WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine Learning for Early Detection of Neurodegenerative Diseases

Subhiksha Seshadri Nallore, Aiswaryaa Velumani, Vikranth Reddimasu

National Institute of technology, Karnataka, Meenakshi Sundararajan Engineering College, Vasireddy Venkatadri Institute of Technology
subhikshaseshadri16@gmail.com , aiswaryaavelumani20@gmail.com , vikranthreddimasu@gmail.com
DOI: https://doi.org/10.55248/gengpi.4.1123.113123

ABSTRACT:

Neurodegenerative disorders, encompassing conditions like Parkinson's disease (PD) and Amyotrophic Lateral Sclerosis (ALS), pose formidable challenges to global healthcare. Timely identification and accurate prognosis are essential for effective intervention and improved patient outcomes. This study investigates the use of machine learning (ML) models specifically designed for early Parkinson's disease diagnosis and predictive modeling of ALS progression.

In the realm of Parkinson's disease, a variety of ML techniques are applied to analyze diverse datasets, including clinical assessments, neuroimaging, and genetic information. Supervised learning algorithms, such as support vector machines and random forests, exhibit promising outcomes in distinguishing individuals with early-stage Parkinson's disease from healthy counterparts. Furthermore, deep learning models, particularly convolutional neural networks, demonstrate high accuracy in detecting subtle patterns within neuroimaging data associated with early Parkinson's disease pathology.

Regarding Amyotrophic Lateral Sclerosis, predictive models play a pivotal role in estimating disease progression and identifying factors influencing its course. ML algorithms, including linear regression and recurrent neural networks, utilize longitudinal clinical data and biomarkers to predict the rate of ALS progression. These models contribute to tailoring treatment strategies and enable clinicians to optimize care plans for individuals grappling with this rapidly progressing neurodegenerative disorder.
Despite notable progress, challenges such as data heterogeneity, model interpretability, and ethical considerations persist. Future research directions involve integrating multimodal data for enhanced accuracy, implementing real-time monitoring for dynamic assessments, and fostering collaboration between researchers, clinicians, and technology developers.

This paper offers valuable insights into the current landscape of ML applications for early Parkinson's disease detection and predictive modeling of ALS progression. By addressing the distinctive aspects of each neurodegenerative disease, ML models have the potential to revolutionize clinical practices, opening avenues for more effective and personalized interventions in the intricate realm of neurodegenerative disorders.

Keywords: Machine Learning, Early Detection, Neurodegenerative Diseases, Parkinson's Disease, Predictive Models, Amyotrophic Lateral Sclerosis, ML Models, Biomarkers, Neuroimaging, Alzheimer's Disease, Deep Learning , Supervised Learning, Unsupervised Learning, Clinical Assessments, Genetic Information, Healthcare Technology

1. Introduction

Neurodegenerative diseases, characterized by progressive degeneration of the structure and function of the nervous system, represent a formidable challenge in healthcare. Conditions such as Alzheimer's disease, Parkinson's disease, and Amyotrophic Lateral Sclerosis not only exact a heavy toll on affected individuals but also strain healthcare systems globally. With an aging population and an increasing incidence of these disorders, understanding the underlying mechanisms and developing effective diagnostic tools are imperative.

Neurodegenerative diseases often manifest insidiously, with symptoms becoming apparent only in advanced stages. The complex interplay of genetic, environmental, and lifestyle factors contributes to their onset, making early diagnosis elusive. Consequently, a critical need exists for innovative approaches that can facilitate the identification of these conditions in their nascent stages, enabling timely interventions and improved patient outcomes.

The significance of early detection in neurodegenerative diseases cannot be overstated. Timely identification allows for the initiation of targeted interventions and personalized treatment plans, potentially altering the course of the disease. In conditions like Alzheimer's disease, where neurodegeneration begins years before clinical symptoms emerge, early detection provides a vital window for therapeutic interventions aimed at slowing or halting disease progression. Similarly, in Parkinson's disease and Amyotrophic Lateral Sclerosis, early diagnosis opens avenues for interventions that may mitigate symptoms and enhance the quality of life for affected individuals.

Moreover, early detection facilitates the inclusion of individuals in clinical trials, enabling researchers to evaluate novel therapies in the early stages of disease development. This proactive approach is essential for the development of disease-modifying treatments, as interventions at later stages often yield limited efficacy.

The integration of machine learning (ML) into healthcare represents a paradigm shift, offering unprecedented opportunities for early disease detection and precision medicine. ML techniques, fueled by the abundance of healthcare data, have demonstrated remarkable capabilities in deciphering complex patterns and relationships within diverse datasets.

In the context of neurodegenerative diseases, ML algorithms can analyze multimodal data, including neuroimaging scans, genetic profiles, and clinical assessments, to identify subtle patterns indicative of early disease pathology. The ability of ML models to learn from vast datasets enables the development of predictive models with high sensitivity and specificity, making them valuable tools for early diagnosis.

This paper explores the intersection of neurodegenerative diseases, early detection, and machine learning, aiming to provide a comprehensive understanding of the current landscape, challenges, and future directions in this dynamic field.

2. Machine Learning for Early Detection of Neurodegenerative Diseases

Machine learning plays a pivotal role in the early detection of neurodegenerative diseases by leveraging diverse datasets to identify subtle patterns associated with disease pathology and clinical manifestations. These applications hold great promise for revolutionizing diagnostic approaches and improving outcomes for individuals affected by Alzheimer's disease, Parkinson's disease, and Amyotrophic Lateral Sclerosis.

2.1 Alzheimer's Disease: Pathophysiology and Clinical Manifestations

2.1.1 Pathophysiology:

Alzheimer's disease (AD) is characterized by the accumulation of beta-amyloid plaques and tau protein tangles in the brain, leading to synaptic dysfunction and neuronal loss. The pathophysiological changes often begin years before noticeable cognitive decline, emphasizing the importance of early detection. Machine learning leverages various data sources to uncover subtle patterns associated with these pathological changes.

2.1.2 Clinical Manifestations:

Clinical manifestations of Alzheimer's include progressive memory loss, cognitive decline, and impaired executive function. Machine learning models analyze neuroimaging data, such as structural and functional MRI scans, to identify early structural alterations and abnormal neural connectivity patterns. Additionally, ML algorithms can process cognitive test results and demographic information to create predictive models for early-stage Alzheimer's detection, offering a non-invasive and potentially more accessible means of diagnosis.

Figure 1 : Clinical symptoms of Alzheimer disease

2.2 Parkinson's Disease: Motor and Non-motor Symptoms

2.2.1 Motor Symptoms:

Parkinson's disease (PD) is characterized by the degeneration of dopamine-producing neurons in the brain. Traditionally, the diagnosis relies on the observation of motor symptoms like tremors, bradykinesia, and rigidity. Machine learning, however, enhances the diagnostic process by analyzing subtle motor patterns through wearable devices and motion sensors. Algorithms can detect early motor abnormalities that might escape human observation, enabling more precise and timely diagnosis.

