[1] **Kaushik Subramanian**

[2] **Adesh JP**

[3] **Amutha AL**

Identification of Parkinson's Disease Using Stacking Classifier

Abstract: - Parkinson's disease, a neurodegenerative condition, impacts millions of individuals across the globe. Timely and precise identification is imperative for efficient administration and therapy. The current study centers around utilizing machine learning algorithms to forecast Parkinson's disease by analyzing biomedical and clinical information. A comprehensive dataset comprising demographic information, medical history, and clinical assessments is collected and preprocessed to handle missing values and ensure data quality. To determine the most significant factors that can predict the disease, methods for selecting and extracting features are employed. The effectiveness of stacking classifier algorithm is assessed in terms of their ability to make accurate predictions. The dataset is split into two parts, one for training and the other for testing. Cross-validation is used to adjust the model's hyperparameters and stay away from overfitting. The model has been thoroughly assessed using conventional classification metrics and support vector machines (SVM). The findings of this research indicate that machine learning has significant potential in accurately predicting Parkinson's disease. Consequently, healthcare providers have the ability to enhance the well-being of individuals with Parkinson's disease and detect those who may be at risk at an earlier stage of the condition. This study adds to the continuous attempts to make use of advanced data analysis methods for the prompt detection and treatment of neurodegenerative disorders such as Parkinson's.

Keywords: Parkinson's Disease, SVM (Support Vector Machine), KNN.

I. INTRODUCTION

Parkinson's Disease, a progressive and intricate neurodegenerative condition, impacts a significant number of individuals globally. The burden of Parkinson's disease is significant for those affected and also challenges both healthcare systems and society on the whole. PD is mainly a disorder that affects movement and is identified by various motor symptoms which includes shaking when not in motion, slow movement, stiffness, and difficulty maintaining balance. The motor symptoms occur due to the progressive decline of dopamine producing neurons in the substantia nigra area of the brain as well as the subsequent reduction of dopamine levels, which is an essential neurotransmitter for controlling movement. PD is more than just a motor function disorder; it encompasses numerous other impairments. It includes a wide range of non-motor symptoms that can greatly affect a patient's quality of daily life, such as problems with thinking, emotional disorders, difficulty sleeping, along with issues with the autonomic nervous system. It is still unclear what causes Parkinson's disease, but a mix of environmental and genetic factors may play a role. The suspected causes of Parkinson's disease include genetic mutations, contact with harmful substances, and certain aspects of one's lifestyle. However, it is still not fully understood exactly how these factors exactly interact and contribute to the development of the disease. The diagnosis of Parkinson's disease primarily depends on clinical assessment and the recognition of typical motor symptoms, usually when a significant number of dopamine-producing neurons have already been damaged. It is extremely important to diagnose and intervene early, as the current treatments mainly concentrate on managing symptoms rather than modifying the disease. There is an urgent requirement for improved methods of diagnosis and treatment to prevent or delay the advancement of diseases.

In this article, we expand on the research conducted by https://rdcu.be/dmaqN. In this particular study, we adopt a comparable methodology, but we aim to make use of alternative machine learning classifiers to enhance the model's performance. Additionally, these classifiers are crucial in facilitating early prediction of PD, enabling the timely implementation of neuroprotective treatments.

[1] *Corresponding author: Department of Computational Intelligence, Faculty of Engineering and Technology, SRMIST, Chennai, India. Email: ks2916@srmist.edu.in

[2] Department of Computational Intelligence, Faculty of Engineering and Technology, SRMIST, Chennai, India. Email: aj9703@srmist.edu.in

[3] Department of Computational Intelligence, Faculty of Engineering and Technology, SRMIST, Chennai, India. Email: amuthaa1@srmist.edu.in

