biomarkers are genetic, biochemical, imaging, and clinical. PD typically appears in elderly individuals, genetic variations may manifest in younger patients, genetic variations may manifest in younger patients. By 2030, the prevalence of PD is expected to increase by almost 30% owing to the increase in aging population [26]. Although there is no known cure for PD, but individuals affected with PD can significantly improve their quality of life and alleviate their symptoms with the aid of timely interventions in treatments such as medication, physical therapies, along with deep brain stimulation [27].

2.3 Role of AI in NDD diagnosis

The co-relation between neuroscience and AI is inextricably linked, with AI offering a broad spectrum of applications across different domains, all aimed at endowing machines with human-like intelligence to effectively perform complex tasks such as speech recognition, gaming, autonomous driving, intelligent traffic management, robotic surgery, image and video analytics, natural language processing

(NLP), and more. Simultaneously, neuroscience, by studying the structure and functionality of the brain, contributes to the efficient detection and diagnosis of various neurological disorders. Among the innovative approaches gaining prominence are ML and deep learning (DL) techniques. These computational methodologies have demonstrated remarkable potential in revolutionizing the diagnosis of NDD, offering a paradigm shift from traditional diagnostic approaches [28-30]. NDDs like AD and PD (among many others) hamper the victim's quality of life and often lead to discomfort and a challenging life, both for the patient and the caretaker(s). Therefore, correct diagnosis at the appropriate time is a pre-requisite for the treatment course of any ailment to ensure maximum help from the healthcare staff. ML algorithms designed for disease detection are computational models that analyze medical data to identify indications of diseases at an early stage, potentially preventing severe outcomes. Over the last decade, advancements in technology have facilitated the swift collection of extensive patient data, including ultrasonography and MRI results, omics profiles from biological samples, electronically recorded clinical, behavioural, and activity data, as well as information sourced from social media [31]. These large health datasets are characterized by high dimensionality, indicating that the number of features or variables documented per observation may occasionally surpass the total count of observations [32]. By leveraging the power of algorithms and neural networks, ML and DL contribute significantly to early detection, efficient classification, and personalized treatment plans for individuals affected by conditions such as AD, PD, and other NDD. This exploration delves into the role of ML and DL in enhancing diagnostic precision, thereby shaping the future landscape of NDD diagnosis and treatment.

2.4 Different types of Machine Learning techniques

ML algorithms streamline the clinical decision-making procedure by autonomously categorizing and forecasting the advancement of diseases through computer-aided diagnosis (CAD) [32]. This replaces the manual interpretation typically conducted by medical professionals. For example, in medical imaging, algorithms can analyze intricate patterns within images to identify anomalies, providing efficient and rapid diagnostic support. By automating the classification and prediction tasks, ML not only expedites the decision-making procedure but also reduces the risk of human error. ML comprises diverse methodologies, and it can be categorized into three fundamental types (as illustrated in Fig. 2.1): Supervised, Unsupervised and reinforcement Learning [33].

Supervised ML: In this, algorithms are trained on labelled dataset, signifying that each input is matched with its intended or desired output. Healthcare experts annotate datasets with specialized human input, including Neuropsychologists for cognitive assessments, Neuroscientists and clinicians for CSF biomarkers or tau protein, and Radiologists for MRI scans. The algorithm identifies features, links them to labels, and predicts new labels for unlabelled data by considering recent input features [34]. Two fundamental categories of supervised ML methods include classification and regression [35]. Classification algorithms, like DT, SVM, naïve Bayes, KNN, and ensemble classifiers, predict categorical responses in areas such as medical image processing or speech recognition [36]. On the other hand, Regression algorithms like linear and LR, support vector regression, and ensemble methods, are tailored for predicting continuous output variables, such as forecasting the rate of cognitive decline over time [37]. Collectively, classification and regression contribute to identifying patient subgroups, clustering similar areas in data, offering insights into personalized patient profiles, and enhancing targeted healthcare interventions.

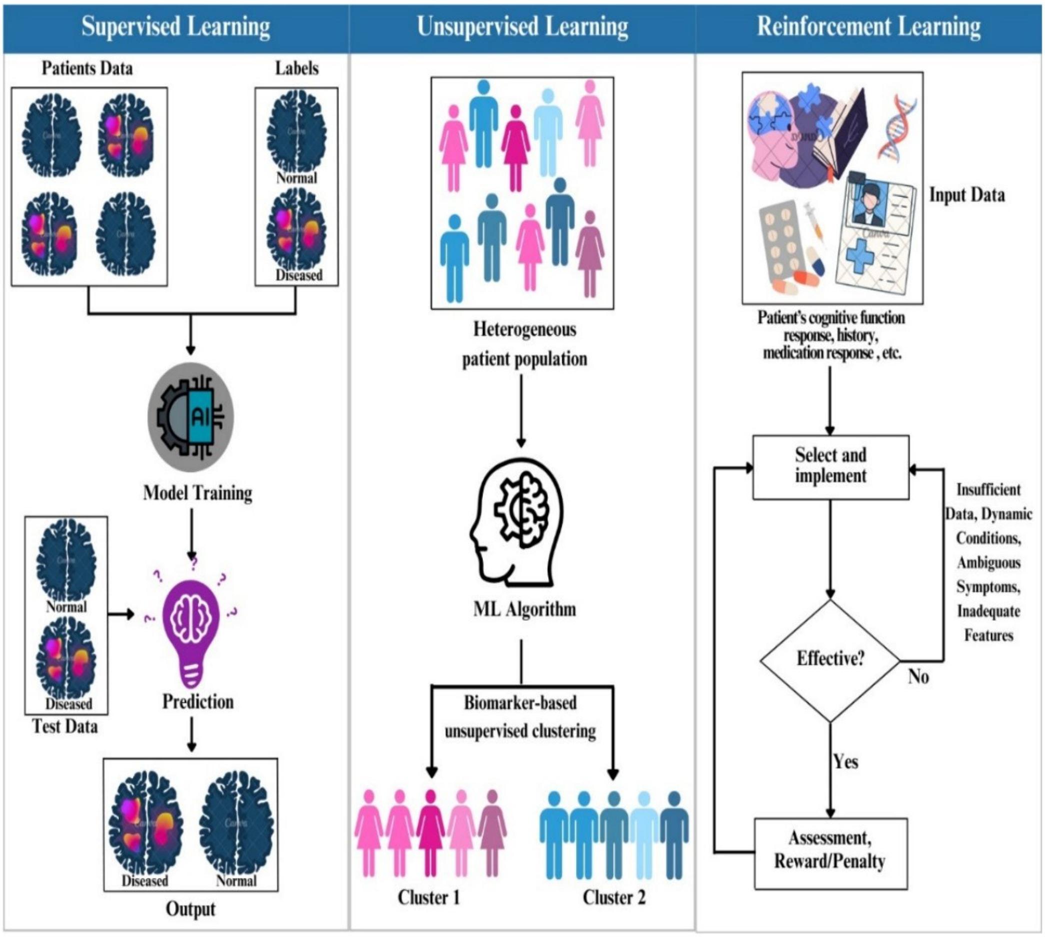

Supervised Learning
Patients Data
Labels
Normal
Diseased
Model Training
Normal
Diseased
Test Data
Prediction
Diseased
Normal
Output
Unsupervised Learning
Heterogeneous patient population
ML Algorithm
Biomarker-based unsupervised clustering
Cluster 1
Cluster 2
Reinforcement Learning
Input Data
Patient's cognitive function response, history, medication response , etc.
Select and implement
Insufficient Data, Dynamic Conditions, Ambiguous Symptoms, Inadequate Features
Effective?
No
Yes
Assessment, Reward/Penalty

Fig. 2.1: Different types of ML techniques: supervised, unsupervised and reinforcement

Unsupervised ML: Unlike supervised ML, unsupervised learning examines datasets without predefined labels, operating autonomously through a data-centric approach, without requiring human intervention [38]. It utilizes clustering algorithms to categorize individuals based on similarities in medical images (MRI, PET) or biomarkers, exposing distinctive disease patterns [39]. Simultaneously, methods for reducing dimensionality of the dataset such as Principal Component

Analysis (PCA) help visualize complex data [40]. PCA transforms dimensionality (higher to lower) of the dataset while conserving the dataset's key attributes. In a study, researchers proposed an ML approach for PD diagnosis using data partitioning and PCA for feature selection. The LR algorithm, SVM with GNB, and weighted KNN classifiers demonstrated lower accuracies compared to the integrated approach involving classifiers, data partitioning, and feature selection, showcasing the efficacy of PCA in classification [41].

Semi-supervised learning: This blends aspects of supervised and unsupervised learning [42]. Initially, the algorithm learns from small labelled data sets and refines predictions. It then leverages the broader insights from unlabelled data to enhance overall performance .

Reinforcement learning: In this, machines or software agent develop decision-making abilities by engaging with their environment. As they explore through trial and error, these agents obtain feedback in the form of a reward or penalty [43]. This process supports adaptive behaviour, fostering continuous improvement over time. For instance, an agent, representing a diagnostic system, learns from patient data such as brain imaging, genetic markers, and clinical history, receiving rewards for accurate predictions and penalties for errors, optimizing its ability to identify disease [44].

2.5 ML algorithms

ML algorithms forms the backbone for developing models that learns from the dataset to make prediction for disease classification. Some of the important algorithms used in this thesis are:

Support Vector Machine: SVM seeks to determine the optimal hyperplane that effectively segregates the datapoints of different classes for classification purpose [45]. The optimal hyperplane is the one that enlarges the margin between different classes, as depicted in Fig.2.2. The datapoints nearest to the hyperplane are regarded as support vectors. SVM employs kernel trick to handle non-linear datasets, which involves mapping the dataset into a higher dimensional space which offers the possibility of linear separation [46], [47]. It is well renowned for its versatility, memory efficiency and effectiveness in handling high-dimensional datasets proficiently.

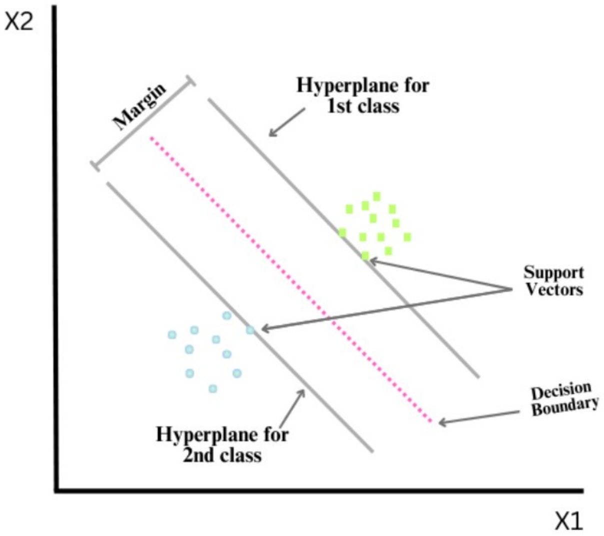


Fig. 2.2 : SVM linear classifier

Logistic Regression: LR is a statistical technique employed for classification tasks that simulates the likelihood or probability of a binary outcome depending upon one or more predictor elements or variables [48], [49]. The resulting probability is constrained to lie between 0 and 1. Fig. 2.3 illustrates the S-shaped curve of LR. This technique employs the sigmoid function for modelling the relationship between dependent and independent variables. It is recognized for its simplicity and probabilistic output.

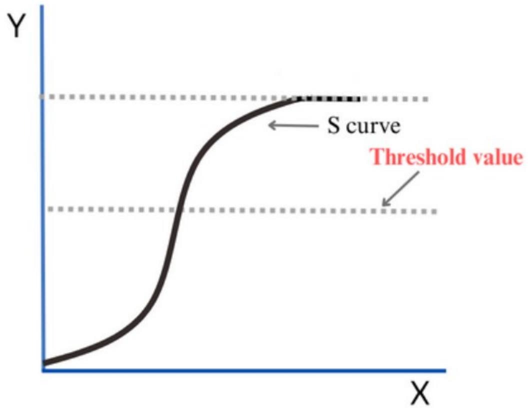


Fig. 2.3 : Logistic regression curve

K-nearest neighbours: KNN is a non-parametric algorithm that lacks an explicit training phase. KNN works on the principle of “similarity” [50]. It assigns a class to ta new datapoint by considering the class of ‘k’ neighbours in the proximity. Thus, selecting an optimal ‘k’ value is crucial for achieving superior performance.

Euclidean distance is typically used for measuring the similarity with the neighbours [51]. Fig. 2.4 illustrates an example of new data point classification using KNN.

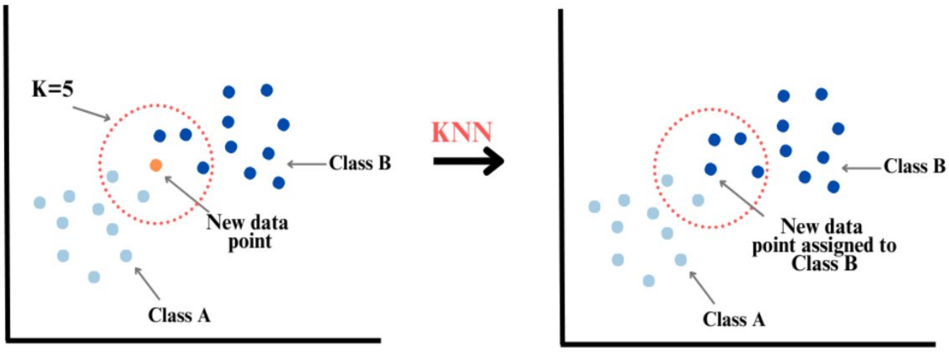


Fig. 2.4 : Classification of a new data point using KNN

Decision Tree: It a highly versatile algorithms that operates by partitioning the dataset into subsets recursively, generating tree like structures ultimately [52], [53]. Each internal node signifies a choice made in compliance with target value and each terminal or leaf node signifies the anticipated outcome [54]. Fig. 2.5 illustrates the framework of DT. It is highly crucial to select an optimal feature at every node in order to achieve the best performance. Information gain is used as a measure to check the effectiveness of the model. DT employs pruning technique to mitigate over-fitting problems.

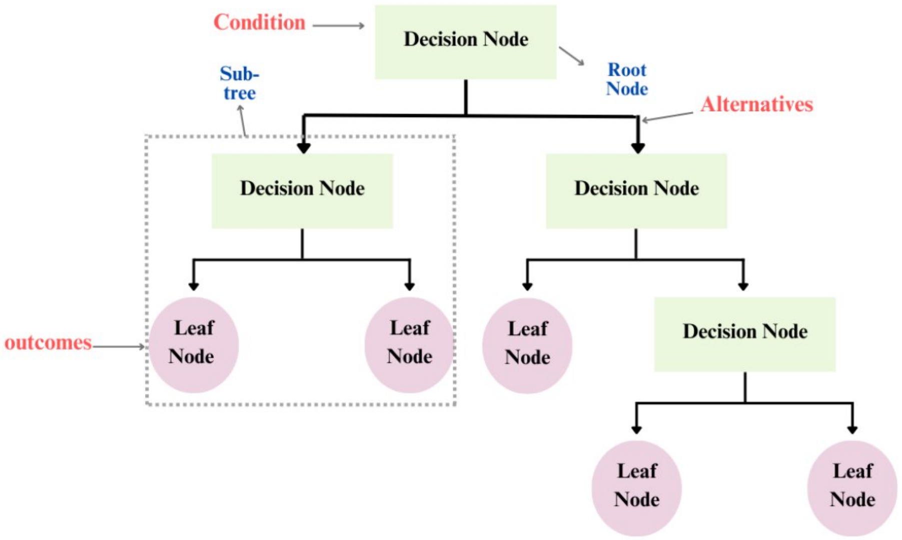


Fig. 2.5 : Framework of Decision tree

Random Forest: RF algorithm develops and aggregates a multitude of DTs for outcome prediction [55]. Each DT is developed by utilising a random subset of the training data. RF basically employs "bagging" technique i.e., bootstrap plus aggregation [56], [57]. Accumulation of more number of DTs boosts up the performance of the RF model, as depicted in Fig. 2.6. It is well known for its accurate predictions, scalability and robustness to overfitting.