2.2.2 Non-motor Symptoms:

Non-motor symptoms, such as sleep disturbances, autonomic dysfunction, and cognitive impairment, are increasingly recognized as crucial indicators in Parkinson's disease. Machine learning models can integrate diverse data, including sleep patterns, autonomic function data, and cognitive assessments, to create holistic diagnostic tools. This comprehensive approach allows for the identification of Parkinson's disease at earlier stages when non-motor symptoms may be more prominent than motor symptoms.

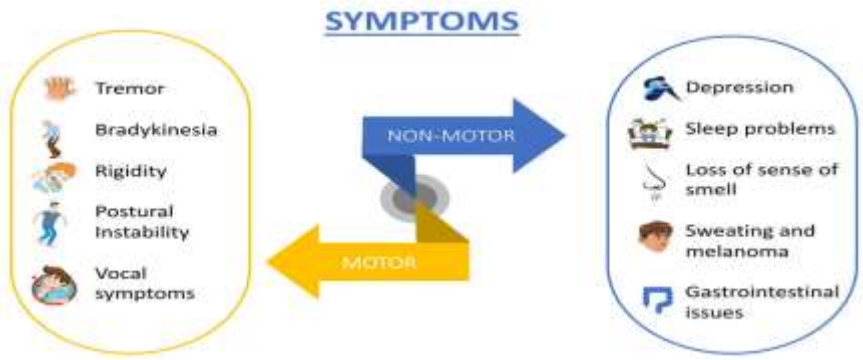

Figure 2: Non-motor symptoms in Parkinson's disease

2.3 Amyotrophic Lateral Sclerosis: Clinical Features and Challenges

2.3.1 Clinical Features:

Amyotrophic Lateral Sclerosis (ALS) is characterized by the degeneration of motor neurons, leading to progressive muscle weakness and atrophy. Early diagnosis of ALS is challenging due to its heterogeneity and the absence of specific biomarkers. Machine learning, however, holds promise in identifying subtle patterns in clinical data, such as electromyography (EMG) signals, genetic information, and clinical assessments.

2.3.2 Challenges:

The challenges in ALS diagnosis stem from the variability in disease progression and the lack of definitive diagnostic tests. Machine learning algorithms can analyze longitudinal clinical data to identify patterns associated with disease progression rates and predict outcomes. This assists in stratifying patients based on prognosis and tailoring interventions accordingly.

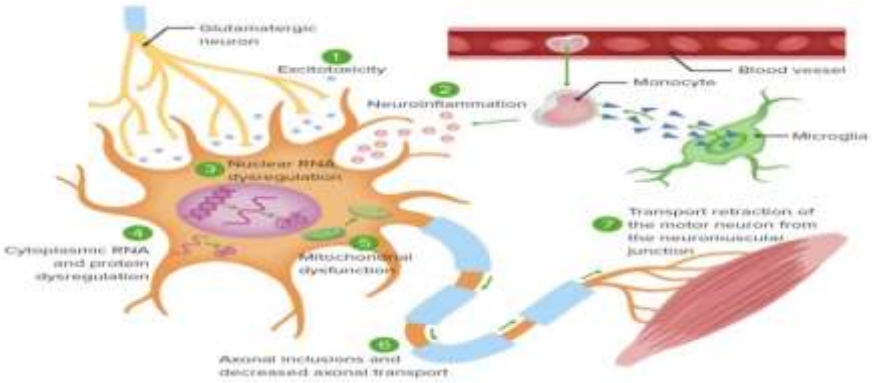

Figure 3 : Amyotrophic Lateral Sclerosis

3. Data Modalities for Early Detection

The integration of diverse data modalities, including neuroimaging, genetic information, and clinical assessments, empowers machine learning algorithms to unravel complex patterns associated with neurodegenerative diseases. This multimodal approach enhances the accuracy and sensitivity of early detection models, offering new avenues for proactive healthcare interventions and personalized treatment strategies.

3.1 Neuroimaging: MRI, PET, and FMRI

3.1.1 Magnetic Resonance Imaging (MRI):

MRI provides high-resolution images of the brain's structural anatomy. In neurodegenerative disease research, machine learning algorithms analyze MRI scans to detect subtle changes indicative of early pathology. Structural MRI can reveal alterations in brain volume and shape, while functional MRI (fMRI) assesses brain activity patterns. ML models trained on these imaging data can identify abnormal patterns associated with neurodegenerative diseases, contributing to early diagnosis.

3.1.2 Positron Emission Tomography (PET):

Positron Emission Tomography (PET) imaging entails administering radioactive tracers to enable the observation of metabolic activities within the brain. In the context of neurodegenerative diseases, PET scans prove valuable for detecting irregular protein accumulations, such as beta-amyloid in Alzheimer's disease or insufficient dopamine levels in Parkinson's disease. Utilizing machine learning algorithms, PET data is scrutinized to discern distinct patterns and measure alterations linked to the disease, thereby assisting in early detection and distinguishing between different diagnoses.

3.1.3 Functional Magnetic Resonance Imaging (FMRI):

FMRI measures brain activity by detecting changes in blood flow. It provides insights into functional connectivity and neural network dynamics. Machine learning models applied to FMRI data can identify early disruptions in functional connectivity patterns associated with neurodegenerative diseases. This modality contributes to understanding the functional changes preceding clinical symptoms, enabling early intervention strategies.

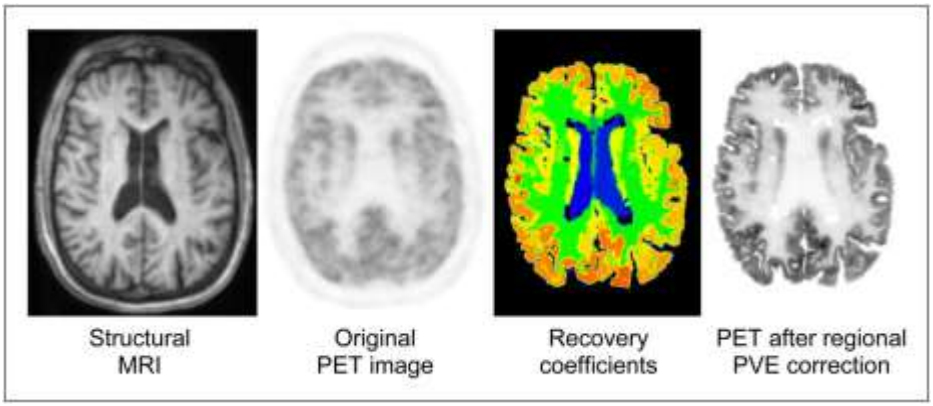

Figure 4 : PET/MRI for Neurologic Applications

3.2 Genetic Information: Biomarkers and Risk Factors

3.2.1 Biomarkers:

Genetic biomarkers play a crucial role in predicting susceptibility to neurodegenerative diseases. Machine learning models analyze genetic data, including single nucleotide polymorphisms (SNPs) and variations in specific genes associated with disease risk. By identifying patterns and interactions within genetic datasets, ML facilitates the identification of individuals at higher risk, enabling proactive monitoring and intervention.