II.LITERATURE SURVEY

There are a lot of related works in predicting Parkinson's disease. In [1] The article explains the creation and verification of a machine learning algorithm that combines genetic information, transcriptomics data, and clinical data to forecast the likelihood of developing Parkinson's disease. The objective in [2] was to explore the potential of functional MRI (fMRI) to predict optimal deep brain stimulation (DBS) parameters for individual patients with Parkinson's disease, using machine learning. The research paper acknowledges several limitations to their findings. Firstly, the order in which fMRI data were acquired was not entirely randomized. Secondly, the brain network responses to changes in frequency, pulse width, and stimulation polarity remain to be examined. The authors of [3] conducted training on various models, including Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), K-Nearest Neighbor (KNN), and Multi Layer Perceptron (MLP), in order to distinguish between patients with Parkinson's disease (PD) and those who are healthy. The dataset utilized in the study included 195 voice recordings from 31 patients who were undergoing examinations. The models underwent training to improve their performance through methods like SMOTE, Feature Selection, and hyperparameter tuning via GridSearchCV.

The results indicated a 70: 70 ratio between MLP and SVM. The project obtained its most favorable outcomes by employing GridSearchCV in conjunction with SMOTE for a 30 % division between training and testing. Overall, MLP had 98.31 % accuracy and SVM had 95 % accuracy. The intention of the framework mentioned in [4] is to enhance the precision of current methods used to forecast the intensity of Parkinson's disease and also to deliver a more effective and dependable outcome. However, since the framework solely depends on the examination of speech signals, it may not always be adequate for accurately predicting the severity of Parkinson's disease in every situation.

The main aim of the study described in [5] was to make use of neural networks and machine learning techniques to forecast the extent of Parkinson's disease by analyzing voice impairment in patients. According to the study, it was discovered that the suggested remedy was able to differentiate between individuals in the early stages of Parkinson's disease. The paper in [6] examines the effects of Parkinson's disease on voice alterations and the potential consequences associated with them. The writers investigate the anatomical and physiological factors behind the voice alterations commonly observed in individuals with Parkinson's disease, while also analyzing how these factors are interconnected. The article examines the vocalization aspect and the potential dysfunction that might arise in individuals with Parkinson's disease prior to it impacting their limbs and fine motor skills. The highlighted aspect emphasizes that alterations in voice may be used as an early indication of a disease or as a sign of how the disease is advancing. In [7] the objective was to use wearable devices integrating inertial sensors and machine learning to detect and predict the occurrence of freezing of gait (FOG) episodes in Parkinson's disease.

In [8] the objective was to enhance parkinson's disease prediction using machine learning and feature selection methods. Naïve Bayes algorithm obtained the best performance in enhancing the detection of Parkinson's disease but larger datasets were needed to validate the effectiveness of the proposed approach. The purpose of the study mentioned in reference [9] was to tackle the intricacies of Parkinson's disease. This was done by combining various types of data (unlabeled, multimodal, and longitudinal) and gaining a thorough comprehension of the paths and aspects through which individual's progress. According to the study, it has been found that clinical data collected over a period of time can be used to anticipate distinct subtypes of Parkinson's disease, based on the progression stage. The research conducted in [10] observed the deterioration of cognitive abilities over various durations, although only a limited number of studies examined cognitive decline for an extended period of more than four years.

The majority of research primarily looked at the assessment of changes in overall cognitive function using the Mini-Mental State Examination. However, there was significant variation in the utilization of neuropsychological tests across different studies. There was only one research study that examined how well patients performed in various cognitive areas such as executive function, language, memory, working memory, attention, and visual-spatial function. This study also followed agreed-upon guidelines to determine if patients had any cognitive impairments. Studies having follow - up intervals of at least four years had been the only ones where noticeable impact sizes were found. The findings highlight the importance of evaluating bigger groups of individuals with Parkinson's disease for extended periods of time, using a thorough set of cognitive tests.

The paper presents a review of previous research that has examined speech patterns in Parkinson's disease along with other synucleinopathies in [11]. The authors look at how these research studies have discovered particular

speech indicators that may be utilized to assist in the identification and tracking of these disorders. The drawbacks of existing methods for examining speech patterns are also emphasized, which involve the requirement of specialized equipment and trained individuals. The next section describes about the stacking classifier used to predict the Parkinson's disease.