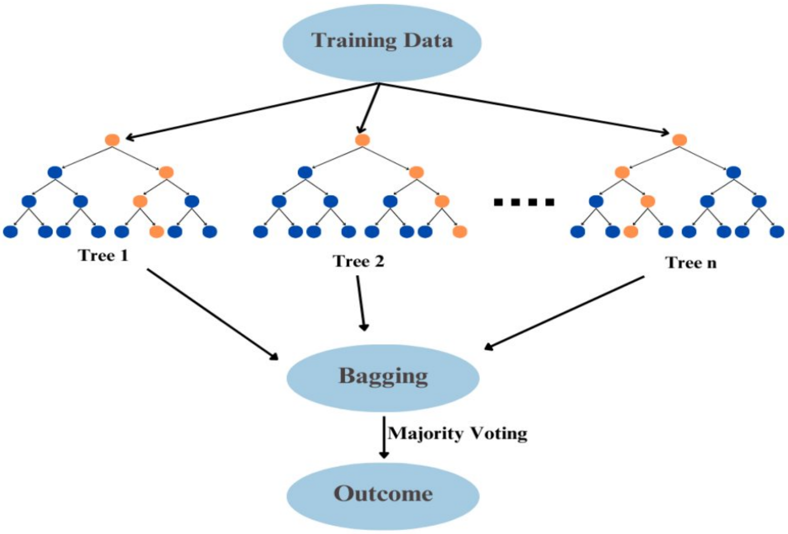


Fig. 2.6 : Framework of Random Forest

Extreme Gradient Boosting: XGB is widely known as the 'gold standard' of ensemble learning. It operates by generating multiple DTs in a series manner and consecutively rectifying the errors of its predecessors by employing an optimisation technique regarded as gradient descent [58], [59]. Then a powerful predictive model is constructed by clustering all the DTs together. XGB has a unique method of mitigating over-fitting issues, which involves pruning, regularisation terms as well as learning rate. Some other exceptional characteristics of this algorithm are high accuracy, capability to handle large and missing datasets, and parallel processing.

Gaussian Naïve Bayes: GNB, a probabilistic classifier, presumes that each feature is independent of other and determine the class of a new datapoint based on this assumption [60]. It operates calculating the likelihood that the data point belongs to each class and then assigning the most likely class to the new data point[61]. GNB is recognised for its easy implementation, scalability and proficient computational efficiency.

Principal Component analysis: PCA is based on the principle of "orthogonal transformation" which generates new set of features referred as component features [62], [63]. These new features are the linear combination of the original features based on the variance they account. PCA basically reduces the complexity of the dataset by extracting key features. Thus, it is highly critical to choose the optimal number of key features for attaining proficient results.

2.6 NDD related data types used by ML algorithms

In numerous neurodegenerative conditions like AD, PD, and motor neuron disease (MND), symptoms often remain latent until a significant neuronal loss has transpired, posing a substantial challenge for early diagnosis [64]. Therefore, there is an increasing interest in leveraging ML models for early detection within the research community [65]. In this section, the major data types related to NDDs (for example – Magnetic Resonance Imaging (MRI), Electroencephalogram (EEG), Biomarker studies, etc.) will be discussed briefly. These data types are then used for ML models training for disease classification.

Neuroimaging: Diagnosis of NDDs posing a great challenge of undetectability is often conducted with the help of neuroimaging techniques such as MRI, CT, and PET scans. Various quantification methods offer complementary insights, and therefore, optimal outcomes are achieved by employing multiple quantification techniques. The findings of [66] indicate the feasibility of automatic quantification methods and computerized decision support systems in clinical practice. They furnish comprehensive information that could aid clinicians in the foreseeable future. Research has been conducted to perform differential diagnosis of NDDs using structural MRI data [66], most investigations into distinct NDDs utilizing structural MRI predominantly rely on visual ratings for characterization [66-70], volumetry [68], [71-74] and morphometry analysis [66], [72], [75-81]. In a separate study [82], PD was targeted for diagnosis using an ML-based framework of neuroimaging. In diagnosing Parkinsonism, two primary considerations involve distinguishing between conditions causing tremor without the deficiency of dopamine (e.g., essential tremor or dystonic tremor) and those leading to an akinetic-rigid syndrome, like multiple system atrophy (MSA) or progressive supranuclear palsy (PSP).

Cognitive performance tests: A cognitive performance test in the realm of NDDs is a thorough assessment devised to gauge multiple aspects of cognitive function, encompassing memory, focused attention, executive function, language, and visuospatial skills, among others [83]. These tests typically comprise a series of standardized tasks, exercises, or queries aimed at evaluating different cognitive domains. Examples of tasks may include recalling lists of words, solving puzzles, following instructions, naming objects, drawing specific shapes, and performing mathematical calculations. The outcomes of these assessments furnish valuable insights into an individual's cognitive strengths and weaknesses, as well as any impairments or deteriorations in cognitive function indicative of a NDD. For instance a study offers initial evidence indicating that performance metrics collected via webcam, utilizing AI algorithms to capture gaze and facial expression data, can reliably identify individual and group disparities in neurobehavioral function [84].

Motor performance tests: The majority of NDDs experience motor deficits at certain phases. Symptoms of motor impairment include muscle spasms extrapyramidal stiffness, bradykinesia, and gait problems including slowing down or being careful when walking [85]. More attention is being placed on diagnostic methods, as well as the creation and selection of therapies that address motor impairments and the ensuing constraints on activities, without categorising individuals based solely on medical diagnoses [86]. Tests that evaluate various aspects of motor control and coordination, for instance Imple Reaction Time, Choice Reaction Time, Movement Time, Fitts' Law tasks, and thorough gait analysis, can be used to assess these impairments [87]. Then the Key aspects including response times, mistake rates, and gait patterns are extracted from the data generated from these tests through a thorough analysis.

EEG: The EEG has emerged as a valuable diagnostic and research tool for dementia, particularly in the context of AD. It aids in both the differential diagnosis and prediction of disease progression. It can be challenging to diagnose AD as its symptoms are similar to other age-related cognitive variations. Therefore, health professionals must conduct a meticulous evaluation to rule out possibilities of other conditions. This can be achieved by conducting comprehensive tests, neurological examinations, blood tests, imaging scans, spinal fluid tests, and psychological tests like mini-mental state examination. EEG signals in patients with dementia often have less intricacy and fewer functional associations, as revealed by non-linear dynamic EEG data analysis. This highlights the potential of EEG for diagnosing and monitoring dementia-related brain function variations.

Transcriptomic data: Transcriptomics involves the examination of RNA transcript levels by employing microarray technology. RNA microarrays, usually consisting of known sets of transcripts, are utilized for this objective [88], [89]. Research in transcriptomics has assisted in characterizing genes with differing expressions and understanding critical biological processes and pathways, significantly aiding neurodevelopmental studies. This progress has fostered the assessment of NDDs from a perspective of transcriptomics, enhancing our comprehension of these conditions. The outcomes of transcriptomics studies aid in developing personalized medicines, and gene therapy, and discovering distinct biomarkers for every illness [90]. RNA sequencing is crucial for diagnosing individuals with NDD who haven't received a genetics diagnosis before [91]. Several algorithms have been used for conducting mRNA analysis for diagnosing NDDs. FRASER stands out as a newly devised algorithm designed to detect anomalies in mRNA splicing with a high degree of precision [91], [92].

Biomarker data: Genetic biomarkers serve as crucial indicators of diseases, rooted in the variations found within an individual's DNA sequence [93]. These variations, commonly referred to as genetic polymorphisms, have the aptitude for influencing the expression or functionality of specific genes, potentially heightening an individual's susceptibility to NDDs. Biomarkers play a critical role in aiding the diagnosis of NDDs, especially in the early stages where symptoms may not be readily discernible. One such example is the utilization of imaging techniques to observe brain changes, aiding in the diagnosis of AD [93]. Researchers are actively investigating specific biomarkers present in blood or cerebrospinal fluid that could facilitate the early identification of various disorders. Measurement of protein concentrations in the cerebrospinal fluid is one avenue being explored to assist in diagnosing illnesses [94-96].

Metabolomic data: Metabolomics is an area of science that's growing quickly. It's all about studying the small molecules in cells, tissues, organs, or whole organisms [97], [98]. These molecules are like unique signatures that tell us a lot about how cells work and what's happening inside them [99], [100]. Metabolomics has shown promise in helping diagnose PD, assessing the likelihood of family members developing it, measuring how drugs work in the body, and making drug development more efficient. Currently, researchers have identified fifty-six metabolites linked to PD in the Human Metabolome Database [101]. Moreover, a unified analysis of metabolomics and proteomics has revealed disruptions in the metabolism of lipid, including an activated metabolism of sphingolipid and reduced apolipoproteins in the plasma of PWP [101], [102].

2.7 Speech impairment as a biomarker for early PD diagnosis

Early and accurate differential diagnosis of NDDs is crucial for several reasons. Firstly, research indicates that early diagnosis, when paired with existing treatments, can significantly delay the advancement of the disease and alleviate the need for hospitalization [66], [103]. Furthermore, as potential disease-modifying drugs are developed, the significance of early diagnosis is expected to increase even further [104]. Secondly, developing new treatments necessitates the precise identification of target populations at an early stage. It has been suggested that the failure of certain pharmaceutical trials in the past could be attributed to the inclusion of overly heterogeneous study populations. Therefore, early and accurate diagnosis plays a vital role in both enhancing patient outcomes and facilitating successful clinical research aimed at developing novel therapies for NDDs [66], [105]. PD has five stages of its progressions and most symptoms remains latent until many neurons has been degenerated. But according to research, stage 0 of PD is marked by vocal cord impairment in more than 90% PD affected individuals. Dysfunctions in the speech mechanisms during any of the basic motor processes necessary for producing speech give rise to speech disorders [106]. In PD, lack of coordination of the muscles producing sound or impaired vibratory activities of vocal cords could be the reason behind phonetic impairment [107]. Abnormalities such as poor articulation, shaking or hoarseness, altered frequencies, diminished quality of sound, decreased rhythm, absence of emotional expressivity, and fluctuations in tone, are the common characteristics of speech impairment [108]. In today's digital era, it is quite easy to measure the voice abnormalities using voice recordings from digital devices or smart phones. These recordings can then be analysed by healthcare professionals for PD detection using automated technologies for higher accuracy. Following detection, doctors can then halt the course of PD by reactivating dopamine-producing neurons in the brain by deep brain stimulation or pharmaceutical therapies ensuring better quality of life for patients. This method can have vital applications in telemedicine, revolutionizing the delivery of medical services in remote areas. There is no cure for PD presently owing to its complex nature, but early medical interventions could help patients to live a normal life.

2.8 Revolutionizing traditional diagnostics with AI innovations

Interpreting medical images, including X-rays, MRI, and CT scans requires a more sophisticated approach than basic equations, as medical imaging diagnoses need to be learned through dedicated training processes. The ML and DL algorithms learn by examining training data and generating predictions when presented with new data, providing enhanced precision and reliability compared to conventional manual interpretation, particularly when managing large datasets [109-113]. Thus, AI models demonstrate efficiency in analyzing extensive imaging data and identifying nuanced patterns, anomalies, and structural changes that might not be immediately discernible to human observers [114]. The heterogeneity in NDD presents challenges in understanding and treatment due to diverse manifestations and disease trajectories among individuals, complicating efforts to decipher common mechanisms and develop targeted treatments [115-117]. ML can anticipate the trajectory of the disease and possible new symptoms, which might not be apparent to humans. It also provides flexibility in the healthcare industry, independent of predetermined rules and assumptions [118], [119]. To detect subtle alterations, it can potentially analyze distinct data types such as medical scans, voice recordings, clinical records, and molecular profiles [120], [121]. For instance, ML algorithms can detect impaired cognitive issues by spotting minute changes in how an individual remembers things over time or by recognizing variations in speech attributes such as the pronunciation of vowels, fundamental frequencies, fluency, and many more [122], [123]. It also assesses individuals' performance in tasks like attention and problem-solving, furnishing physicians with richer insights beyond self-reported symptoms [124]. AI has the capability to discern novel molecular biomarkers associated with the pathology of NDD, by evaluating the multi-omics data from large-scale studies. Conventional methods typically consider only a small group of individuals with limited attributes, resulting in oversimplifying the complexities of NDD. Subtype and Stage Inference (SuStaIn), an ML method, overcomes this problem by taking into account diverse data of patients to determine different phenotypes of the disease and advancement in stages. In a study SuStaIn successfully identified different groups and their unambiguous brain degeneration pattern, affirming its potential to categorize different subtypes in genetic frontotemporal dementia [125]. The fusion of IoT and AI, especially ML, has revolutionized the healthcare industry. IoT sensors are capable of tracking individuals, monitoring the activity of patients, and predicting their health status. This technology generates vast amounts of medical data, predicts disease, and enables real-time monitoring of patients [126]. To meet the rising demand for remote healthcare, an ML-based application, utilizing sensors and AI, known as AIoMT (Artificial Intelligence of Medical Things), has been developed [127]. In the face of disease progression, changing patient dynamics, and limited specialist availability, ML and DL models present encouraging solutions to address diagnostic challenges [128].

2.9 AI-ML tools in Telemedicine

Telemedicine refers to any medical activity that happens when the doctor and patient are not in the same place. This could include talking over the phone, video calls, or using other communication tools. It's been around for a long time, like when doctors used radios to advise to ship captains far out at sea. As diseases like NDDs progress, patients' motor and cognitive functions keep getting worse over time. This makes it really hard for them and their caregivers to travel to hospitals for medical help. Things like not having good transportation, living far away from hospitals, and not having enough money can make this even harder. So, keeping in touch between patients and doctors becomes a big problem for giving care, keeping track of the disease, and helping out when needed. This is where telehealth and telemedicine come in handy. Telehealth means using electronic devices to give health services. In this case, it can help make sure patients with long-term NDDs get consistent care. Telehealth includes things like telemedicine, which is having appointments with doctors over video calls, tele-coaching, and telecare [129]. Telemedicine has been proven to help manage patients with dementia. It allows doctors to monitor the progression of the disease by giving cognitive tests and staying in touch with patients virtually. This became especially important during the recent COVID-19 pandemic when regular visits to hospitals were difficult. ML algorithms can be utilised for remotely analysing MRI scans for AD pattern detections, motor symptoms for PD [130], and movement and sleep patterns in HD. This will allow for early diagnosis and personalised treatments in order to enhance patient care and results in remote areas. Patients and caregivers find telemedicine convenient, especially when they can fill out questionnaires on their own. However, it can be challenging for some patients who may not have access to or know how to use technology, especially those living in long-term care facilities.

CHAPTER 3

METHODOLOGY

3.1 Data extraction

The two speech datasets were extracted from ML repository of UCI. The first dataset comprised 195 voice recording with 24 different voice attributes or features such as shimmer, jitter. MDVP, fundamental frequency and many more. Out of 195 voice recordings, 147 belonged to Parkinson's patients and 48 to normal individuals, represented as 1 and 0 respectively in the "status" column. On the other hand, the second dataset comprised 756 voice instances each with 754 attributes such as PPE, DFA, RPDE, TWQT features and many more. Out of 756 instances, 564 recordings belonged to Parkinson's patients (107 men and 81 women) and 192 to the control group i.e., normal individuals (23 men and 41 women), represented as 1 and 0 respectively in the "class" column. The microphone was pre-set to 44.1KHz frequency for audio capturing and continuous vocalisation of the vowel by each subject was recorded thrice.

3.2 Datasets Pre-processing

Both the datasets underwent pre-processing in order to clean the data, handle missing values and have elaborate understanding for the datasets and the trends in their features. Correlation heat map was generated each dataset to have an understanding of correlation between different attributes. Next, the feature (X) and target (Y) were separated. Feature variable comprised all the attributes except for name or id, status or class and target variable comprised "status" or "class". Subsequently, each dataset was segmented into training and testing subsets.

3.3 Model Training

Seven different ML algorithms, comprising SVM, LR, KNN, DT, RF, GNB, and XGB, were used as ML models and underwent training using training dataset.

3.4 Model assessment

Post training, the efficiency of each model was assessed using the testing dataset.