3.2.2 Risk Factors:

Machine learning applications extend beyond individual genetic markers to analyze complex interactions among multiple genes and environmental factors. These models can identify combinations of genetic and environmental variables that contribute to increased neurodegenerative disease risk. The integration of diverse data sources enhances the accuracy of risk prediction models, enabling targeted early interventions.

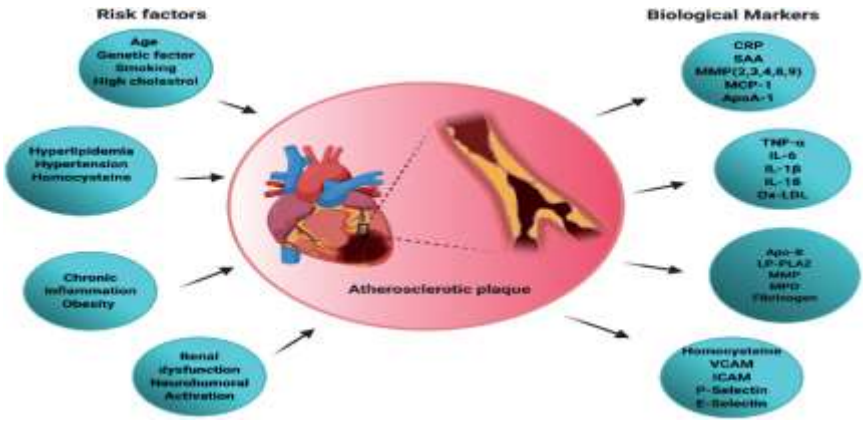

Figure 5 : Schematic representation of different risk factors and biomarkers

3.3 Clinical Assessments: Cognitive Tests, Motor Function Assessments

3.3.1 Cognitive Tests:

Neuropsychological assessments, including cognitive tests, are vital for diagnosing and monitoring neurodegenerative diseases. Machine learning algorithms process cognitive test results to identify subtle changes indicative of early cognitive decline. These models can detect patterns associated with specific neurodegenerative diseases, aiding in early differentiation and personalized treatment planning.

3.3.2 Motor Function Assessments:

Motor function assessments, such as gait analysis and fine motor skill assessments, provide valuable information for the early detection of diseases like Parkinson's. Wearable devices equipped with accelerometers and gyroscopes collect quantitative data on motor patterns. Machine learning algorithms analyze this data to detect subtle abnormalities indicative of early-stage motor dysfunction, contributing to early diagnosis and intervention.

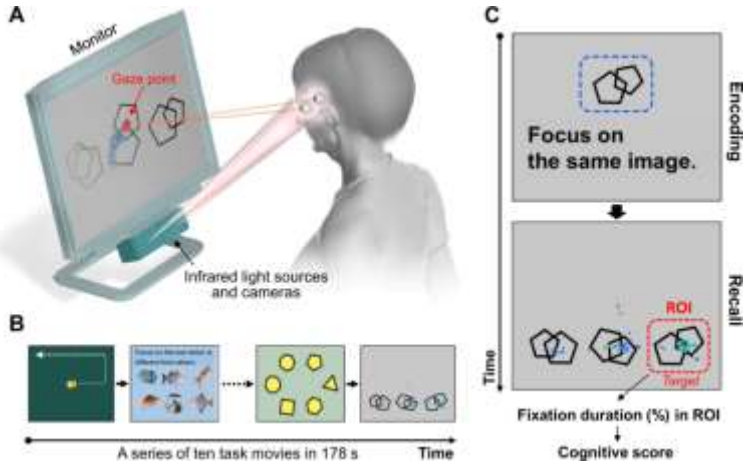

Figure 6 : Novel Method for Rapid Assessment of Cognitive

4. Machine Learning Techniques for Early Detection of Neurodegenerative Diseases

A diverse array of machine learning techniques, including supervised learning, unsupervised learning, deep learning, and ensemble methods, contributes to the development of robust models for the early detection of neurodegenerative diseases. The choice of technique often depends on the nature of the data and the specific challenges associated with each disease.

4.1 Supervised Learning: Classification Algorithms

4.1.1 Overview:

Supervised learning involves training a model on a labeled dataset, where the algorithm learns to map input features to predefined output labels. In the context of neurodegenerative diseases, classification algorithms, such as Support Vector Machines (SVM), Random Forest, and Logistic Regression, are

employed. These algorithms learn from labeled data, where instances are categorized as either indicative or non-indicative of the disease. The trained model can then classify new data, aiding in the early identification of specific disease markers.

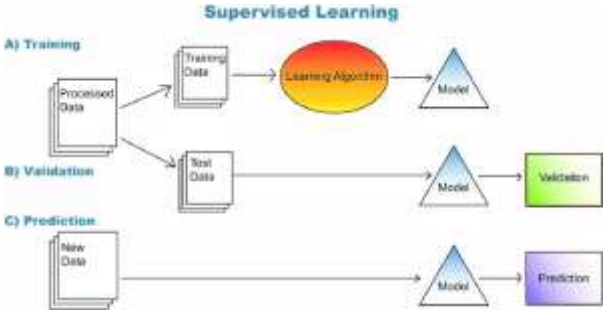

Figure 7 : Graphic representation of supervised machine learning

4.1.2 Application:

For example, in Alzheimer's disease detection, supervised learning models can be trained on neuroimaging data labeled with disease status. The algorithm learns to recognize patterns associated with early-stage Alzheimer's, facilitating the classification of new imaging data and aiding in early diagnosis.

4.2 Clustering and Anomaly Unsupervised Learning: Detection

4.2.1 Overview:

Unsupervised learning does not rely on labeled data but instead identifies patterns or groups within the data. Clustering algorithms, such as k-means and hierarchical clustering, group similar instances together. Anomaly detection algorithms identify instances that deviate significantly from the norm. In neurodegenerative disease research, unsupervised learning helps discover subtle patterns within datasets that may indicate early disease stages.

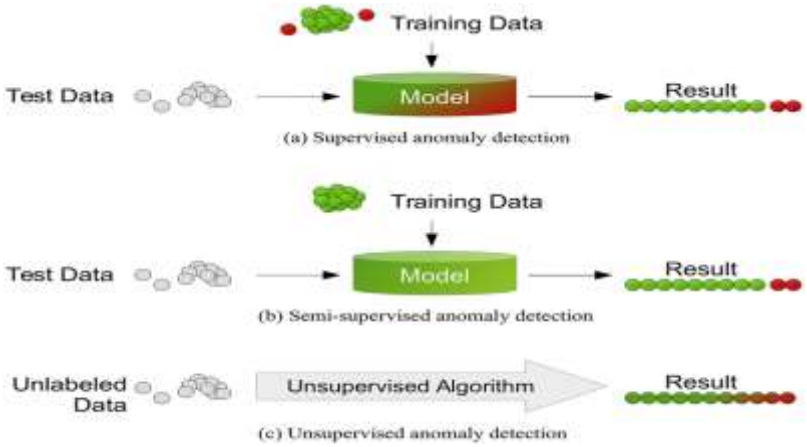

Figure 8: Unsupervised anomaly detection algorithms

4.2.2 Application:

For instance, unsupervised learning can be applied to neuroimaging data to identify subgroups of patients with distinct imaging patterns. This may reveal early-stage characteristics that were not previously apparent, aiding in the understanding of disease heterogeneity and facilitating targeted interventions.