III.PREDICTION OF PARKINSON'S DISEASE USING SVM AND KNN

The proposed approach involves integrating various types of data, such as maximum vocal fundamental frequency, minimum vocal fundamental frequency, signal fractal scaling exponent, correlation dimension etc. Once the model is trained, it will be evaluated on a held-out test set of Parkinson's data. The model's accuracy will be measured by its ability to correctly classify whether a person has Parkinson's or not in the test set. The expected outcome of this study is a highly accurate classifier based model that can be used to identify Parkinson's disease in real time. The dataset is loaded from csv file to a pandas data frame and this dataset is stored in Parkinson's data. The data is grouped by status that is the people who are affected and unaffected by Parkinson's and the mean of every column is taken. The data is pre-processed and then the status columns are separated from the dataset. Since status column is our target variable, we need it separately as we are going to classify whether a person is healthy or not using status.

Data standardization is done to ensure that all the columns fall under a particular range but won't change the meaning conveyed by the columns. The fit function will help the standard scale function what is the nature of data. KNN Algorithm is employed in this model as it has 0 false positives and 0 false negatives. SVC Algorithm is also employed even though it has 4 false positives but has an accuracy of 95%. The algorithms mentioned previously are stacked on top of one another using the Stacking Classifier method. We do this to train multiple models and based on their combined output, it builds a new model with improved performance.

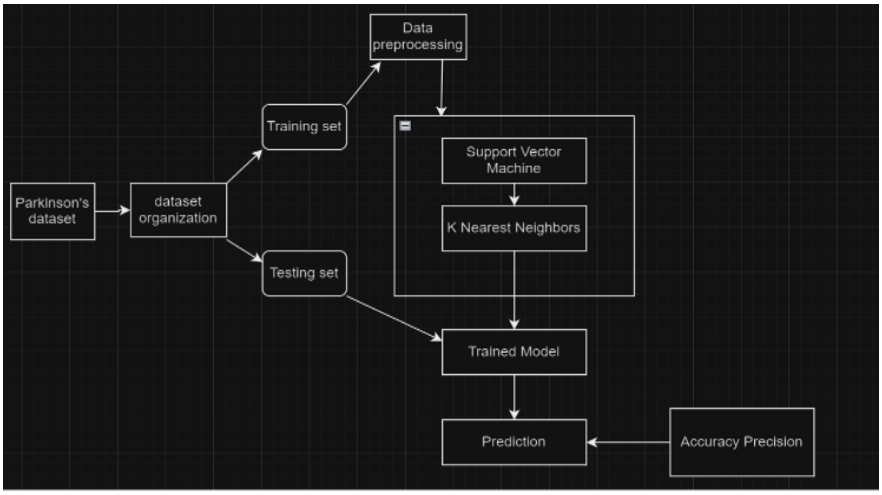

Figure 1. Stacking Classifier

3.1 Support Vector Machines:

Vladimir Vapnik and Alexey Chervonenkis were the pioneers who at first introduced support vector machines (Chervonenkis, 2013) (Vapnik, Learn, Guyon, 1and1 1995 n.d.). SVM is a machine learning technique that has the capability to address both nonlinear and linear issues. It delivers exceptional results for both classification and regression tasks. The SVM classification method searches for the best separable hyperplane to categorize the dataset into 2 classes (Smola and Schölkopf, 2004). The model is capable of predicting and determining instances of noisy data issues. A supervised learning algorithm is known as a Support Vector Machine. An SVM is used to classify the data and create a hyperplane with N dimensions, dividing it into k categories.

These models resemble neural networks in many respects. A dataset with N dimensions is considered. The support vector machine maps the training data onto a space with N dimensions. The data points used for training are subsequently split into k regions using hyperplanes that have n distinct dimensions, based on their respective labels. After the testing stage, the test points are displayed on the N-dimensional plane. The points are accurately categorized according to their specific regions. After splitting our dataset into a train set as well as a test set, we went on to train the SVM model using the train set.