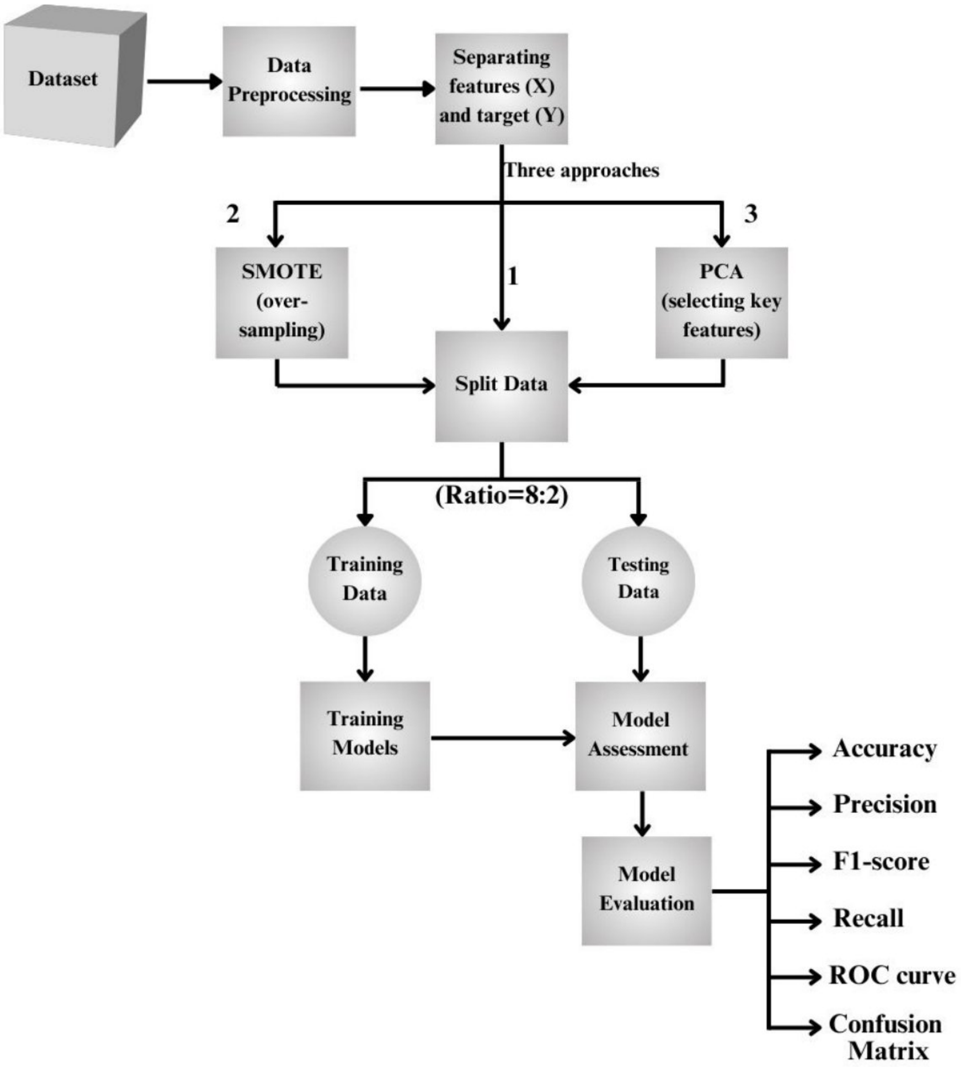


Fig. 3.1: Outline of the methodology

3.5 Model evaluation

The effectiveness of each ML model was evaluated by employing scoring metrics such as accuracy, precision, recall, F1 score, ROC curve and confusion matrix. The formulae for these scoring criteria are shown in Equations 3.1–3.4. True Positives (TP), True Negatives (TN), False Negatives (FN), and False Positives (FP) are the terms in the metrics equation.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3.1)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (3.2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3.3)$$

$$\text{F1} = 2\frac{Precision.Recall}{Precision+Recall} = \frac{2TP}{2TP+FP+FN} \quad (3.4)$$

ROC curve is the graphical portrayal of the performance. It plots true positives rate (TPR) against the false positives rates (FPR). Their formulae are depicted below in equations 5 and 6 respectively. Best model will showcase highest area under the curve (AUC) and higher TPR for a lower FPR.

3.6 Three main approaches

As illustrated in Fig. 3.1, three different approaches were employed during the methodology for each datasets and are described below in detail:

Approach 1: Training the ML models on complete datasets

- Retrieval of audio dataset from UCI website
- Conducting datasets pre-processing to handle missing values, duplicates and gain understanding about attribute patterns. Dropping the columns "name" and "status" in dataset 1 and "id" and "class" in dataset 2.
- Partitioning the dataset into 2 subsets: 80% as training subset and 20% as testing subset
- Standardize the features of training subsets using StandardScaler from the scikit-learn library

- Utilise the data to train ML models
- Assessing and evaluating model performance using various scoring metrics

Approach 2: Over-sampling of the dataset

- Retrieval of audio dataset from UCI website
- Conducting datasets pre-processing to handle missing values, duplicates and gain understanding about attribute patterns. Dropping the columns "name" and "status" in dataset 1 and "id" and "class" in dataset 2.
- Employ SMOTE (Synthetic Minority Over-sampling Technique) from the imlearn library to rectify the imbalance in the number of voice recording for normal and PD patient. Overs-sampling both the classes to equal number of recording, 294 recordings for each class in dataset 1 and 1128 recordings for each class in dataset 2
- Partitioning the dataset into 2 subsets: 80% as training subset and 20% as testing subset
- Standardize the features of training subsets using StandardScaler from the scikit-learn library
- Utilise the data to train ML models
- Assessing and evaluating model performance using various scoring metrics

Approach 3: Training the ML models on 5 key features extracted by PCA algorithm

- Retrieval of audio dataset from UCI website
- Conducting datasets pre-processing to handle missing values, duplicates and gain understanding about attribute patterns. Dropping the columns "name" and "status" in dataset 1 and "id" and "class" in dataset 2.
- Use PCA to identify the top five features from all of the characteristics for model training
- Partitioning the dataset into 2 subsets: 80% as training subset and 20% as testing subset
- Standardize the features of training subsets using StandardScaler from the scikit-learn library
- Utilise the data to train ML models
- Assessing and evaluating model performance using various scoring metrics

CHAPTER 4

RESULTS AND DISCUSSION

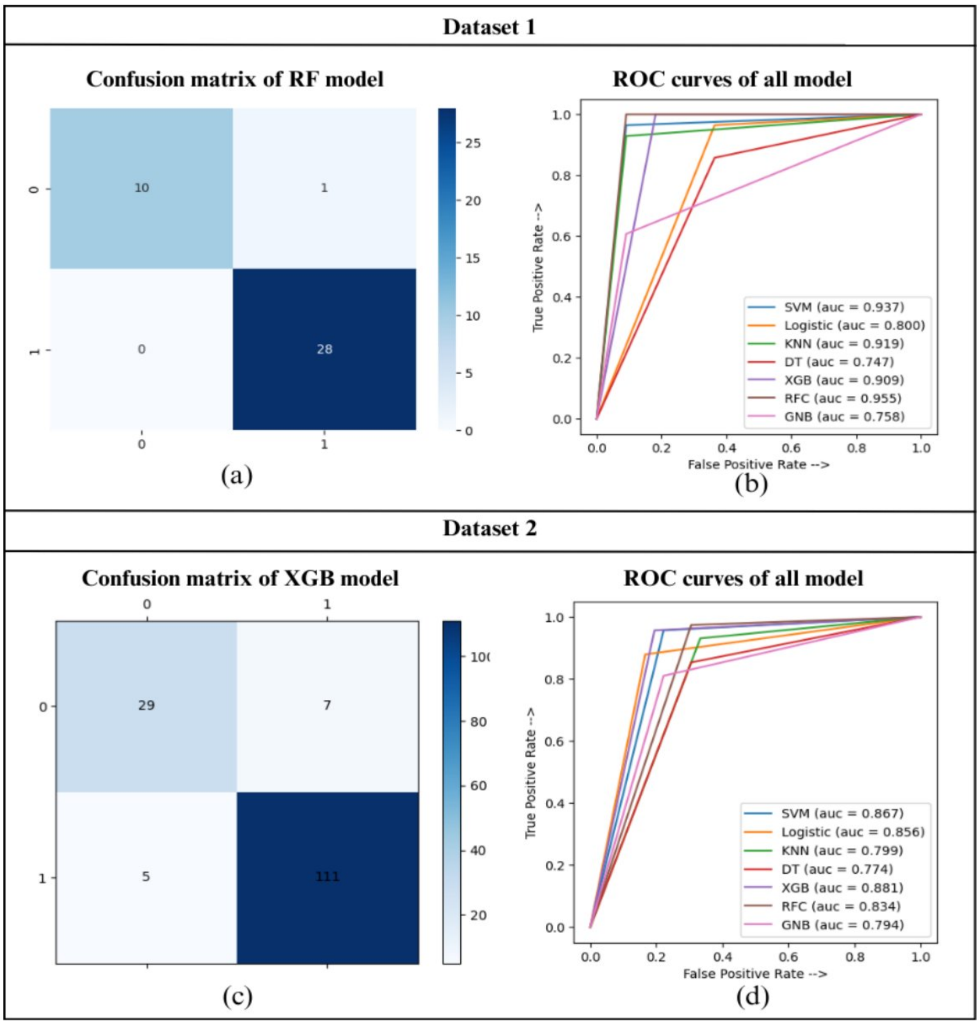
4.1 Findings of approach 1 : Training models on entire dataset

The dataset was partitioned into two subsets, training and testing, in a ratio of 8:2. Followed by data standardization and model training. Performance of the models are evaluated using scoring metrics.

Table 4.1: Results of scoring metrics of ML models trained on dataset 1 and 2 in approach 1

Metric	ML models						
	SVM	*LR*	*DT*	*RF*	*KNN*	*GNB*	*XGB*
Dataset 1							
Accuracy	0.949	0.872	0.795	0.974	0.923	0.692	0.949
Precision	0.964	0.871	0.857	0.966	0.963	0.944	0.933
Recall	0.964	0.964	0.857	1.000	0.939	0.607	1.000
F1- score	0.964	0.915	0.857	0.982	0.945	0.739	0.966
Dataset 2							
Accuracy	0.914	0.867	0.816	0.908	0.868	0.803	0.921
Precision	0.932	0.944	0.900	0.911	0.900	0.922	0.941
Recall	0.957	0.879	0.853	0.974	0.931	0.810	0.957
F1- score	0.944	0.911	0.876	0.942	0.915	0.862	0.949

The performance metrics of ML models across two datasets in approach 1 are presented in Table 4.1. For dataset 1, RF model standout as the best model with an impressive 97.4% accuracy, 96.6% precision and recall of 1.00, highlighting its excellence in classification tasks. 'Auto' for maximum features, 225 estimators, maximum depth of 8, and 'entropy' as the criterion are the best hyperparameters for this model. SVM and XGB rank as the second most effective models, each demonstrating an accuracy of 94.9% and F1-score of 0.96. KNN also performs well with 92.3 % accuracy, 96.3% precision and F1-score of 0.945. GNB model underperforms all models, with only 69.2% accuracy and 0.607 recall, suggesting its shortcomings in handling the dataset complexities. For dataset 2, XGB model excels with 92.1% accuracy, 94.1 precision and F1-score of 0.949. SVM model closely follows the XGB model with 91.4% accuracy, 93.2% precision and F1-score of 0.944, underscoring their reliability. RF model also performs well with 90.8% accuracy and highest recall of 0.974, indicating its robustness. LR and KNN models exhibits similar performance in terms accuracy and F1-score. DT and GNB models shows the weakest performance with 81.6% and 80.3% accuracy, respectively.



(a) (b) (c) (d)

Fig. 4.1: Graphical representation of approach 1 results for both datasets 1 and 2

Fig. 4.1. above showcases the essential performance visualisation of all the model across both the datasets in approach 1. Subfigure (a) presents the confusion matrix of RF model, demonstrating outstanding performance with 28 TP, 10 TN, 1 FP and 0 FN and subfigure (c) presents the confusion matrix of XGB model, highlighting its effectiveness, with 111 TP, 29 TN, 7 FP and 5 FN. Subfigure (b) and (d) represents ROC curves of all models across dataset 1 and 2, respectively. In dataset 1, RF classifier exhibits the highest AUC (0.955), closely trailed by SVM model with AUC of 0.937. This highlights their exceptional performance. XGB also performs well with 0.909 AUC. In dataset 2, highest AUC (0.881) is achieved by XGB model, followed by SVM model with AUC of 0.867. LR and RF models also displays respectable AUC values of 0.856 and 0.834, respectively, while DT and GNB trails with AUC values of 0.774 and 0.794, respectively.

4.2 Findings of approach 2 : Training models on over-sampled dataset

Table 4.2: Results of scoring metrics of ML models trained on dataset 1 and 2 in approach 2

Metric	ML models						
	SVM	***LR***	***DT***	***RF***	***KNN***	***GNB***	***XGB***
Dataset 1							
Accuracy	0.966	0.864	0.898	0.949	0.932	0.864	0.983
Precision	1.000	0.889	0.925	0.933	1.000	0.957	0.967
Recall	0.931	0.827	0.862	0.966	0.862	0.759	1.000
F1- score	0.964	0.857	0.893	0.949	0.926	0.864	0.983
Dataset 2							
Accuracy	0.978	0.934	0.841	0.942	0.876	0.810	0.973
Precision	0.991	0.981	0.839	0.972	1.000	0.802	0.991
Recall	0.966	0.888	0.853	0.914	0.758	0.836	0.957
F1- score	0.978	0.932	0.846	0.942	0.862	0.819	0.973

Results of performance metrics of all models across two datasets in approach 2 are depicted in Table 4.2, revealing their competencies and weaknesses. In dataset 1, The XGB classifier model leads with an impressive accuracy of 98.30% and F-1 score of 0.983, coupled with a perfect recall score of 1.00, demonstrating exceptional performance on the balanced dataset, surpassing the effectiveness of other models in terms of classification. The model was optimised using hyperparameters such as random state set as 300, learning rate of 0.5 and maximum depth at 5. SVM is yet another a top contender, with 96.6% accuracy and 100%, highlighting its robustness. RF classifier also performs well with 94.9 % accuracy and 0.966 recall. LR and GNB models underperforms, each with 86.4% accuracy. In dataset 2, SVM model still maintains its lead with accuracy (97.8%), precision (99.1%) and F-1 score

(0.978). XGB model closely follows, sustaining high accuracy of 97.3% and F-1 score of 0.973, although with a little lower recall. RF remains a robust model with consistent 94.2 % accuracy and 0.942 F-1 score. KNN also performs well with 100% precision, but GNB still remains the least effective model.