4.3 Deep Learning: Neural Networks for Feature Extraction

4.3.1 Overview:

Deep learning, particularly neural networks, excels at automatically learning hierarchical representations from complex data. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are commonly used architectures for feature extraction. In neurodegenerative disease detection, deep learning models can automatically extract intricate patterns from neuroimaging or genetic data, contributing to highly accurate and nuanced early detection models.

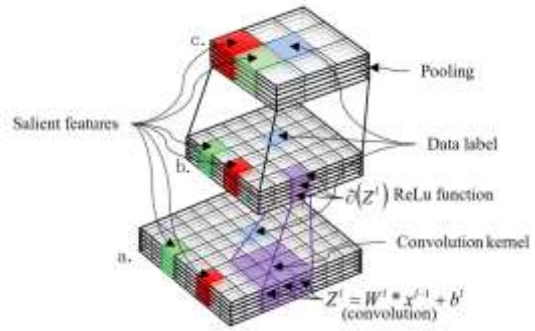

Figure 9 : The processing of feature extraction in deep convolution neural network (CNN).

4.3.2 Application:

In the context of Alzheimer's disease, deep learning models can process 3D brain images to automatically learn features indicative of pathological changes. The model may discover subtle alterations in brain structures that are challenging for traditional algorithms to identify, enhancing the sensitivity of early detection.

4.4 Ensemble Methods: Boosting and Bagging

4.4.1 Overview:

Ensemble methods combine multiple machine learning models to improve overall performance. Boosting (e.g., AdaBoost) combines weak learners sequentially, assigning more weight to misclassified instances. Bagging (e.g., Random Forest) builds multiple models in parallel on different subsets of the data and combines their predictions. In neurodegenerative disease research, ensemble methods enhance the robustness and generalizability of models.

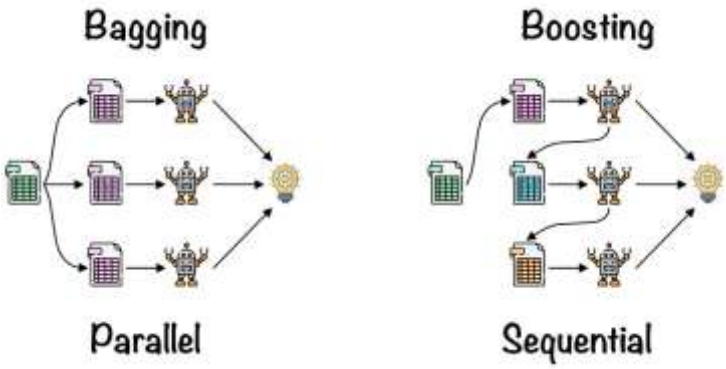

Figure 10 : Ensemble Learning: Bagging & Boosting

4.4.2 Application:

For instance, in Parkinson's disease detection, ensemble methods can combine information from diverse data modalities—neuroimaging, genetic data, and clinical assessments. This integrated approach reduces the risk of overfitting to a specific dataset and enhances the model's ability to generalize to new, unseen data, improving the reliability of early detection models.

5. Proposed Methodology for Parkinson's disease

The proposed methodology involves gathering audio data related to Parkinson's patients' voice modulations from both the PPMI [21] and UCI datasets. This dataset encompasses information regarding jitter, shimmer, and MDVP of vowel phonations. The collected data undergoes preprocessing, analysis, and visualization to gain a comprehensive understanding of its attributes. Subsequently, four models—Logistic Regression, Support Vector Machine (SVM), Random Forest Regressor, and K Nearest Neighbors—are trained on 75% of the dataset. These models are specifically trained to classify the given audio data as either indicative of Parkinson's disease (PD) or as healthy, relying on variations in frequency.

The models are then tested on the remaining 25% of the data, and their performance is evaluated using metrics such as sensitivity, precision, accuracy, confusion matrix [22], and ROC-AUC score. Figure 11 visually represents the generic process implemented in this study. It outlines the stages of data ingestion from the PPMI database, the separation of data into testing and training sets, the training of the four models, and the validation of results using the test data.

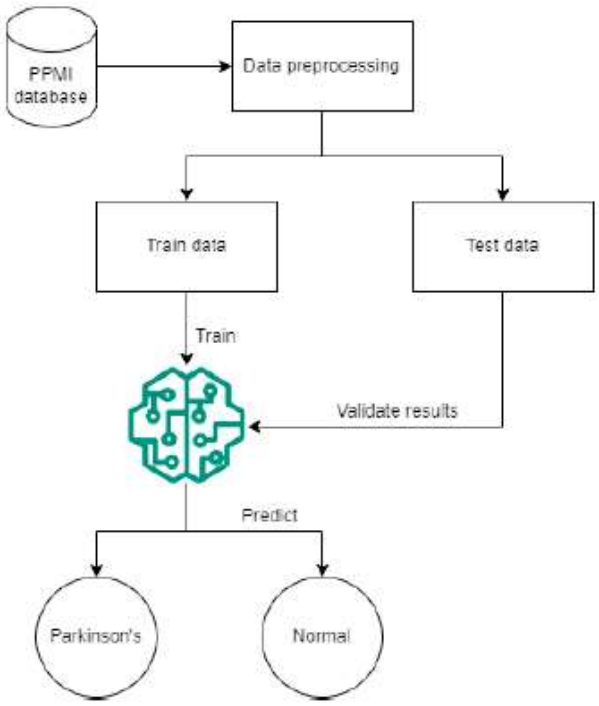

Figure 11 : Proposed Architecture

The objective of this research paper is to pinpoint the key attributes crucial for Parkinson's Disease (PD) classification and to investigate the influence of data imbalance in medical classification. With these goals in mind, three distinct approaches have been applied. First, training is conducted on the entire dataset, serving as a baseline test for PD classification. Second, training is performed on Principal Component Analysis (PCA)-identified attributes. Third, training is carried out on a set of 109 records obtained after balancing the dataset. The algorithms employed in each of these approaches are delineated below:

Algorithm for approach 1: Models are trained on 22 attributes of statistics

- Collect MDVP audio data from PPPMI and UCI databases.
- Perform data analysis to identify skew, imbalance, and variable distribution in the dataset.
- Scale the data to a common range using Standard Scaler.
- Split the dataset into testing and training sets, with 75% of the data allocated for training.
- Train SVM, logistic regression, random forest, and KNN models using the 22 attribute of the data.

Algorithm for approach 2: Principal module Analysis (PCA) is applied to identify 5 key attributes

- Collect MDVP audio data from PPPMI and UCI databases.
- Conduct data analysis to identify skew, imbalance, and variable distribution.
- Scale the data to a common range using Standard Scaler.
- Identify variance in each column and apply Principal Component Analysis (PCA) to pinpoint 5 key attributes out of the 22.
- Split the dataset into testing and training sets, with 75% of the data used for training.
- Retrain SVM, logistic regression, random forest, and KNN models.
- Compare classification results using the confusion matrix, ROC-AUC curve, and accuracy metrics.