3.2 KNN Algorithm:

The traditional K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithm that is primarily employed for classification tasks. The algorithm relies on a variable parameter k, which signifies the count of the "closest neighbors." The KNN algorithm operates by figuring out the closest data point(s) or neighbor(s) from a set of training data for a given query. The data points that are closest to the query point are found based on their nearest distances from the point. Once the k closest data points are identified, the system applies a majority voting method to determine the class that has the highest frequency. According to the ruling, the classification of the query as final is based on the class that had the highest frequency.

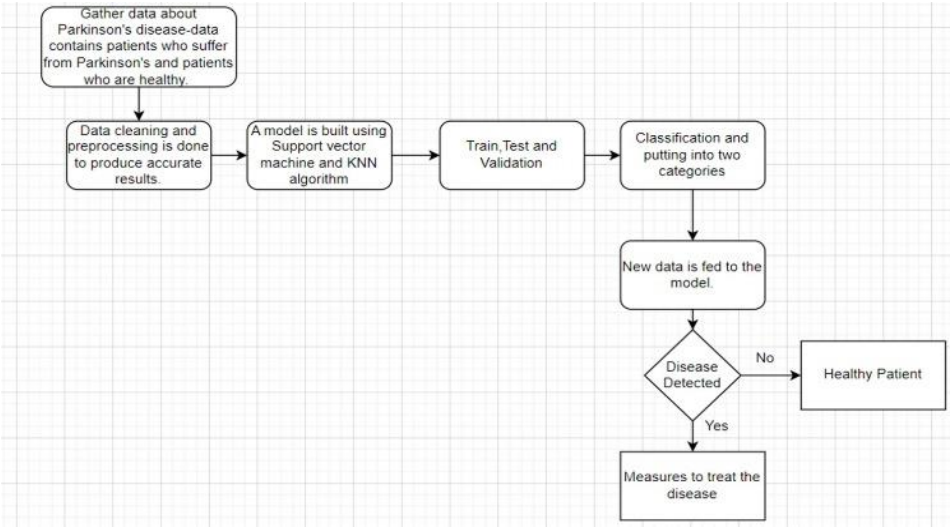

Figure 2. Architecture Diagram

3.3 Dataset Description:

The dataset consists of various biomedical voice measurements collected from 31 individuals, 23 of whom have Parkinson's disease (PD). Every row in the table represents a different voice recording from one of 195 individuals, with each column representing a specific voice measurement. The information in the "status column" is meant to differentiate between individuals who are healthy and those who have PD, with a value of 0 indicating health and 1 indicating PD. The characteristics in the dataset are “name” (with recording number), MDVP:FO represents average vocal fundamental frequency, MDVP:Fhi represents maximum vocal fundamental frequency, and MDVP:Flo refers to the minimum vocal fundamental frequency, and other attributes include several measures of variation in fundamental frequency.