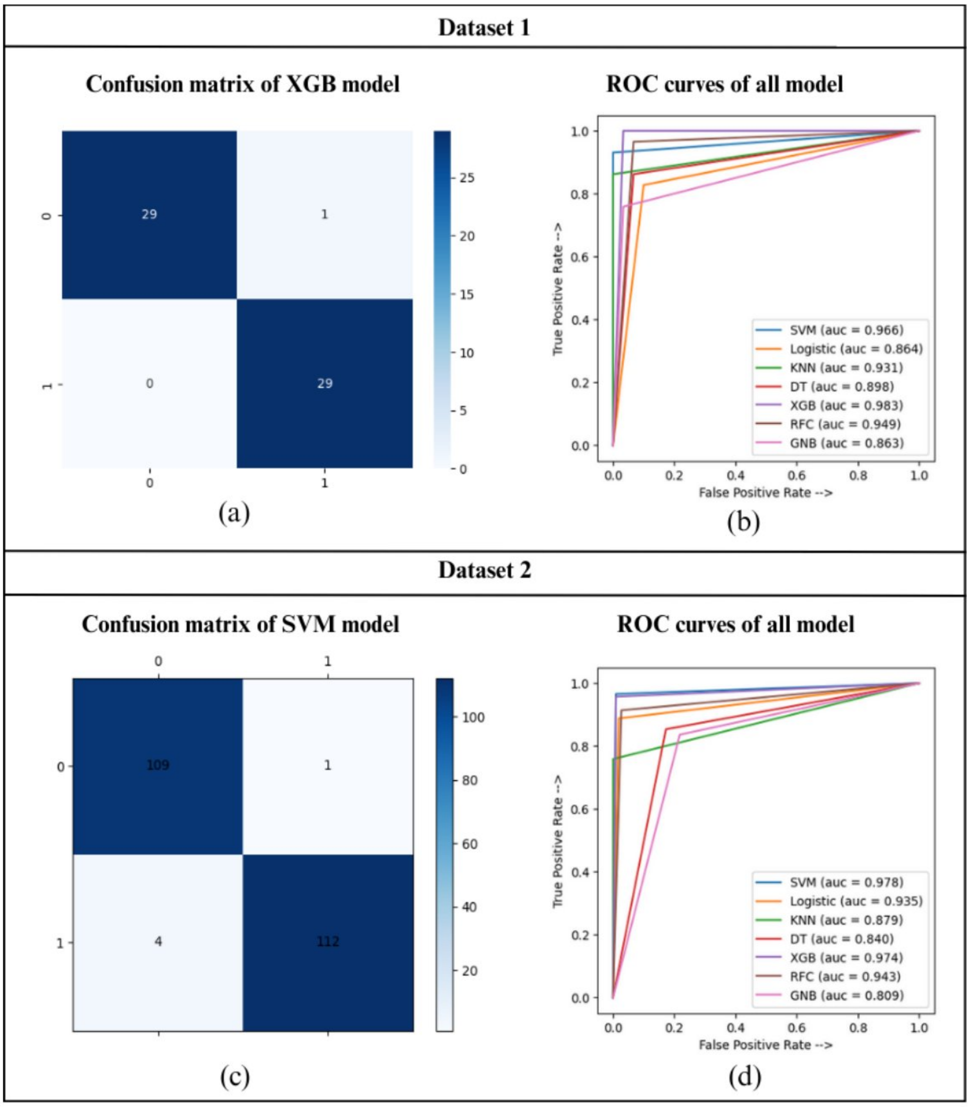


Fig. 4.2: Graphical representation of approach 2 results for both datasets 1 and 2

A comprehensive performance visualization of various models across datasets 1 and 2 in approach 2 is provided in Fig. 4.2. The subfigure (a) presents the confusion matrix of XGB model, the best model, in dataset 1 , showing an exemplary performance with 29 TP, 29 TN, 1 FP and 0 FN. On the other hand, subfigure (c) showcases the confusion matrix of SVM model of dataset 2, revealing 112 TP, 109 TN, 1 FP and 4 FN. Subfigure (b) and (d) displays the ROC curves of all the models in approach2 across datasets 1 and 2, respectively. In dataset 1, highest AUC of 0.983 is achieved by XGB model, highlighting superior performance of XGB in classification tasks. SVM also performs well with AUC of 96.6. In dataset 2, SVM model achieved the highest AUC of 0.978, closely followed by XGB model with AUC of 0.974, indicating their outstanding competence in discriminating between TP and TN classes.

4.3 Findings of approach 3 : Training models on key features only

The scoring metrics of ML models in approach 3 depicted in Table 4.3 reveals significant difference in their performances. For dataset 1, RF and XGB models outperformed other models with 97.4% accuracy and 0.98 F1- score each, highlighting their powerful generalisation expertise and robustness. KNN and DT model also performs commendably, with 96.4 % and 96.3% precision, showcasing their ability to effectively handle FP and FN results. For dataset 2, SVM turns out to be the best model in term of accuracy (88.1%) and F1-score (0.929). It also maintains high recall across both the datasets, indicating its efficiency in locating pertinent instances. KNN also performs well with 84.2% accuracy and recall of 0.917. RF and XGB models displayed similar proficiency in performance with 83.5% accuracy and precision of 86.4% and 88.7%, respectively. Whereas GNB model underperforms across both the datasets, suggesting its limitation for handling the complexity of the datasets.

Table 4.3: Results of scoring metrics of ML models trained on dataset 1 and 2 in approach 3

Metric	**ML models**						
	SVM	***LR***	***DT***	***RF***	***KNN***	***GNB***	***XGB***
Dataset 1							
Accuracy	0.897	0.872	0.923	0.974	0.949	0.872	0.974
Precision	0.962	0.871	0.963	0.965	0.964	0.960	0.965
Recall	0.893	0.964	0.929	1.000	0.964	0.857	1.000
F1- score	0.926	0.915	0.945	0.982	0.964	0.906	0.982
Dataset 2							
Accuracy	0.881	0.822	0.737	0.835	0.842	0.690	0.836
Precision	0.887	0.867	0.865	0.864	0.888	0.785	0.887
Recall	0.975	0.917	0.793	0.943	0.917	0.843	0.909
F1- score	0.929	0.891	0.827	0.901	0.902	0.813	0.897

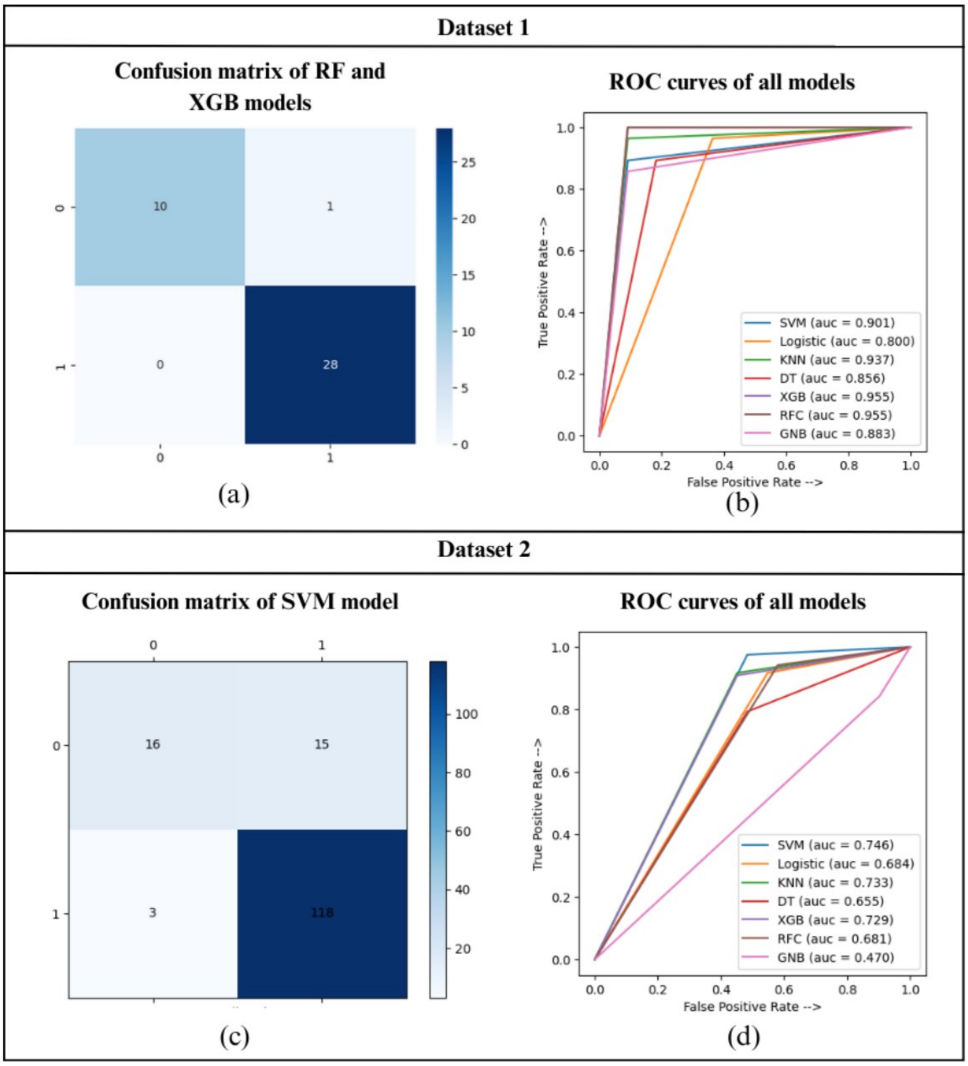


Fig. 4.3: Graphical representation of approach 3 results for both datasets 1 and 2

Key performance visualisation of all models across both the dataset in approach 3 is presented in Fig. 4.3. above. Subfigure (a) displays the confusion matrix of XGB model on dataset 1, revealing 28 TP, 10 TN, 1 FP and 0 FN, whereas subfigure (c) displays the confusion matrix of SVM model on dataset 1, revealing 118 TP, 16 TN, 15 FP and 3 FN. ROC curves of all models on dataset 1 and 2 are depicted by subfigures (b) and (d), respectively. The AUC values are particularly noteworthy. In dataset 1, the highest AUC of 0.955 is attained by RF and XGB models, highlighting their superior classification ability. KNN and SVM also performs well. In dataset 2, SVM leads with an AUC of 0.746, followed by XGB at 0.729. Most models witnessed a decline in performance, with GNB having the least AUC (0.470).

CHAPTER 5

CONCLUSION

PD is one of the most widespread NDD, with its prevalence increasing with age. Nearly 10 million individuals are affected PD across globe, typically in older age groups. It presents significant challenges due to absence of cure. So it becomes primary concern to diagnose PD in preliminary stages in order to delay or prevent its progression into a potentially severe conditions. The fusion of healthcare and AI has marked the commencement of a new era for diagnosing and managing diseases, particularly NDD. Leveraging the vast amounts of data available today, AI is poised to revolutionize healthcare by automating diagnosis tasks. ML algorithms designed for disease detection are computational models that analyse medical data to identify indications of diseases at an early stage. In this study, we conducted comparative analysis of seven different ML models, comprising SVM, LR, DT, RF, KNN, XGB and GNB, utilising two different speech datasets with multiple attributes. Three different approaches were adopted during the methodology: Conducting model training on the entire dataset, over-sampling the dataset to equalise the number of recordings in both the classes, and training the models only on 5 key attributes extracted by PCA. Outcomes reveal that the SVM, RF and XGB are the most reliable models, exhibiting superior performance across both the datasets in all the approaches. Conversely, LR and GNB models exhibited lowest efficiencies in all tests, highlighting their limitations in handling the complexities of the datasets. It is also observed that SVM and XGB models performed exceptionally well when trained on balanced datasets, in the second approach, scoring highest scores in all the scoring metrics. XGB showcased highest accuracy of 98.3% in dataset 1, while SVM achieved highest accuracy of 97.8% in dataset 2. Third approach also revealed promising results, indicating the importance feature selection for model training. SVM and XGB models has the potential to offer accuracy of clinical level when trained on best hyper-parameters and large datasets. These could also serve potential application in telemedicine or remote healthcare, by analysing and interpreting the voice recordings or other biomarkers captured by patients using smart technologies, enabling early, accurate and real-time diagnoses and personalized care. Thus enhancing the outcomes in remote healthcare settings.

REFERENCES

[1] R. Xia and Z. H. Mao, "Progression of motor symptoms in Parkinson's disease," *Neuroscience Bulletin*, vol. 28, no. 1. 2012. doi: 10.1007/s12264-012-1050-z.

[2] A. L. Mahul-Mellier *et al.*, "The process of Lewy body formation, rather than simply α-synuclein fibrillization, is one of the major drivers of neurodegeneration," *Proc Natl Acad Sci U S A*, vol. 117, no. 9, 2020, doi: 10.1073/pnas.1913904117.

[3] M. C. Rodriguez-Oroz *et al.*, "Initial clinical manifestations of Parkinson's disease: features and pathophysiological mechanisms," *The Lancet Neurology*, vol. 8, no. 12. 2009. doi: 10.1016/S1474-4422(09)70293-5.

[4] D. E. Hobson, "Clinical manifestations of Parkinson's disease and parkinsonism," *Canadian Journal of Neurological Sciences*, vol. 30, no. SUPPL. 1. 2003. doi: 10.1017/s0317167100003188.

[5] F. Khaliq, J. Oberhauser, D. Wakhloo, and S. Mahajani, "Decoding degeneration: The implementation of machine learning for clinical detection of neurodegenerative disorders," *Neural Regeneration Research*, vol. 18, no. 6. 2023. doi: 10.4103/1673-5374.355982.

[6] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain Cogn*, vol. 56, no. 1, 2004, doi: 10.1016/j.bandc.2004.05.002.

[7] A. A. Moustafa *et al.*, "Motor symptoms in Parkinson's disease: A unified framework," *Neuroscience and Biobehavioral Reviews*, vol. 68. 2016. doi: 10.1016/j.neubiorev.2016.07.010.

[8] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins, and R. H. Ghomi, "Parkinson's Disease Diagnosis Using Machine Learning and Voice," in *2018 IEEE Signal Processing in Medicine and Biology Symposium, SPMB 2018 - Proceedings*, 2018. doi: 10.1109/SPMB.2018.8615607.

[9] F. Amato, L. Borzì, G. Olmo, and J. R. Orozco-Arroyave, "An algorithm for Parkinson's disease speech classification based on isolated words analysis," *Health Inf Sci Syst*, vol. 9, no. 1, 2021, doi: 10.1007/s13755-021-00162-8.

[10] G. M. Schulz, T. Peterson, C. M. Sapienza, M. Greer, and W. Friedman, "Voice and speech characteristics of persons with Parkinson's disease pre-and post-pallidotomy surgery: Preliminary findings," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 5, 1999, doi: 10.1044/jslhr.4205.1176.

[11] C. Soto and S. Pritzkow, "Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases," *Nature Neuroscience*, vol. 21, no. 10. 2018. doi: 10.1038/s41593-018-0235-9.

[12] C. A. Ross and M. A. Poirier, "Protein aggregation and neurodegenerative disease," *Nat Med*, vol. 10, no. 7, 2004, doi: 10.1038/nm1066.

[13] M. Goedert, "Alzheimer's and Parkinson's diseases: The prion concept in relation to assembled Aβ, tau, and α-synuclein," *Science*, vol. 349, no. 6248. 2015. doi: 10.1126/science.1255555.

[14] B. I. Giasson *et al.*, "Oxidative damage linked to neurodegeneration by selective α-synuclein nitration in synucleinopathy lesions," *Science (1979)*, vol. 290, no. 5493, 2000, doi: 10.1126/science.290.5493.985.

[15] R. Fischer and O. Maier, "Interrelation of oxidative stress and inflammation in neurodegenerative disease: Role of TNF," *Oxidative Medicine and Cellular Longevity*, vol. 2015. 2015. doi: 10.1155/2015/610813.

[16] A. R. Cardoso *et al.*, "Essential genetic findings in neurodevelopmental disorders," *Human genomics*, vol. 13, no. 1. 2019. doi: 10.1186/s40246-019-0216-4.

[17] M. Chin-Chan, J. Navarro-Yepes, and B. Quintanilla-Vega, "Environmental pollutants as risk factors for neurodegenerative disorders: Alzheimer and Parkinson diseases," *Front Cell Neurosci*, vol. 9, no. APR, 2015, doi: 10.3389/fncel.2015.00124.

[18] M. Nabi and N. Tabassum, "Role of Environmental Toxicants on Neurodegenerative Disorders," *Frontiers in Toxicology*, vol. 4. 2022. doi: 10.3389/ftox.2022.837579.

[19] I. Parenti, L. G. Rabaneda, H. Schoen, and G. Novarino, "Neurodevelopmental Disorders: From Genetics to Functional Pathways," *Trends in Neurosciences*, vol. 43, no. 8. 2020. doi: 10.1016/j.tins.2020.05.004.

[20] M. D. Johnson *et al.*, "Neuromodulation for brain disorders: Challenges and opportunities," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 3. 2013. doi: 10.1109/TBME.2013.2244890.

[21] C. Nieder, A. L. Grosu, and L. E. Gaspar, "Stereotactic radiosurgery (SRS) for brain metastases: A systematic review," *Radiation Oncology*, vol. 9, no. 1. 2014. doi: 10.1186/1748-717X-9-155.

[22] T. Yasuda and H. Mochizuki, "The regulatory role of α-synuclein and parkin in neuronal cell apoptosis; Possible implications for the pathogenesis of Parkinson's disease," *Apoptosis*, vol. 15, no. 11, 2010, doi: 10.1007/s10495-010-0486-8.