Algorithm for approach 3: Imbalance exclusion in dataset

- Collect MDVP audio data from PPPMI and UCI databases.

- Perform data analysis to identify skew, imbalance, and variable distribution.
- Address the dataset imbalance, which initially consists of 109 records of Parkinson's patients (PWP) and 40 records of normal individuals, as depicted in Figure 2(a). Resolve the imbalance by upsampling the minority class to achieve 109 records each, as shown in Figure 2(b).
- Scale the data to a common range using Standard Scaler.
- Split the dataset into testing and training sets, with 75% of the data designated for training.
- Retrain SVM, logistic regression, random forest, and KNN models.
- Compare classification results using the confusion matrix, ROC-AUC curve, and accuracy metrics.

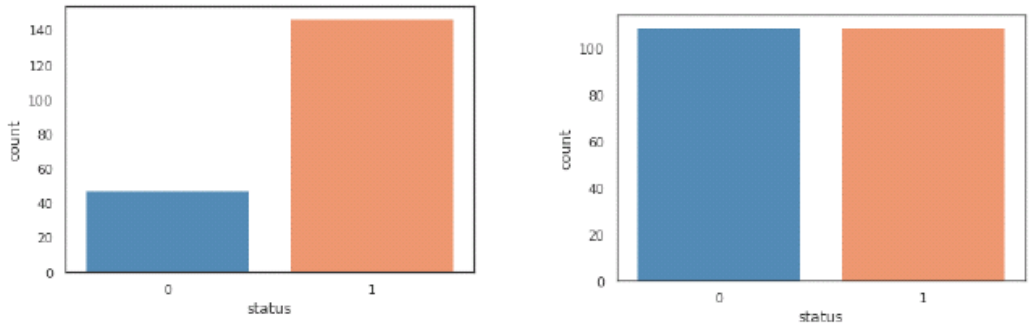

Figure 12. (a) Imbalanced data with 40 normal records; (b) Balanced data after up sampling

5.1. Dataset

Biomedical voice measurements [24] were obtained from a cohort of 31 individuals, of which 23 were diagnosed with Parkinson's disease (PD). The age range of the patients is between 46 and 85 years, while normal readings were obtained from individuals aged 23 years. On average, six phonations were recorded 195 times for each person, with the duration ranging from 1 to 36 seconds. The attributes of the 195 records are detailed in Table 1 below:

5.2. Data preprocessing

Data wrangling [26] procedures have been applied to cleanse the data and address any missing attributes in the dataset. Figure 13 illustrates the ratio of noise to harmonic tone (NHR) and the ratio of harmonic tone to noise (HNR) for individuals with Parkinson's disease (PWP). As the disease advances through its stages, there is a noticeable rise in speech-related noise, leading to an elevated NHR, as depicted in Figure 3 (b). The data's skewness and the low value of NHR (0.3) suggest a diminished voice quality.

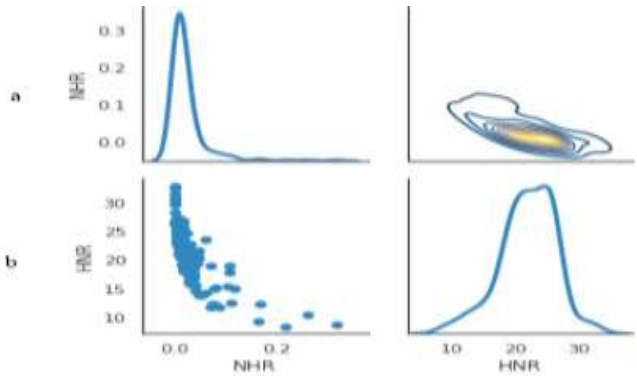

Figure 13 : (a) NHR plot; (b) HNR plot

Figure 14 illustrates a box plot encompassing all 22 attributes present in the dataset. This plot visually represents the distribution and skewness of data across a median quartile. The figure uses blue for normal records and orange for Parkinson's patients (PWP) records. Notably, the NHR data points for PWP exhibit the highest number of outliers, attributed to heightened speech-related noise. Likewise, the HNR records display a significant number of data outliers below the median for PWP records.

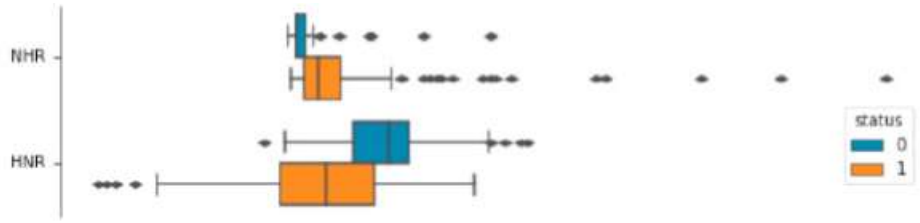

Figure 14 : Box plot of NHR and HNR

Figure 15 presents a pair plot specifically focusing on shimmer data. This plot aims to emphasize the divergence in voice shimmer between individuals with Parkinson's disease (PWP) and healthy patients. Notably, it reveals a linear correlation between Shimmer: APQ3 and Shimmer: DDA, while Shimmer: APQ5 and Shimmer: APQ3 exhibit a left-skewed distribution.

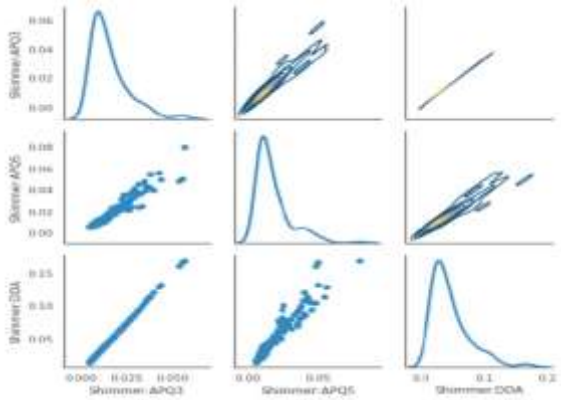

Figure 15 : Pair plot of shimmer data

6. Model training

This research paper investigates the performance of Logistic Regression, Random Forest classifier, Support Vector classifier, and K nearest neighbors' models in three distinct approaches:

- Utilizing the complete dataset comprising 195 records and 22 attributes.
- 2. Employing a dataset with 195 records and reduced to 5 attributes after Principal Component Analysis (PCA).
- Working with a balanced dataset consisting of 109 records and 22 attributes.