[23] R. C. Hughes, "Parkinson's Disease and its Management," *BMJ*, vol. 308, no. 6923, 1994, doi: 10.1136/bmj.308.6923.281.

[24] A. Kouli, K. M. Torsney, and W.-L. Kuan, "Parkinson's Disease: Etiology, Neuropathology, and Pathogenesis," in *Parkinson's Disease: Pathogenesis and Clinical Aspects*, 2018. doi: 10.15586/codonpublications.parkinsonsdisease.2018.ch1.

[25] O. M. A. El-Agnaf *et al.*, "Detection of oligomeric forms of α-synuclein protein in human plasma as a potential biomarker for Parkinson's disease," *The FASEB Journal*, vol. 20, no. 3, 2006, doi: 10.1096/fj.03-1449com.

[26] R. C. Chen *et al.*, "Prevalence, incidence, and mortality of PD: A door-to-door survey in Ilan County, Taiwan," *Neurology*, vol. 57, no. 9, 2001, doi: 10.1212/WNL.57.9.1679.

[27] A. H. V. Schapira and J. Obeso, "Timing of treatment initiation in Parkinson's disease: A need for reappraisal?," *Annals of Neurology*, vol. 59, no. 3. 2006. doi: 10.1002/ana.20789.

[28] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift fur Medizinische Physik*, vol. 29, no. 2. 2019. doi: 10.1016/j.zemedi.2018.11.002.

[29] S. Vieira, W. H. L. Pinaya, and A. Mechelli, "Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications," *Neuroscience and Biobehavioral Reviews*, vol. 74. 2017. doi: 10.1016/j.neubiorev.2017.01.002.

[30] M. R. Arbabshirani, S. Plis, J. Sui, and V. D. Calhoun, "Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls," *Neuroimage*, vol. 145, 2017, doi: 10.1016/j.neuroimage.2016.02.079.

[31] E. E. Schadt, M. D. Linderman, J. Sorenson, L. Lee, and G. P. Nolan, "Computational solutions to large-scale data management and analysis," *Nature Reviews Genetics*, vol. 11, no. 9. 2010. doi: 10.1038/nrg2857.

[32] M. A. Myszczynska *et al.*, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nature Reviews Neurology*, vol. 16, no. 8. 2020. doi: 10.1038/s41582-020-0377-8.

[33] M. Kubat, *An Introduction to Machine Learning*. 2021. doi: 10.1007/978-3-030-81935-4.

[34] C. Singh, "Machine Learning in Pattern Recognition," *European Journal of Engineering and Technology Research*, vol. 8, no. 2, 2023, doi: 10.24018/ejeng.2023.8.2.3025.

[35] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS.B*, vol. 4, 2017, doi: 10.20544/horizons.b.04.1.17.p05.

[36] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica (Ljubljana)*, vol. 31, no. 3. 2007.

[37] Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges," *Sensors*, vol. 23, no. 9. 2023. doi: 10.3390/s23094178.

[38] J. Ha, M. Kambe, and J. Pe, *Data Mining: Concepts and Techniques*. 2011. doi: 10.1016/C2009-0-61819-5.

[39] Y. Liu, S. Mazumdar, and P. A. Bath, "An unsupervised learning approach to diagnosing Alzheimer's disease using brain magnetic resonance imaging scans," *Int J Med Inform*, vol. 173, 2023, doi: 10.1016/j.ijmedinf.2023.105027.

[40] T. Fujiwara, J. K. Chou, Shilpika, P. Xu, L. Ren, and K. L. Ma, "An Incremental Dimensionality Reduction Method for Visualizing Streaming Multidimensional Data," *IEEE Trans Vis Comput Graph*, vol. 26, no. 1, 2020, doi: 10.1109/TVCG.2019.2934433.

[41] V. Mittal and R. K. Sharma, "Machine learning approach for classification of Parkinson disease using acoustic features," *J Reliab Intell Environ*, vol. 7, no. 3, 2021, doi: 10.1007/s40860-021-00141-6.

[42] X. Goldberg, "Introduction to semi-supervised learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, 2009, doi: 10.2200/S00196ED1V01Y200906AIM006.

[43] D. J. Joshi, I. Kale, S. Gandewar, O. Korate, D. Patwari, and S. Patil, "Reinforcement Learning: A Survey," in *Advances in Intelligent Systems and Computing*, 2021. doi: 10.1007/978-981-33-4859-2_29.

[44] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, 2015, doi: 10.1038/nature14236.

[45] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, 2020, doi: 10.1016/j.neucom.2019.10.118.

[46] S. A. Kasnavi, M. Aminafshar, M. M. Shariati, N. Emam Jomeh Kashan, and M. Honarvar, "The effect of kernel selection on genome wide prediction of discrete traits by Support Vector Machine," *Gene Rep*, vol. 11, 2018, doi: 10.1016/j.genrep.2018.04.006.

[47] D. Kancherla, J. D. Bodapati, and N. Veeranjaneyulu, "Effect of different kernels on the performance of an SVM based classification," *International Journal of Recent Technology and Engineering*, vol. 7, no. 5, 2019.

[48] N. R. Panda, J. K. Pati, J. N. Mohanty, and R. Bhuyan, "A Review on Logistic Regression in Medical Research," *National Journal of Community Medicine*, vol. 13, no. 4. 2022. doi: 10.55489/njcm.134202222.

[49] S. Le Cessie and J. C. Van Houwelingen, "Ridge Estimators in Logistic Regression," *Appl Stat*, vol. 41, no. 1, 1992, doi: 10.2307/2347628.

[50] Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," in *Proceedings - 2017 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering, ICITISEE 2017*, 2017. doi: 10.1109/ICITISEE.2017.8285514.

[51] J. Han, M. Kamber, and J. Pei, *Data Mining Concept and Techniques, 3rd ed.* 2012.

[52] C. Kingsford and S. L. Salzberg, “What are decision trees?,” *Nature Biotechnology*, vol. 26, no. 9. 2008. doi: 10.1038/nbt0908-1011.

[53] S. B. Kotsiantis, “Decision trees: A recent overview,” *Artificial Intelligence Review*, vol. 39, no. 4. 2013. doi: 10.1007/s10462-011-9272-4.

[54] F. Aaboub, H. Chamlal, and T. Ouaderhman, “Statistical analysis of various splitting criteria for decision trees *,” *J Algorithm Comput Technol*, vol. 17, 2023, doi: 10.1177/17483026231198181.

[55] L. Breiman, “Random forests. Mach. Learn,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, no. 45, 2001.

[56] Y. Amit and D. Geman, “Shape Quantization and Recognition with Randomized Trees,” *Neural Comput*, vol. 9, no. 7, 1997, doi: 10.1162/neco.1997.9.7.1545.

[57] J. R. Quinlan, “Bagging, boosting, and C4.5,” in *Proceedings of the National Conference on Artificial Intelligence*, 1996.

[58] T. Chen and C. Guestrin, “XGBoost : Reliable Large-scale Tree Boosting System,” *ArXiv*, 2016.

[59] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672.2939785.

[60] H. Zhang, “The optimality of Naive Bayes,” in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, 2004.

[61] N. G. Ramadhan and F. D. Adhinata, “Sentiment analysis on vaccine COVID-19 using word count and Gaussian Naïve Bayes,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 3, 2022, doi: 10.11591/ijeecs.v26.i3.pp1765-1772.

[62] I. T. Jollife and J. Cadima, “Principal component analysis: A review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065. 2016. doi: 10.1098/rsta.2015.0202.

[63] L. I. Smith, “A tutorial on Principal Components Analysis Introduction,” *Statistics (Ber)*, 2002.

[64] L. R. Fischer *et al.*, “Amyotrophic lateral sclerosis is a distal axonopathy: Evidence in mice and man,” *Exp Neurol*, vol. 185, no. 2, 2004, doi: 10.1016/j.expneurol.2003.10.004.

[65] M. A. Myszczynska *et al.*, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nature Reviews Neurology*, vol. 16, no. 8. 2020. doi: 10.1038/s41582-020-0377-8.

[66] J. Koikkalainen *et al.*, "Differential diagnosis of neurodegenerative diseases using structural MRI data," *Neuroimage Clin*, vol. 11, 2016, doi: 10.1016/j.nicl.2016.02.019.

[67] A. R. Varma *et al.*, "Diagnostic patterns of regional atrophy on MRI and regional cerebral blood flow change on SPECT in young onset patients with Alzheimer's disease, frontotemporal dementia and vascular dementia," *Acta Neurol Scand*, vol. 105, no. 4, 2002, doi: 10.1034/j.1600-0404.2002.1o148.x.

[68] J. S. Meyer, J. Huang, and M. H. Chowdhury, "MRI confirms mild cognitive impairments prodromal for Alzheimer's, vascular and Parkinson-Lewy body dementias," *J Neurol Sci*, vol. 257, no. 1–2, 2007, doi: 10.1016/j.jns.2007.01.016.

[69] E. J. Burton *et al.*, "Medial temporal lobe atrophy on MRI differentiates Alzheimer's disease from dementia with Lewy bodies and vascular cognitive impairment: A prospective study with pathological verification of diagnosis," *Brain*, vol. 132, no. 1, 2009, doi: 10.1093/brain/awn298.

[70] R. Barber, A. Gholkar, P. Scheltens, C. Ballard, I. G. McKeith, and J. T. O'Brien, "Medial temporal lobe atrophy on MRI in dementia with Lewy bodies," *Neurology*, vol. 52, no. 6, 1999, doi: 10.1212/wnl.52.6.1153.

[71] K. Ishii *et al.*, "Comparison of regional brain volume and glucose metabolism between patients with mild dementia with lewy bodies and those with mild Alzheimer's disease," *Journal of Nuclear Medicine*, vol. 48, no. 5, 2007, doi: 10.2967/jnumed.106.035691.

[72] M. Á. Muñoz-Ruiz *et al.*, "Structural MRI in Frontotemporal Dementia: Comparisons between Hippocampal Volumetry, Tensor-Based Morphometry and Voxel-Based Morphometry," *PLoS One*, vol. 7, no. 12, 2012, doi: 10.1371/journal.pone.0052531.

[73] R. Barber, C. Ballard, I. G. McKeith, A. Gholkar, and J. T. O'Brien, "MRI volumetric study of dementia with Lewy bodies: A comparison with AD and vascular dementia," *Neurology*, vol. 54, no. 6, 2000, doi: 10.1212/WNL.54.6.1304.

[74] G. B. Frisoni *et al.*, "Hippocampal and entorhinal cortex atrophy in frontotemporal dementia and Alzheimer's disease," *Neurology*, vol. 52, no. 1, 1999, doi: 10.1212/wnl.52.1.91.

[75] S. Klöppel *et al.*, "Automatic classification of MR scans in Alzheimer's disease," *Brain*, vol. 131, no. 3, 2008, doi: 10.1093/brain/awm319.

[76] G. D. Rabinovici *et al.*, "Distinct MRI atrophy patterns in autopsy-proven Alzheimer's disease and frontotemporal lobar degeneration," *Am J Alzheimers Dis Other Demen*, vol. 22, no. 6, 2008, doi: 10.1177/1533317507308779.

[77] J. L. Whitwell *et al.*, “Focal atrophy in dementia with Lewy bodies on MRI: A distinct pattern from Alzheimer’s disease,” *Brain*, vol. 130, no. 3, 2007, doi: 10.1093/brain/awl388.

[78] M. Ballmaier *et al.*, “Comparing gray matter loss profiles between dementia with Lewy bodies and Alzheimer’s disease using cortical pattern matching: Diagnosis and gender effects,” *Neuroimage*, vol. 23, no. 1, 2004, doi: 10.1016/j.neuroimage.2004.04.026.

[79] R. Barber, I. McKeith, C. Ballard, and J. O’Brien, “Volumetric MRI study of the caudate nucleus in patients with dementia with Lewy bodies, Alzheimer’s disease, and vascular dementia,” *J Neurol Neurosurg Psychiatry*, vol. 72, no. 3, 2002, doi: 10.1136/jnnp.72.3.406.

[80] E. J. Burton *et al.*, “Patterns of cerebral atrophy in Dementia with Lewy bodies using voxel-based morphometry,” *Neuroimage*, vol. 17, no. 2, 2002, doi: 10.1016/S1053-8119(02)91197-3.

[81] M. P. Laakso *et al.*, “Hippocampus and entorhinal cortex in frontotemporal dementia and Alzheimer’s disease: A morphometric MRI study,” *Biol Psychiatry*, vol. 47, no. 12, 2000, doi: 10.1016/S0006-3223(99)00306-6.

[82] G. Singh, M. Vadera, L. Samavedham, and E. C. H. Lim, “Machine Learning-Based Framework for Multi-Class Diagnosis of Neurodegenerative Diseases: A Study on Parkinson’s Disease,” in *IFAC-PapersOnLine*, 2016. doi: 10.1016/j.ifacol.2016.07.331.

[83] Z. S. Nasreddine *et al.*, “The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment,” *J Am Geriatr Soc*, vol. 53, no. 4, 2005, doi: 10.1111/j.1532-5415.2005.53221.x.

[84] T. W. Frazier *et al.*, “Development of webcam-collected and artificial-intelligence-derived social and cognitive performance measures for neurodevelopmental genetic syndromes,” *Am J Med Genet C Semin Med Genet*, vol. 193, no. 3, 2023, doi: 10.1002/ajmg.c.32058.

[85] L. Ricciardi *et al.*, “Freezing of gait in Parkinson’s disease: The paradoxical interplay between gait and cognition,” *Parkinsonism Relat Disord*, vol. 20, no. 8, 2014, doi: 10.1016/j.parkreldis.2014.04.009.

[86] M. Montero-Odasso *et al.*, “Motor Phenotype in Neurodegenerative Disorders: Gait and Balance Platform Study Design Protocol for the Ontario Neurodegenerative Research Initiative (ONDRI),” *Journal of Alzheimer’s Disease*, vol. 59, no. 2, 2017, doi: 10.3233/JAD-170149.

[87] D. H. Aslan, M. E. Hernandez, M. L. Frechette, A. T. Gephart, I. M. Soloveychik, and J. J. Sosnoff, “The neural underpinnings of motor learning in people with neurodegenerative diseases: A scoping review,” *Neuroscience and Biobehavioral Reviews*, vol. 131. 2021. doi: 10.1016/j.neubiorev.2021.10.006.

[88] R. A. Bradshaw and P. D. Stahl, *Encyclopedia of Cell Biology*, vol. 1–4. 2015. doi: 10.1016/c2019-1-02520-x.

[89] F. D. Mast, A. V. Ratushny, and J. D. Aitchison, “Systems cell biology,” *Journal of Cell Biology*, vol. 206, no. 6. 2014. doi: 10.1083/jcb.201405027.

[90] P. Garg and P. Srivastava, “Decoding Transcriptomics of Neurodevelopmental Disorders: A Computational Approach,” in *Integrative Approaches to Biotechnology*, 2023. doi: 10.1201/9781003324706-18.

[91] J. Dekker *et al.*, “Web-accessible application for identifying pathogenic transcripts with RNA-seq: Increased sensitivity in diagnosis of neurodevelopmental disorders,” *Am J Hum Genet*, vol. 110, no. 2, 2023, doi: 10.1016/j.ajhg.2022.12.015.

[92] C. Mertes *et al.*, “Detection of aberrant splicing events in RNA-seq data using FRASER,” *Nat Commun*, vol. 12, no. 1, 2021, doi: 10.1038/s41467-020-20573-7.

[93] L. Wang and L. Zhang, “Circulating Exosomal miRNA as Diagnostic Biomarkers of Neurodegenerative Diseases,” *Frontiers in Molecular Neuroscience*, vol. 13. 2020. doi: 10.3389/fnmol.2020.00053.

[94] K. Blennow and H. Zetterberg, “The past and the future of Alzheimer’s disease CSF biomarkers-a journey toward validated biochemical tests covering the whole spectrum of molecular events,” *Frontiers in Neuroscience*, vol. 9, no. SEP. 2015. doi: 10.3389/fnins.2015.00345.