Table 1 : Dataset attributes

Attribute	**Purpose**
Name	Patient name and recording number are stored in ASCII CSV format.
MDVP: Fo (Hz)	Represents the fundamental frequency of the pitch period.
MDVP: Fhi (Hz)	Denotes the upper limit of the fundamental frequency or the maximum threshold of voice modulation.
MDVP: Flo (Hz)	Indicates the lower limit or minimal vocal fundamental frequency.
MDVP: Jitter, Abs, RAP, PPQ, DDP	These encompass various measures from Kay Pentax's multi-dimensional voice program (MDVP), a traditional method assessing the frequency of vibrations in vocal folds during the pitch period to vibrations at the start of the next cycle called pitch mark [25].
Jitter and Shimmer	Quantify the absolute difference between frequencies of each cycle, normalized to the average.
NHR and HNR	Represent signal-to-noise and tonal ratio measures, offering insight into the robustness of the environment to noise.
Status	A binary indicator where 0 signifies a healthy person, while 1 indicates a person with a specific condition (PWP).
D2	Utilized for identifying dysphonia in speech through correlation dimension using fractal objects; it is a nonlinear, dynamic attribute.

RPDE	Measures Recurrence Period Density Entropy, providing insight into the extent to which the signal is periodic.
DFA	Denotes Detrended Fluctuation Analysis, measuring the extent of stochastic self-similarity of noise in speech signals.
PPE	Pitch Period entropy, employed to assess abnormal variations in speech on a logarithmic scale.

6.1. Logistic regression for classification

Logistic regression [27] stands as a widely employed supervised machine learning algorithm that forecasts categorically dependent variables based on a set of independent variables. It employs a curve fitting method to anticipate a probabilistic value within the 0 to 1 range, serving as the outcome for a categorical or discrete input. In contrast to Linear regression [28], which fits a line to linearly predict one or more dependent variables, logistic regression anticipates an S-shaped logistic curve for values within the 0 to 1 range. This proves advantageous for analyzing audio data, as attributes influencing the classification of Parkinson's Disease (PD) are not linearly correlated; instead, they follow an exponential pattern. The activation function of the logistic classification is depicted in Figure 16.

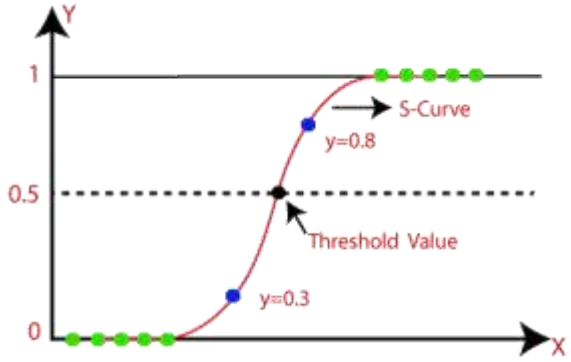

Figure 16 : Logistic regression classifier

6.2. Random Forest classifier

The Random Forest, a supervised machine learning algorithm suitable for both classification and regression tasks, is employed in this study. The implementation involves a random forest classifier [29], wherein multiple decision trees are trained on subsets of the dataset. The algorithm considers the average of these trees to enhance predictive accuracy. It operates as a democratic model, treating no single decision tree as superior. Instead, the collective majority vote from all models is considered to generate an average prediction for the output. As the number of trees increases, the likelihood of overfitting decreases. The architecture of the random forest classifier utilized in this research paper is depicted in Figure 17 below.

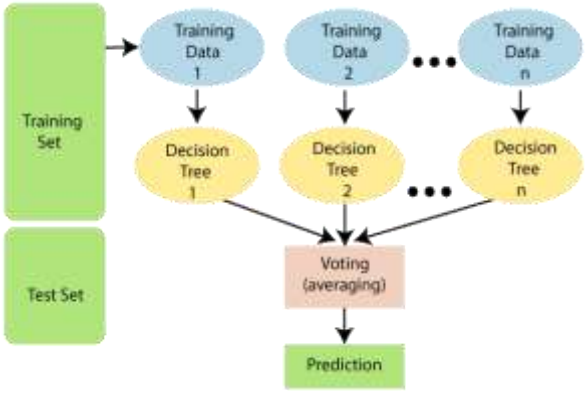

Figure 17 : Random Forest classifier Architecture

6.3. Support Vector Machine (SVM)

The Support Vector Machine (SVM) [30] stands as a supervised machine learning algorithm that establishes a hyperplane for the separation of N features by mapping them to a multidimensional space. The structure of the SVM model is depicted in Figure 18.

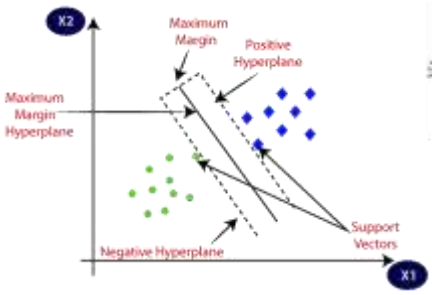

Figure 18 : Support Vector Machine (SVM)

Given that Parkinson's disease (PD) voice data lacks linear separability, an SVM kernel is employed to elevate the data into a higher-dimensional space. SVM demonstrates effectiveness in handling PD data, thanks to its memory efficiency and the formation of support vectors derived from a subset of training data points.

6.4. K nearest neighbors (KNN)

The K-nearest neighbors (KNN) algorithm [31] is a non-parametric, supervised machine learning approach that clusters data according to inherent similarities. It excels when applied to balanced audio data consisting of 109 records, primarily due to its suitability for small dataset sizes. Efficiently creating two clusters for Parkinson's disease and healthy data, KNN is considered a lazy learning algorithm, signifying that no assumptions about the data are made, allowing for the learning of novel patterns directly from the training data.

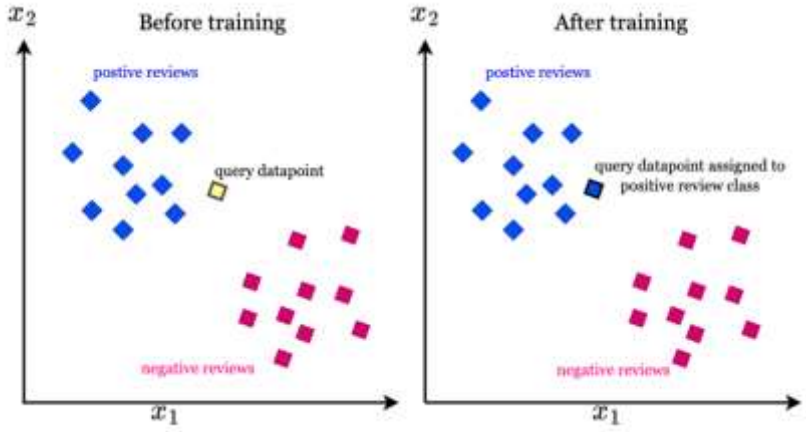

Figure 19 : K nearest neighbors (KNN)

7. Model evaluation

In the process of determining the optimal model, we evaluate the outcomes of three different approaches and nine trained models. The selected metrics for comparison include the ROC-AUC curve, confusion matrix, accuracy, precision, recall, and F1 score. The formulas for these metrics are depicted in equations 1-3, with TP representing True Positives, FP for False Positives, TN for True Negatives, and FN for False Negatives.

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Accuracy = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (3)$$

The Receiver Operating Characteristics (ROC) curve [32] is a probability curve, and the Area Under the Curve (AUC) quantifies the area beneath this curve. AUC serves as a metric for the model's ability to distinguish between classes or its separability. This metric gauges the trade-off between clinical sensitivity and specificity for a given set of tests.