[95] S. Rastogi *et al.*, “The evolving landscape of exosomes in neurodegenerative diseases: Exosomes characteristics and a promising role in early diagnosis,” *International Journal of Molecular Sciences*, vol. 22, no. 1. 2021. doi: 10.3390/ijms22010440.

[96] N. Mattsson *et al.*, “CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment,” *JAMA*, vol. 302, no. 4, 2009, doi: 10.1001/jama.2009.1064.

[97] G. J. Patti, O. Yanes, and G. Siuzdak, “Innovation: Metabolomics: the apogee of the omics trilogy,” *Nature Reviews Molecular Cell Biology*, vol. 13, no. 4. 2012. doi: 10.1038/nrm3314.

[98] A. Zhang, H. Sun, and X. Wang, “Serum metabolomics as a novel diagnostic approach for disease: A systematic review,” *Analytical and Bioanalytical Chemistry*, vol. 404, no. 4. 2012. doi: 10.1007/s00216-012-6117-1.

[99] Y. Chen, E. M. Li, and L. Y. Xu, “Guide to Metabolomics Analysis: A Bioinformatics Workflow,” *Metabolites*, vol. 12, no. 4. 2022. doi: 10.3390/metabo12040357.

[100] G. A. N. Gowda, S. Zhang, H. Gu, V. Asiago, N. Shanaiah, and D. Raftery, “Metabolomics-based methods for early disease diagnostics,” *Expert Review of Molecular Diagnostics*, vol. 8, no. 5. 2008. doi: 10.1586/14737159.8.5.617.

[101] C. B. Santos-Rebouças, J. Cordovil Cotrin, and G. C. dos Santos Junior, “Exploring the interplay between metabolomics and genetics in Parkinson’s disease: Insights from ongoing research and future avenues,” *Mech Ageing Dev*, vol. 216, 2023, doi: 10.1016/j.mad.2023.111875.

[102] L. Hu *et al.*, “Integrated Metabolomics and Proteomics Analysis Reveals Plasma Lipid Metabolic Disturbance in Patients With Parkinson’s Disease,” *Front Mol Neurosci*, vol. 13, 2020, doi: 10.3389/fnmol.2020.00080.

[103] H. H. Feldman *et al.*, “Treatment with galantamine and time to nursing home placement in Alzheimer’s disease patients with and without cerebrovascular disease,” *Int J Geriatr Psychiatry*, vol. 24, no. 5, 2009, doi: 10.1002/gps.2141.

[104] E. R. Siemers *et al.*, “Phase 3 solanezumab trials: Secondary outcomes in mild Alzheimer’s disease patients,” *Alzheimer’s and Dementia*, vol. 12, no. 2, 2016, doi: 10.1016/j.jalz.2015.06.1893.

[105] F. Falahati, E. Westman, and A. Simmons, “Multivariate data analysis and machine learning in Alzheimer’s disease with a focus on structural magnetic resonance imaging,” *Journal of Alzheimer’s Disease*, vol. 41, no. 3. 2014. doi: 10.3233/JAD-131928.

[106] F. L. Darley, A. E. Aronson, and J. R. Brown, “Differential diagnostic patterns of dysarthria.,” *J Speech Hear Res*, vol. 12, no. 2, 1969, doi: 10.1044/jshr.1202.246.

[107] P. Gillivan-Murphy, P. Carding, and N. Miller, “Vocal tract characteristics in Parkinson’s disease,” *Current Opinion in Otolaryngology and Head and Neck Surgery*, vol. 24, no. 3. 2016. doi: 10.1097/MOO.0000000000000252.

[108] S. Yang *et al.*, “The physical significance of acoustic parameters and its clinical significance of dysarthria in Parkinson’s disease,” *Sci Rep*, vol. 10, no. 1, 2020, doi: 10.1038/s41598-020-68754-0.

[109] A. P. Kiraly *et al.*, “Deep convolutional encoder-decoders for prostate cancer detection and classification,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017. doi: 10.1007/978-3-319-66179-7_56.

[110] J. Zhang, A. Saha, Z. Zhu, and M. A. Mazurowski, “Hierarchical Convolutional Neural Networks for Segmentation of Breast Tumors in MRI With Application to Radiogenomics,” *IEEE Trans Med Imaging*, vol. 38, no. 2, 2019, doi: 10.1109/TMI.2018.2865671.

[111] R. Cao *et al.*, “Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet,” *IEEE Trans Med Imaging*, vol. 38, no. 11, 2019, doi: 10.1109/TMI.2019.2901928.

[112] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, “Computer-aided detection of prostate cancer in MRI,” *IEEE Trans Med Imaging*, vol. 33, no. 5, 2014, doi: 10.1109/TMI.2014.2303821.

[113] M. EL-Geneedy, H. E. D. Moustafa, F. Khalifa, H. Khater, and E. AbdElhalim, “An MRI-based deep learning approach for accurate detection of Alzheimer’s disease,” *Alexandria Engineering Journal*, vol. 63, 2023, doi: 10.1016/j.aej.2022.07.062.

[114] M. Tsuneki, “Deep learning models in medical image analysis,” *Journal of Oral Biosciences*, vol. 64, no. 3. 2022. doi: 10.1016/j.job.2022.03.003.

[115] C. Li, M. L. Dubbelaar, X. Zhang, and J. C. Zheng, “Editorial: Understanding the heterogeneity and spatial brain environment of neurodegenerative diseases through conventional and future methods,” *Frontiers in Cellular Neuroscience*, vol. 17. 2023. doi: 10.3389/fncel.2023.1211273.

[116] R. Duara and W. Barker, “Heterogeneity in Alzheimer’s Disease Diagnosis and Progression Rates: Implications for Therapeutic Trials,” *Neurotherapeutics*, vol. 19, no. 1. 2022. doi: 10.1007/s13311-022-01185-z.

[117] F. Albrecht *et al.*, “Unraveling Parkinson’s disease heterogeneity using subtypes based on multimodal data,” *Parkinsonism Relat Disord*, vol. 102, 2022, doi: 10.1016/j.parkreldis.2022.07.014.

[118] D. Bzdok, N. Altman, and M. Krzywinski, “Points of Significance: Statistics versus machine learning,” *Nature Methods*, vol. 15, no. 4. 2018. doi: 10.1038/nmeth.4642.

[119] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, “Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment,” *Medicina (Lithuania)*, vol. 56, no. 9, 2020, doi: 10.3390/medicina56090455.

[120] M. F. Bergeron, S. Landset, F. Tarpin-Bernard, C. B. Ashford, T. M. Khoshgoftaar, and J. W. Ashford, “Episodic-Memory Performance in Machine Learning Modeling for Predicting Cognitive Health Status Classification,” *Journal of Alzheimer’s Disease*, vol. 70, no. 1, 2019, doi: 10.3233/JAD-190165.

[121] Y. M. Choe, M. S. Byun, J. H. Lee, B. K. Sohn, D. Y. Lee, and J. W. Kim, “Subjective memory complaint as a useful tool for the early detection of Alzheimer’s disease,” *Neuropsychiatr Dis Treat*, vol. 14, 2018, doi: 10.2147/NDT.S174517.

[122] R. Sadeghian, J. David Schaffer, and S. A. Zahorian, “Speech processing approach for diagnosing dementia in an early stage,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017. doi: 10.21437/Interspeech.2017-1712.

[123] G. Gagliardi, “Natural language processing techniques for studying language in pathological ageing: A scoping review,” *International Journal of Language and Communication Disorders*. 2023. doi: 10.1111/1460-6984.12870.

[124] S. Mezrar and F. Bendella, “Machine learning and Serious Game for the Early Diagnosis of Alzheimer’s Disease,” *Simul Gaming*, vol. 53, no. 4, 2022, doi: 10.1177/10468781221106850.

[125] A. L. Young *et al.*, “Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference,” *Nat Commun*, vol. 9, no. 1, 2018, doi: 10.1038/s41467-018-05892-0.

[126] K. M. Giannakopoulou, I. Roussaki, and K. Demestichas, “Internet of Things Technologies and Machine Learning Methods for Parkinson’s Disease Diagnosis, Monitoring and Management: A Systematic Review,” *Sensors*, vol. 22, no. 5. 2022. doi: 10.3390/s22051799.

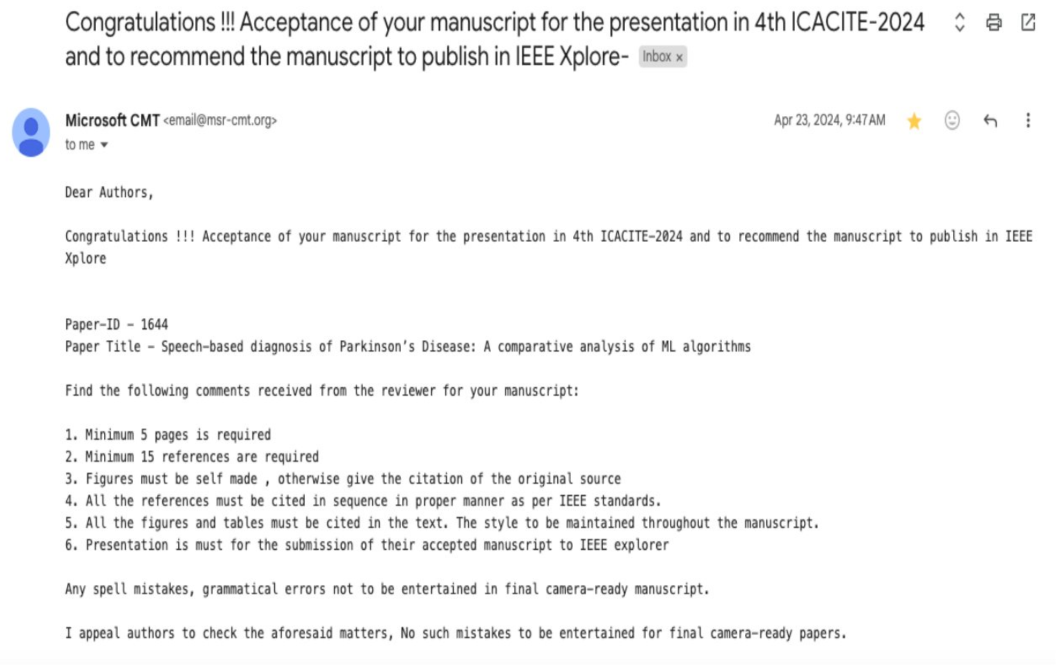
[127] S. A. Siddiqui, A. Ahmad, and N. Fatima, “IoT-based disease prediction using machine learning,” *Computers and Electrical Engineering*, vol. 108, 2023, doi: 10.1016/j.compeleceng.2023.108675.

[128] A. Baerheim, “The diagnostic process in general practice: Has it a two-phase structure?,” *Family Practice*, vol. 18, no. 3. 2001. doi: 10.1093/fampra/18.3.243.

[129] F. De Marchi, E. Contaldi, L. Magistrelli, R. Cantello, C. Comi, and L. Mazzini, “Telehealth in neurodegenerative diseases: Opportunities and challenges for patients and physicians,” *Brain Sci*, vol. 11, no. 2, 2021, doi: 10.3390/brainsci11020237.

[130] R. I. Griffiths *et al.*, “Automated assessment of bradykinesia and dyskinesia in Parkinson’s disease,” *J Parkinsons Dis*, vol. 2, no. 1, 2012, doi: 10.3233/JPD-2012-11071.

PROOF OF ACCEPTANCE



Congratulations !!! Acceptance of your manuscript for the presentation in 4th ICACITE-2024 and to recommend the manuscript to publish in IEEE Xplore- Inbox ×

Microsoft CMT <email@msr-cmt.org>
to me

Apr 23, 2024, 9:47 AM

Dear Authors,

Congratulations !!! Acceptance of your manuscript for the presentation in 4th ICACITE-2024 and to recommend the manuscript to publish in IEEE Xplore

Paper-ID - 1644
Paper Title - Speech-based diagnosis of Parkinson's Disease: A comparative analysis of ML algorithms

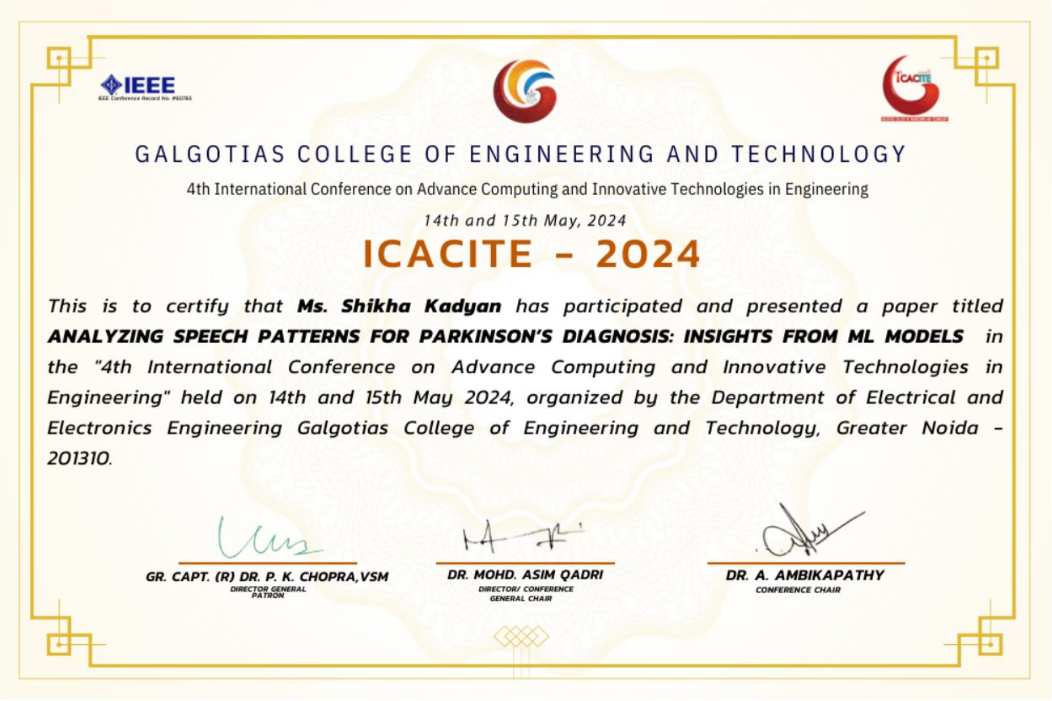
Find the following comments received from the reviewer for your manuscript:

1. Minimum 5 pages is required
2. Minimum 15 references are required
3. Figures must be self made , otherwise give the citation of the original source
4. All the references must be cited in sequence in proper manner as per IEEE standards.
5. All the figures and tables must be cited in the text. The style to be maintained throughout the manuscript.
6. Presentation is must for the submission of their accepted manuscript to IEEE explorer

Any spell mistakes, grammatical errors not to be entertained in final camera-ready manuscript.

I appeal authors to check the aforesaid matters, No such mistakes to be entertained for final camera-ready papers.