The outcomes of models following the application of approach 1, where models are trained on the 22 attributes of the MDVP dataset, are presented in Table 2 below.

Table 2. Outcomes of Strategy 1: 22 attribute instruction

Metric	Logistic Regression	Random Forest	SVM	KNN
Accuracy	83.67%	91.83 %	85.71 %	85.71 %
Precision	1.0	0.95	1.0	0.95
Recall	0.83	0.86	0.84	0.86
ROC AUC curve	0.636	0.701	0.682	0.701

The Random Forest classifier is well-suited for the entire dataset due to its ensemble nature. It assesses the average output from 100 decision trees before making a final prediction. Each attribute holds equal weight during the classification process. The confusion matrix for this model is depicted in Figure 20 below, where the model accurately classifies 7 instances as true negatives (no PD), 4 instances as false negatives, 38 instances as true positives (PWP), and records 0 false positives.

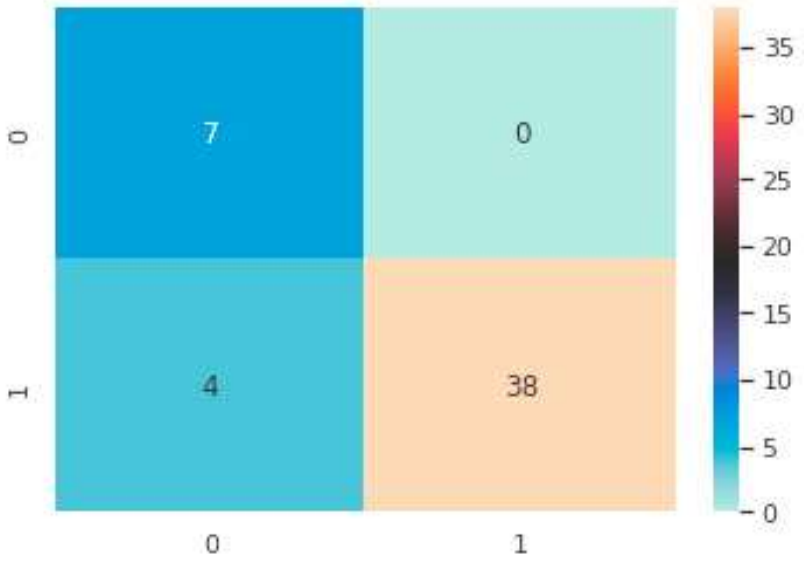

Figure 20 : Approach 1's confusion matrix for the Random Forest Model

Table 3 illustrates the outcomes of approach 2 following the implementation of Principal Component Analysis (PCA). This process yields five major attributes: MDVP, Shimmer, Jitter, PPE, and RPDE. Subsequent to training and evaluating models based on these five attributes, the results are as follows:

Table 2 : Outcomes of Strategy 1: 22 attribute instruction

Metric	Logistic Regression	Random Forest	SVM	KNN
Accuracy	83.67%	83.67%	91.75 %	83.67 %
Precision	1.0	1.0	1.0	0.92
Recall	0.83	0.90	0.86	0.90
ROC AUC curve	0.636	0.818	0.727	0.779

The Support Vector classifier with L1-support and a linear kernel proves advantageous for PCA-transformed data, as it efficiently identifies the optimal hyperplane with reduced processing time and enhanced accuracy. The confusion matrix for this model is presented in Figure 21 below, where the SVM accurately classifies 5 instances as true negatives (no PD), 6 instances as false negatives, 38 instances as true positives (PWP), and registers 0 false positives.

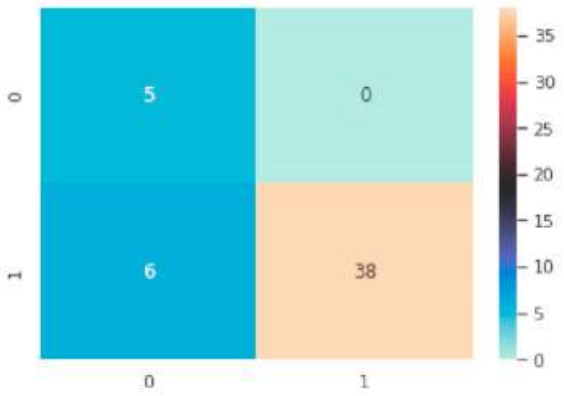

Figure 21 : Approach 2's confusion matrix for the Support Vector Model

Table 4 presented below illustrates the outcomes for approach 3, involving the utilization of a balanced dataset. Models were trained on an equal number of records from both normal and Parkinson's patients' data. This balancing strategy ensures that equal importance is attributed to both PWP and non-Parkinson's patients. The results are detailed as follows:

Table 3 : Results of Approach 2: 5 attributes after PCA

Metric	Logistic Regression	Random Forest	SVM	KNN
Accuracy	85.71%	85.71 %	81.63 %	91.83 %
Precision	0.89	0.89	0.82	0.95
Recall	0.92	0.92	0.94	0.95
ROC AUC curve	0.811	0.811	0.817	0.883

Among the models applied to the balanced dataset, the K-nearest neighbors (KNN) model exhibits superior performance, achieving the highest precision and recall, both at 0.95. The balanced distribution of data facilitates the quicker identification of similarities between PWP and non-Parkinson's patients. The classification results are depicted in the confusion matrix shown in Figure 22. The KNN model classifies the data into 9 true negatives (no PD), 2 false negatives, 36 true positives (PWP), and 2 false positives.

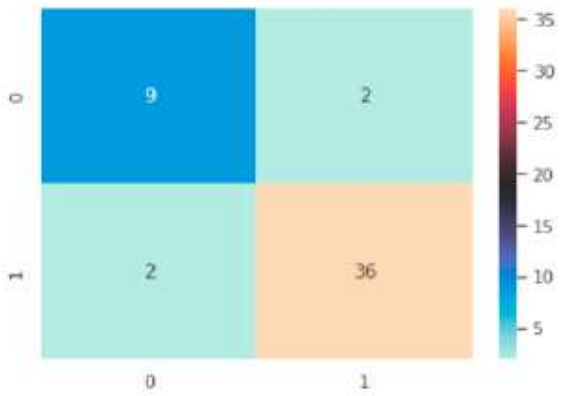

Figure 22 : Approach 3's confusion matrix for the KNN model

10. Conclusion:

The classification of Parkinson's disease using vowel phonation data achieves an accuracy of 91.835% and a sensitivity of 0.95 for the Random Forest classifier. The Random Forest model yields favorable results, benefiting from the equitable consideration of all 22 attributes in the MDVP dataset. Additionally, this paper presents the outcomes of the SVM model, demonstrating an accuracy of 91.836% and a sensitivity of 0.94 after applying PCA to the dataset. Both SVM and Random Forest models exhibit robust performance against outliers and notably predict no false positives in the results.

The K-nearest neighbor (KNN) model also demonstrates efficacy for a balanced dataset, particularly due to its capacity to classify into two categories without presumptions about the data. Consequently, we recommend the utilization of the Random Forest model for the classification of Parkinson's disease. This approach proves non-invasive, straightforward, and accurate, offering long-term relief to individuals with Parkinson's disease on a global scale.