CERTIFICATE OF CONFERENCE



IEEE
IEEE Conference Record No. #60783

GALGOTIAS COLLEGE OF ENGINEERING AND TECHNOLOGY

4th International Conference on Advance Computing and Innovative Technologies in Engineering

14th and 15th May, 2024

ICACITE - 2024

*This is to certify that **Ms. Shikha Kadyan** has participated and presented a paper titled **ANALYZING SPEECH PATTERNS FOR PARKINSON'S DIAGNOSIS: INSIGHTS FROM ML MODELS** in the "4th International Conference on Advance Computing and Innovative Technologies in Engineering" held on 14th and 15th May 2024, organized by the Department of Electrical and Electronics Engineering Galgotias College of Engineering and Technology, Greater Noida - 201310.*

GR. CAPT. (R) DR. P. K. CHOPRA,VSM
DIRECTOR GENERAL
PATRON

DR. MOHD. ASIM QADRI
DIRECTOR/ CONFERENCE
GENERAL CHAIR

DR. A. AMBIKAPATHY
CONFERENCE CHAIR

ACCEPTED IEEE PAPER

Analyzing speech patterns for Parkinson's diagnosis: Insights from ML models

Shikha Kadyan
Molecular Neuroscience and Functional Genomics Laboratory, Dept. of Biotechnology
Delhi Technological University
New Delhi-110042, India
shikhakadyan07@gmail.com

Pravir Kumar
Molecular Neuroscience and Functional Genomics Laboratory, Dept. of Biotechnology
Delhi Technological University
New Delhi-110042, India
pravirkumar@dtu.ac.in

***Abstract*—Parkinson's disease (PD) is the most ubiquitous neurological disease in the globe which affects the human neurological system. It is a primary concern to detect PD in its early stages to slow down its progress with proper treatment and foster a better quality of life for affected individuals. PD primarily impacts motor and cognitive function, recent studies also revealed that 90% of Parkinson's patients encounter speech difficulties in the preliminary phase of the disease. In the framework of this study, we have conducted a comparative analysis of various machine learning (ML) models, including Support Vector Machine (SVM), Gaussian Naïve Bayes (GNB), K-Nearest Neighbors (KNN), Decision Tree (DT), Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGB), and Principal Component Analysis (PCA), for the precise identification of PD in early phase using a speech dataset. Three different approaches are employed for classification. The XGB model performed remarkably well, with 98.30% accuracy and 96.67% precision. The results hold significant promise for enhancing early-stage PD diagnosis in healthcare centres as well as within the home environment.**

Keywords— Speech recognition, machine learning models, telemedicine, SVM, KNN, XGB, GNB, hyperparameters, Principal component analysis, ROC-AUC.

I. Introduction

The distinctive characteristic of Parkinson's disease (PD) is the steady decrease of neuronal cells that produce dopamine in the substantia nigra (Snp), located in the midbrain. These neurons perform a vital function in coordinating movement at the muscular level [1]. A decline in their number results in dopamine deficiency in the basal ganglia, causing impaired motor functions such as resting tremors, stiffness of the muscles, bradykinesia and imbalanced posture [2]. Because symptoms and disease progression of the disease differ, PD may remain undiagnosed for an extended period. Although a cure for PD is presently unavailable, treatments focus on managing symptoms and elevating the overall quality of life for patients. Therefore, the early and accurate diagnosis of PD holds profound significance in facilitating timely intervention and support for patients. New approaches are required for PD diagnosis. Hence, cost-effective, straightforward, and credible methods ought to be employed for accurate diagnosis and treatment assurance. Machine learning (ML) models are used to categorise individuals as either as healthy individuals or PD patients. Research indicated that evaluating vocal irregularities can serve as an indicator for early PD detection [3]. Reduced intensity, pitch, monotonous loudness, tense quiet, rapid speech bursts, ambiguous consonant enunciation, erratic tempo, and dysphonia, particularly characterised by hoarse and hushed voices, constitute typical speech impairments in PD [4]. According to reports, in the earliest phases of the disease, voice and speech difficulties impact about 90% of PD patients [5]. Early vocal cord impairment detection and monitoring in PD can be beneficial owing to its fairly simple measurement and the possibility of remote evaluation. Considering this, it would be appropriate for applications involving telemedicine or remote healthcare [6]. Timely and efficient diagnosis of PD can be achieved by listening and drawing conclusions from audio recordings in case medical intervention is not possible on the spot, as in the case of remote and rural areas. This can be achieved by healthcare experts or ML-based systems by analysing recordings captured via digital devices and smartphone applications.

In order to catch PD in the preliminary stage, this study focuses on evaluating various ML models by analyzing a dataset of 195 voice recordings of different subjects. This would ensure that the patient receives an appropriate treatment plan for a better prognosis and quality of life. The findings of this study unveil that in terms of accuracy of performance, other models failed to beat the Extreme Gradient Boosting (XGB) model displaying an exceptional accuracy of 98.30% post-training on 22 distinct attributes from the over-sampled voice data.

II. Methodology

The reservoir of audio recording used in this methodology were the ML archive of the University of California, Irvine (UCI) and Parkinson's Progression Markers Initiative (PPMI). The focus was directed towards the identification of variations lying in the voice attributes of PD patients. The dataset comprised 22 peculiar voice attributes such as fundamental frequency, shimmer, jitter, multi-dimensional voice programmer (MDVP) measurements, dysphonia, and many more. To have an elaborate understanding of the dataset pre-processing was performed. Random Forest Classifier (RF), Gaussian Naïve Bayes classifier (GNB), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGB and Decision Tree (DT) were the seven distinct ML models subjected to training using 80% of the voice recordings to spot the distinctions in voice attributes when compared between PD patients and healthy individuals, frequency variation being the most reliable factor. Following training, the models underwent testing and performance evaluation.

This research aims to determine the crucial factors in classifying individuals with Parkinson's and assess how the data imbalances impact medical classification. To achieve this aim, three different approaches were employed: Training the models on the all the attributes of the dataset to establish a baseline for classification, Over-sampling the dataset followed by training the models and Training the models on the major five attributes identified through PCA (Principal Component Analysis). Three approaches are delineated as follows:

Approach 1: Training the models on 22 data-derived attributes

- Collection of audio recording data from PPMI or UCI databases
- Conduct data analysis to identify outliers, skewness, duplicate entries, missing values, and variable distribution patterns. Removal of columns named "Name" and "Status"
- Split dataset into testing and training sets, with 80% designated for training
- Apply StandardScaler from the scikit-learn library to standardize the features of data
- Train the various models including SVM, LR, DT, RF, KNN, GNB, and XGB using the data
- Examine classification outcomes by analyzing the confusion matrix, scoring metrics, and ROC-AUC curve

Approach 2: Over-sampling the dataset

- Collection of audio recording data from PPMI or UCI databases
- Conduct data analysis to identify outliers, skewness, duplicate entries, missing values, and variable distribution patterns. Removal of columns named "Name" and "Status"
- The dataset exhibits an imbalance, comprising 147 records of individuals with PD and 48 records of normal individuals. To mitigate this imbalance, employ SMOTE (Synthetic Minority Over-sampling Technique) from the imlearn library, oversampling both classes to 294 records each
- Split dataset into testing and training sets, with 80% designated for training
- Apply StandardScaler from the scikit-learn library to standardize the features of data
- Train the various models including SVM, LR, DT, RF, KNN, GNB, and XGB using the data
- Examine classification outcomes by analyzing the confusion matrix, scoring metrics, and ROC-AUC curve

Approach 3: Model training on five key attributes identified by PCA algorithm

- Collection of audio recording data from PPMI or UCI databases
- Conduct data analysis to identify outliers, skewness, duplicate entries, missing values, and variable distribution patterns. Removal of columns named "Name" and "Status"
- Apply PCA to extract the five most significant features out of all for training the models
- Split dataset as 20% testing and 80% training sets.
- Apply StandardScaler from the scikit-learn library to standardize the features of data
- Train the various models including SVM, LR, DT, RF, KNN, GNB, and XGB using the data
- Examine classification outcomes by analyzing the confusion matrix and scoring metrics

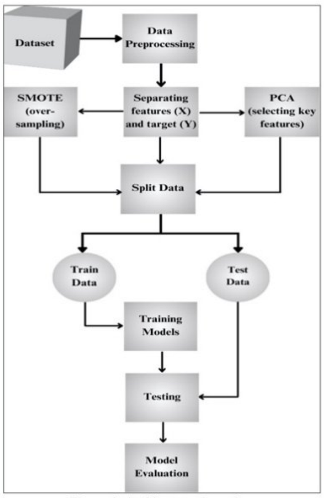


Figure 1. Architecture Overview

Fig. 1 demonstrates an outline of the standard methodology employed. It showcases the steps involved in extracting data from the PPMI or UCI database, dividing the data into two subsets one as test and another as train, training seven models using the data, verifying the outcomes using the test data and finally evaluating models using scoring metrics.

A. Data Collection

The dataset is easily accessible and can be procured via the ML archive of UCI (https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/parkinsons.data) or the PPMI website encompassing diverse biomedical voice metrics gathered from 31 individuals, of whom 23 are diagnosed with PD [7]. Every entry in the dataset complies to a distinct voice recording, identified by the "name" column, while the columns delineate specific vocal measurements, which include 22 different attributes. The main objective is to make a distinction among a population pool comprising individuals who are healthy and those affected with PD, with the "status" column representing 0 as healthy and 1 as PD. The data, formatted in ASCII CSV, comprises a total of 195 voice recordings, with an average of approximately six recordings per individual.

B. Data Processing

Data processing involves cleansing the data, managing missing attributes and dropping redundant columns within the dataset. After that correlation of different features was observed using a correlation heat map. Subsequently, the target and features were separated followed by segmenting the data into the test dataset and the training dataset. The target variable indicated the status i.e., whether an individual had PD or not and features comprised all the attributes except for name and status. Finally, the dataset underwent standardization using the standard scaler from sklearn library.

C. Model Training

The model undergoes training employing a spectrum of ML algorithms.

Support Vector Machine (SVM)

SVM discerns optimal hyperplane that efficiently separates data points associated with separate groups, as illustrated in Fig. 2. SVM optimizes the margin between these distinct classes, rendering it resilient to outliers and proficient in managing datasets with high dimensionality .

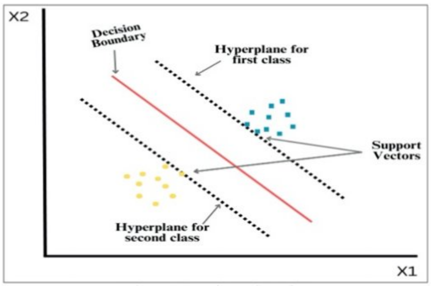


Figure 2. SVM hyperplane plot

In addressing a training dataset characterized by a nonlinear decision surface, as exemplified by PD voice data, SVM strategically employs the kernel technique, a method that entails transforming the data into a feature space with high dimensions [8]. In this space, a linear equation effectively delineates the distinct classes.

Logistic Regression for classification (LR)

LR is mainly used for binary classification. Unlike linear regression, it predicts class probabilities using the logistic function employed to input features [9]. The graphs for linear regression and logistic regression have been illustrated below in Fig. 3. The resulting S-shaped curve determines class assignments based on a threshold, typically 0.5. If the anticipated probability exceeds this threshold, the occurrence is allocated to class 1; or else, it's allocated to class 0. Ideal for audio data, this approach is well-suited because the attributes influencing the classification of PD do not exhibit linear correlation but instead adhere to an exponential pattern.

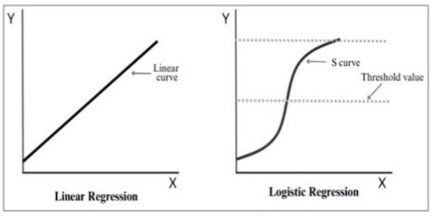


Figure 3. Comparison between the curve of Linear Regression and Logistic Regression classifier

K-Nearest Neighbors (KNN)

KNN, an ML algorithm, classifies through leveraging the closest neighbours' majority vote from the training dataset. Distance, often Euclidean, is calculated between the input point and all training data [10]. It excels with a well-balanced audio dataset containing 109 records, attributed to its compact size. Efficiently, 2 clusters are formed one for PD and another for healthy data. Being a sluggish learning system, KNN avoids preconceptions, enabling learning novel patterns from the training data.

Decision Tree Classifier (DT)

The decision tree classifier, an ML algorithm, decides outcomes using input features. It recursively divides datasets into subsets, as shown in Fig. 4,choosing the optimal feature at each node for maximum information gain or minimum impurity [11]. This iterative process halts upon meeting a specified condition, forming a tree structure. The end nodes, or leaves, signify the final classification results.

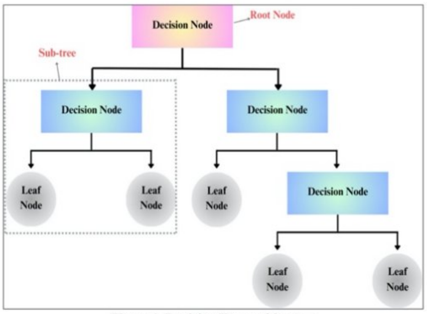


Figure 4. Decision Tree architecture

Random Forest Classifier (RF)

RF classifier, a well-known ensemble learning method, generates multiple decision trees to ascertain the most common class prediction. Every tree is formed by harnessing a subset of the dataset along with random features, thus boosting the model's proficiency to generalize [12]. As the forest accumulates more trees, its performance boosts, simultaneously retaining robustness against outliers and reducing susceptibility to overfitting, thereby setting it apart from individual decision trees.

Gaussian Naïve Bayes Classifier (GNB)

GNB classifier, a probabilistic model within ML, follows the Gaussian (bell curve) distribution and relies on the feature independence hypothesis. Given the class label, it is presumed that all the features standalone from one another to determine the class of a data point, the GNB classifier first computes how likely the data point corresponds to each class, and then chooses the most probable one [13]. The classifier's simplicity, scalability, high computational efficiency and probability-based framework high computational efficiency make it beneficial for datasets with numerous features.

Extreme Gradient Boosting Classifier (XGB)

Xgboost, a powerful ensemble learning algorithm, is renowned for its exceptional performance accuracy for classification purposes. It works by orderly constructing DTs in a series fashion and subsequently identifies and addresses errors of predecessors using a technique known as gradient descent optimization. This results in fine-tuning a differential loss function so that errors can be minimised throughout the training phase [14]. It then clusters all these DTs together to build a potent predictive model. Regularization terms, along with pruning and learning rate bolster XGB to mitigate overfitting problems. High speed, accuracy, streamlined parallel processing, proficiency in handling large datasets, and missing values are a few other peculiar features of this algorithm.

Principal Component analysis (PCA)

PCA reduces the dimensionality of a data in order to simplify the data while retaining vital features [15]. The method by which it works is "orthogonal transformation" which utilises the original collection of features to create a new set of orthogonal features known as component features. These component features are the linear combinations of original features, ranked based on the variance they account for within the data. PCA optimizes computational efficiency, diminishes redundancy, and safeguards essential information crucial for training.

D. Model Evaluation

Following model training on the dataset, the models were assessed to gauge their performance. Accuracy, confusion matrix, F1 score, precision, Receiver Operating Characteristic (ROC) curve and recall, are the various scoring metrics opted to compare different models and make informed decisions about model selection and tuning. Equations 1-5 depict the formulas for these scoring metrics. The equation of the metrics involves TP (True Positives), TN (True Negatives), FN (False Negatives), and FP (False Positives).

$$\text{Precision} = \frac{TP}{TP+FP} \tag{1}$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{2}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{4}$$

$$\text{F1} = 2\frac{Precision.Recall}{Precision+Recall} = \frac{2TP}{2TP+FP+FN} \tag{5}$$

Accuracy assesses the ratio of accurate predictions, with higher values indicating superior overall predictive performance. Specificity gauges the proportion of accurately predicted actual negatives, reflecting the ability to identify normal individuals. Precision signifies the relevance of predicted positives. The F1 score combines precision and recall in a single value, ranging between 0 and 1, high score signifies high model accuracy. The ROC curve visually depicts the probability curve, and the extent beneath this curve is determined by the Area under the Curve (AUC), serving as a quantification of the algorithm's effectiveness.

III. Result and Discussion

The dataset underwent segmentation into training and testing sets, adhering to an 8:2 ratio, where 80% constituted the training dataset and the remaining constituted the testing dataset. Then it was standardized using a standard scaler before performing model training.

Precision, accuracy, recall, and F1-score were utilized as scoring metrics for optimal model selection.

TABLE I. RESULTS OF APPROACH 1 SCORING METRICS

Metric	**ML models**						
	SVM	***LR***	***DT***	***RF***	***KNN***	***GNB***	***XGB***
Accuracy	0.949	0.872	0.795	0.974	0.923	0.692	0.949
Precision	0.964	0.871	0.857	0.966	0.963	0.944	0.933
Recall	0.964	0.964	0.857	1.000	0.939	0.607	1.000
F1- score	0.964	0.915	0.857	0.982	0.945	0.739	0.966

The results obtained from approach 1 are depicted above in Table I. RF classifier, a model composed of multiple decision trees, demonstrated outstanding results, attaining a 97.43% accuracy rate and the highest precision of 96.55%, surpassing the efficacy of other models. Hyperparameter tuning was applied to optimize the RF model. The optimal settings for the RF model included parameters like 'auto' for maximum features, 225 estimators, maximum depth of 8, and 'entropy' as the criterion.

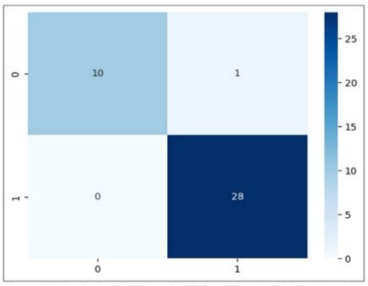


Figure 5. Confusion Matrix of RF classifier model in approach 1

Fig. 5 illustrates the confusion matrix, showing 28 true positives (Parkinson's patients), 10 true negatives (non-Parkinson's), 1 false positive, and 0 false negatives.

Fig. 6 displays the ROC AUC curves of all models, where a larger area indicates superior performance.

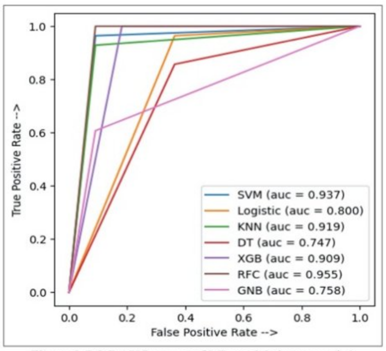


Figure 6. ROC-AUC curves of ML models in approach 1

The outcomes from approach 2 are presented in Table II below. An equal amount of data recordings from PD patients as well as unaffected individuals were employed to train the models. This approach of balancing assures that both groups receive fair consideration in the model's learning process.

TABLE II. RESULTS OF APPROACH 2: BALANCED DATASET

Metric	**ML models**						
	SVM	***LR***	***DT***	***RF***	***KNN***	***GNB***	***XGB***
Accuracy	0.966	0.864	0.898	0.949	0.932	0.864	0.983
Precision	1.000	0.889	0.925	0.933	1.000	0.957	0.967
Recall	0.931	0.827	0.862	0.966	0.862	0.759	1.000
F1- score	0.964	0.857	0.893	0.949	0.926	0.864	0.983

The XGB classifier model showcased remarkable performance on the balanced dataset, attaining an impressive accuracy of 98.30% alongside the highest recall score of 1.00, surpassing the effectiveness of other models. The ideal configuration for the XGB classifier included a learning rate of 0.5, a random state set to 300, and a maximum depth of 5. During the classification process, every attribute is equally vital.

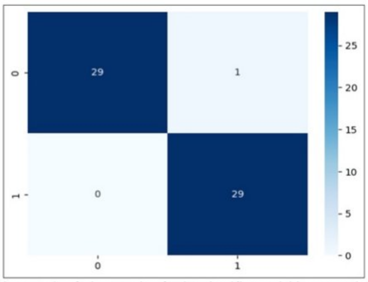


Figure 7. Confusion Matrix of XGB classifier model in approach 2

Fig. 7 presents the confusion matrix of this model, wherein the model classifies 29 true positives (Patient with Parkinson's), 29 true negatives (no PD), 1 false positive and 0 false negatives. In Fig. 8, All of the models' ROC AUC curves are shown; larger areas denotes greater efficiency.

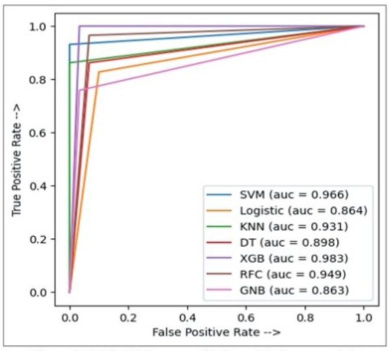


Figure 8. ROC-AUC curves of ML models in approach 2

The results obtained from approach 3 are showcased in Table III. In this particular approach, the PCA algorithm is leveraged to extract five most significant attributes. Subsequently, models are trained and evaluated based on these five attributes, leading to the following outcomes:

TABLE III. RESULTS OF APPROACH 3 SCORING METRICS

Metric	**ML models**						
	SVM	***LR***	***DT***	***RF***	***KNN***	***GNB***	***XGB***
Accuracy	0.897	0.872	0.923	0.974	0.949	0.872	0.974
Precision	0.962	0.871	0.963	0.965	0.964	0.960	0.965
Recall	0.893	0.964	0.929	1.000	0.964	0.857	1.000
F1- score	0.926	0.915	0.945	0.982	0.964	0.906	0.982

The RF and XGB classifier models emerge as the top performers in approach 3, displaying impressive accuracy rates of 97.43% and precision scores of 96.55%. This underscores the effectiveness of utilizing a reduced set of features, highlighting the efficiency of feature selection and dimensionality reduction techniques in the modelling phase.

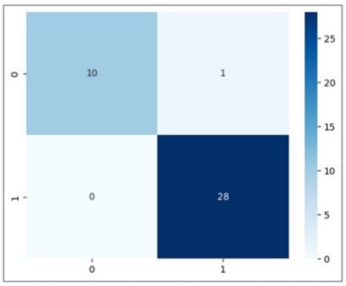


Figure 9. Confusion Matrix of RF and XGB classifier model in approach 3

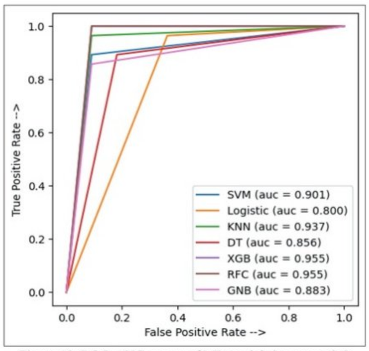


Figure 10. ROC-AUC curves of ML models in approach 3

The confusion matrix of both the RF classifier model as well as the XGB classifier model in approach 3 was the same and is depicted by Fig. 9, wherein the model classifies 28 true positives (Patient with Parkinson's), 10 true negatives (no PD), 1 false positive and 0 false negatives. In Fig. 10, the ROC curves of every model in approach 3 are depicted, with larger areas denoting better performance. The ROC curves of both the RF and XGB classifier models overlap, showcasing the highest AUC values.

IV. Conclusion

The automated ML algorithms allow for PD detection in the preliminary phase with the highest accuracy and precision. Our study compares the efficacy of various ML classifies in PD diagnosis, employing complex and noisy voice data. Results reveal that the XGB classifier model trained with an over-sampled dataset in approach 2 demonstrates the best performance, scoring 1.00 for recall, 98.30% for accuracy, and 97.67% for precision. This model has the potential to offer an accuracy of 100% i.e., clinical-grade accuracy through hyper-parameter tuning when trained on a substantial dataset. It could serve as a valuable asset in telemedicine or remote healthcare applications.

V. Acknowledgement

We extend our heartfelt appreciation to the esteemed senior management of Delhi Technological University (DTU) and the Department of Biotechnology (DBT) under the Government of India for their steadfast encouragement and invaluable guidance.

References

[1] L. Naranjo, C. J. Pérez, J. Martín, and Y. Campos-Roca, "A two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications," *Comput Methods Programs Biomed*, vol. 142, 2017, doi: 10.1016/j.cmpb.2017.02.019.

[2] H. Braak and E. Braak, "Pathoanatomy of Parkinson's disease," in *Journal of Neurology, Supplement*, 2000. doi: 10.1007/pl00007758.

[3] B. Harel, M. Cannizzaro, and P. J. Snyder, "Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease: A longitudinal case study," *Brain Cogn*, vol. 56, no. 1, 2004, doi: 10.1016/j.bandc.2004.05.002.

[4] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins, and R. H. Ghomi, "Parkinson's Disease Diagnosis Using Machine Learning and Voice," in *2018 IEEE Signal Processing in Medicine and Biology Symposium, SPMB 2018 - Proceedings*, 2018. doi: 10.1109/SPMB.2018.8615607.

[5] G. M. Schulz, T. Peterson, C. M. Sapienza, M. Greer, and W. Friedman, "Voice and speech characteristics of persons with Parkinson's disease pre-and post-pallidotomy surgery: Preliminary findings," *Journal of Speech, Language, and Hearing Research*, vol. 42, no. 5, 1999, doi: 10.1044/jslhr.4205.1176.

[6] F. Amato, I. Rechichi, L. Borzi, and G. Olmo, "Sleep Quality through Vocal Analysis: a Telemedicine Application," in *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events, PerCom Workshops 2022*, 2022. doi: 10.1109/PerComWorkshops53856.2022.9767372.

[7] M. A. Little, P. E. McSharry, S. J. Roberts, D. A. E. Costello, and I. M. Moroz, "Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection," *Biomed Eng Online*, vol. 6, 2007, doi: 10.1186/1475-925X-6-23.

[8] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, "SVM and SVM ensembles in breast cancer prediction," *PLoS One*, vol. 12, no. 1, 2017, doi: 10.1371/journal.pone.0161501.

[9] M. Maalouf, "Logistic regression in data analysis: An overview," *International Journal of Data Analysis Techniques and Strategies*, vol. 3, no. 3. 2011. doi: 10.1504/IJDATS.2011.041335.

[10] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction," *Sci Rep*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-10358-x.

[11] Y. Y. Song and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch Psychiatry*, vol. 27, no. 2, 2015, doi: 10.11919/j.issn.1002-0829.215044.

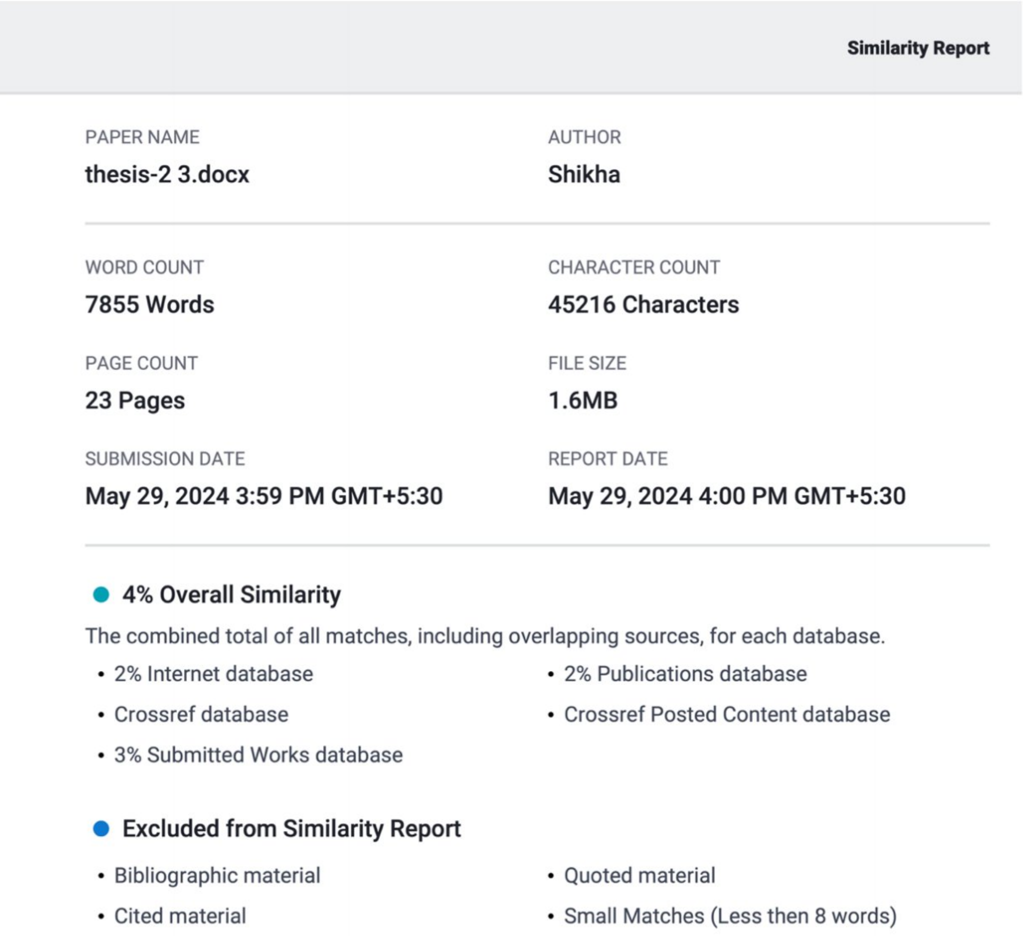
[12] D. Petkovic, R. Altman, M. Wong, and A. Vigil, "Improving the explainability of Random Forest classifier – User centered approach," in *Pacific Symposium on Biocomputing*, 2018. doi: 10.1142/9789813235533_0019.

[13] M. V. Anand, B. Kiranbala, S. R. Srividhya, K. C., M. Younus, and M. H. Rahman, "Gaussian Naïve Bayes Algorithm: A Reliable Technique Involved in the Assortment of the Segregation in Cancer," *Mobile Information Systems*, vol. 2022, 2022, doi: 10.1155/2022/2436946.

[14] A. Subasi, S. S. Panigrahi, B. S. Patil, M. A. Canbaz, and R. Klén, "Advanced pattern recognition tools for disease diagnosis," in *5G IoT and Edge Computing for Smart Healthcare*, 2022. doi: 10.1016/B978-0-323-90548-0.00011-5.

[15] B. M. S. Hasan and A. M. Abdulazeez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, 2021, doi: 10.30880/jscdm.2021.02.01.003.

PLAGIARISM REPORT



Similarity Report

PAPER NAME
thesis-2 3.docx

AUTHOR
Shikha

WORD COUNT
7855 Words

CHARACTER COUNT
45216 Characters

PAGE COUNT
23 Pages

FILE SIZE
1.6MB

SUBMISSION DATE
May 29, 2024 3:59 PM GMT+5:30

REPORT DATE
May 29, 2024 4:00 PM GMT+5:30

● 4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 2% Internet database
- Crossref database
- 3% Submitted Works database
- 2% Publications database
- Crossref Posted Content database

● Excluded from Similarity Report

- Bibliographic material
- Cited material
- Quoted material
- Small Matches (Less then 8 words)

CURRICULUM VITAE

SHIKHA KADYAN

Roll No. - 2K22/MSCBIO/46 shikha_2k22mscbio46@dtu.ac.in

EDUCATION

M.Sc. Biotechnology	2022-2024	Delhi Technological University	CGPA-8.94
B.Sc. Life Sciences	2019-2022	Miranda House, University of Delhi	CGPA- 8.97
CBSE (Class XII)	2018	DPS Panipat City	88.6%
CBSE (Class X)	2016	DPS Panipat City	95%

INTERNSHIPS

- **Drug discovery, design, and development** under the supervision of Dr. Mirza S. Baig, IIT Indore (Jan 2022-March 2022).
- **DSKC research internship** on *Mycobacterium tuberculosis*, Miranda House.

COURSES AND CERTIFICATES

- Short-term course on Cell Culture, Miranda House – May 2022
- Bioinformatics for Beginners by Miranda House (2020)

WORKSHOPS

- Primer designing and qPCR experiments workshop – 7/10/2023, Benekind Diagnostics
- Hands-on workshop on Phage Biology – Discovery and Analysis 20/01/23-25/01/23 CIIDRET, South Campus, University of Delhi

ACADEMICS ACHIEVEMNTS AND AWARDS

- Consistent maintenance of first division throughout graduation and post-graduation

SKILLS

TLC, HPLC, Paper chromatography, DNA isolation, Plaque Assay, Centrifugation, PCR

DIGITAL SKILLS

C++, SQL, DBSM, Operating System, Canva, MS Word, MS Excel and MS PowerPoint

POSITIONS OF RESPONSIBILITY

- Editorial, Design, Operations and DB Calling, RAKT, NSS DTU member
- Marketing team head of the life sciences department, MH (2020-2021)
- Member of the organizing committee in the international webinar (2021) "Drug Discovery Approaches for emerging and drug-resistant pathogens"
- Member of organizing committee in National webinar (2020) “Battling Environmental Woes: Peace Policy for future”

EXTRACURRICULAR ACTIVITIES AND ACHIEVEMENTS

- Research and Editorial team member, MH Vatavaran (2021-2022).
- Member of Women Developmental Cell, Miranda House (2019-2022).
- Member of organizing committee in National webinar “Battling Environmental Woes: Peace Policy for Future” (2020)