Looking ahead, we propose incorporating audio and REM sleep data to enhance classification results, recognizing that audio data alone may not suffice as a comprehensive biomarker for Parkinson's disease. These findings aim to encourage the use of mobile-recorded audio for PD classification through telemedicine.

References:

[1]. [1] Prabhavathi, K., Patil, S. (2022). "Tremors and Bradykinesia. In: Arjunan, S.P., Kumar, D.K. (eds) Techniques for Assessment of Parkinsonism for Diagnosis and Rehabilitation". Series in BioEngineering. Springer. 135–149 https://doi.org/10.1007/978-981-16-3056-9_9

[2]. [2] Braak, H., Braak, E. (2000) "Pathoanatomy of Parkinson's disease" J Neurol 247, II3–II10. https://doi.org/10.1007/PL00007758

[3]. [3] F. Amato, I. Rechichi, L. Borzì and G. Olmo, (2022), "Sleep Quality through Vocal Analysis: A Telemedicine Application," 2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops), 706-711, doi: 10.1109/PerComWorkshops53856.2022.9767372.

[4]. [4] Neighbors C, Song SA. "Dysphonia" (2022) StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing.

[5]. [5] Serge Pinto, Canan Ozsancak, Elina Tripoliti, Stéphane Thobois, Patricia Limousin-Dowsey, Pascal Auzou, "Treatments for dysarthria in Parkinson's disease", (2004) The Lancet Neurology, 3(9): 547-556, ISSN 1474-4422, https://doi.org/10.1016/S1474-4422(04)00854-3.

[6]. [6] Nicoló G. Pozzi, Ioannis U. Isaias (2022), "Chapter 19 - Adaptive deep brain stimulation: Retuning Parkinson's disease", Elsevier 184: 273-284. https://doi.org/10.1016/B978-0-12-819410-2.00015-1

[7]. [7] Alatas Bilal, Moradi Shadi, Tapak Leili, Afshar Saeid (2022), "Identification of Novel Noninvasive Diagnostics Biomarkers in the Parkinson's Diseases and Improving the Disease Classification Using Support Vector Machine", BioMed Research International, Hindawi

[8]. [8] P. Raundale, C. Thosar and S. Rane (2021), "Prediction of Parkinson's disease and severity of the disease using Machine Learning and Deep Learning algorithm," 2021 2nd International Conference for Emerging Technology (INCET), pp. 1-5, doi: 10.1109/INCET51464.2021.9456292.

[9]. [9] F. Cordella, A. Paffi and A. Pallotti (2021) "Classification-based screening of Parkinson's disease patients through voice signal," 2021 IEEE International Symposium on Medical Measurements and Applications (MeMeA), pp. 1-6, doi: 10.1109/MeMeA52024.2021.9478683.

[10]. [10] Ali, L., Chakraborty, C., He, Z. et al. (2022) "A novel sample and feature dependent ensemble approach for Parkinson's disease detection". Neural Comput & Applic. https://doi.org/10.1007/s00521-022-07046-2

[11]. [11] F. Huang, H. Xu, T. Shen and L. Jin (2021), "Recognition of Parkinson's Disease Based on Residual Neural Network and Voice Diagnosis," 2021 IEEE 5th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 381-386, doi: 10.1109/ITNEC52019.2021.9586915.

[12]. [12] D. Trivedi H. Jaeger and M. Stadtschnitzer. (2019) "Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls." https://doi.org/10.5281/zenodo.2867216

[13]. [13] M. Wodzinski, A. Skalski, D. Hemmerling, J. R. Orozco-Arroyave and E. Nöth, (2019) "Deep Learning Approach to Parkinson's Disease Detection Using Voice Recordings and Convolutional Neural Network Dedicated to Image Classification," 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 717-720, doi: 10.1109/EMBC.2019.8856972.

[14]. [14] T. J. Wroge, Y. Özkanca, C. Demiroglu, D. Si, D. C. Atkins and R. H. Ghomi, (2018), "Parkinson's Disease Diagnosis Using Machine Learning and Voice", 2018 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1-7, doi: 10.1109/SPMB.2018.8615607.

[15]. [15] W. Wang, J. Lee, F. Harrou and Y. Sun, "Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning," in IEEE Access, vol. 8, pp. 147635-147646, 2020, doi: 10.1109/ACCESS.2020.3016062.

[16]. [16] R. Alkhatib, M. O. Diab, C. Corbier and M. E. Badaoui, "Machine Learning Algorithm for Gait Analysis and Classification on Early Detection of Parkinson," in IEEE Sensors Letters, vol. 4, no. 6, pp. 1-4, June 2020, Art no. 6000604, doi: 10.1109/LSENS.2020.2994938.

[17]. [17] C. Ricciardi et al., "Machine learning can detect the presence of Mild cognitive impairment in patients affected by Parkinson's Disease," 2020 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2020, pp. 1-6, doi: 10.1109/MeMeA49120.2020.9137301.

[18]. [18] X. Yang, Q. Ye, G. Cai, Y. Wang and G. Cai, (2022), "PD-ResNet for Classification of Parkinson's Disease from Gait," in IEEE Journal of Translational Engineering in Health and Medicine, vol. 10, pp. 1-11, 2022, Art no. 2200111, doi: 10.1109/JTEHM.2022.3180933.

[19]. [19] A. U. Haq et al., "Feature Selection Based on L1-Norm Support Vector Machine and Effective Recognition System for Parkinson's Disease Using Voice Recordings," in IEEE Access, vol. 7, pp. 37718-37734, 2019, doi: 10.1109/ACCESS.2019.2906350.

[20]. [20] Mei Jie, Desrosiers Christian, Frasnelli Johannes, (2021), “Machine Learning for the Diagnosis of Parkinson's Disease: A Review of Literature”, in Frontiers in Aging Neuroscience, vol. 13, doi: 10.3389/fnagi.2021.633752.

[21]. [21] https://www.ppmi-info.org/access-data-specimens/download-data

[22]. [22] Amalia Luque, Alejandro Carrasco, Alejandro Martín, Ana de las Heras, (2019), “the impact of class imbalance in classification performance metrics based on the binary confusion matrix”, Pattern Recognition, Volume 91, Pages 216-231, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2019.02.023.

[23]. [23] J. R. Barr, M. Sobel and T. Thatcher (2022), "Upsampling, a comparative study with new ideas," 2022 IEEE 16th International Conference on Semantic Computing (ICSC), pp. 318-321, doi: 10.1109/ICSC52841.2022.00059.

[24]. [24] Little, M.A., McSharry, P.E., Roberts, S.J. et al. (2007) “Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection” BioMed Eng OnLine 6 (23). https://doi.org/10.1186/1475-925X-6-23

[25]. [25] Little, Max A et al. (2009) “Suitability of dysphonia measurements for telemonitoring of Parkinson's disease.” IEEE transactions on bio-medical engineering vol. 56 (4): 1015. doi:10.1109/TBME.2008.2005954