name	MDVP:Fo	MDVP:Fhi	MDVP:Flo	MDVP:Jitte	MDVP:Jitter(	MDVP:R	MDVP:P	Jitter:D	MDVP:Shin	MDVP:Shimm	Shimmer:A
phon_R01_	119.992	157.302	74.997	0.00784	7E-05	0.0037	0.00554	0.0111	0.04374	0.426	0.02182
phon_R01_	122.4	148.65	113.819	0.00968	8E-05	0.00465	0.00696	0.0139	0.06134	0.626	0.03134
phon_R01_	116.682	131.111	111.555	0.0105	9E-05	0.00544	0.00781	0.0163	0.05233	0.482	0.02757
phon_R01_	116.676	137.871	111.366	0.00997	9E-05	0.00502	0.00698	0.0151	0.05492	0.517	0.02924
phon_R01_	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.00908	0.0197	0.06425	0.584	0.0349
phon_R01_	120.552	131.162	113.787	0.00968	8E-05	0.00463	0.0075	0.0139	0.04701	0.456	0.02328
phon_R01_	120.267	137.244	114.82	0.00333	3E-05	0.00155	0.00202	0.0047	0.01608	0.14	0.00779
phon_R01_	107.332	113.84	104.315	0.0029	3E-05	0.00144	0.00182	0.0043	0.01567	0.134	0.00829
phon_R01_	95.73	132.068	91.754	0.00551	6E-05	0.00293	0.00332	0.0088	0.02093	0.191	0.01073
phon_R01_	95.056	120.103	91.226	0.00532	6E-05	0.00268	0.00332	0.008	0.02838	0.255	0.01441
phon_R01_	88.333	112.24	84.072	0.00505	6E-05	0.00254	0.0033	0.0076	0.02143	0.197	0.01079
phon_R01_	91.904	115.871	86.292	0.0054	6E-05	0.00281	0.00336	0.0084	0.02752	0.249	0.01424
phon_R01_	136.926	159.866	131.276	0.00293	2E-05	0.00118	0.00153	0.0036	0.01259	0.112	0.00656
phon_R01_	139.173	179.139	76.556	0.0039	3E-05	0.00165	0.00208	0.005	0.01642	0.154	0.00728
phon_R01_	152.845	163.305	75.836	0.00294	2E-05	0.00121	0.00149	0.0036	0.01828	0.158	0.01064
phon_R01_	142.167	217.455	83.159	0.00369	3E-05	0.00157	0.00203	0.0047	0.01503	0.126	0.00772
phon_R01_	144.188	349.259	82.764	0.00544	4E-05	0.00211	0.00292	0.0063	0.02047	0.192	0.00969
phon_R01_	168.778	232.181	75.603	0.00718	4E-05	0.00284	0.00387	0.0085	0.03327	0.348	0.01441
phon_R01_	153.046	175.829	68.623	0.00742	5E-05	0.00364	0.00432	0.0109	0.05517	0.542	0.02471
phon_R01_	156.405	189.398	142.822	0.00768	5E-05	0.00372	0.00399	0.0112	0.03995	0.348	0.01721

Figure 3. Sample of the Dataset

3.4 Model summary:

The study described the use of SVM and KNN for early disease identification in Parkinson's disease. The suggested method involves stuffing and classifying the obtained data using classification method.by combining inputs from Parkinson's sufferers and healthy individuals, the model produces a reliable results for the data sets obtained as input.

IV.RESULTS

The results obtained from the KNN model when used on the dataset produces an accuracy of 82%, Matthews Correlation coefficient of 52% and F1 score of 83%. The results from the SVM model produces an accuracy of 87%, Matthews Correlation coefficient of 56% and F1 score of 84%. The Matthews Correlation coefficient stands out as a valuable classification metric used for condensing the information contained in an error matrix.

Tuning the parameters of SVM can be challenging and time consuming. KNN's performance can go down as the dimensionality of feature space increases. To counteract these effects stacking classifier is employed. A stacking classifier is an ensemble machine learning technique that amalgamates various classification models into a single and more powerful model. This amalgamation often results in enhanced performance because the composite model leverages the strengths of each constituent model. Stacking classifier is used to combine the predictions of SVM model with those of KNN. The individual models are trained on distinct partitions of the dataset, after which their predictions are amalgamated through the utilization of a meta-classifier.

The stacking classifier using KNN and SVM used on the test dataset produces an accuracy of 82%, Matthews Correlation Coefficient of 56% and F1 score of 84%. This has led to improved performance of the model compared to individual base models. It also has made the model more robust to overfitting. These factors show significant improvement of the stacking classifier model when compared to KNN and SVM model when used alone.

4.1 Confusion Matrix

The confusion matrix is a matrix which provides a summary of just how well a machine learning model performs on a specific dataset for testing purposes. It is commonly employed to evaluate the effectiveness of classification models, that seek to forecast a categorical tag for every given input example. The matrix shows the count of accurate positive predictions, accurate negative predictions, incorrect positive predictions, and incorrect negative predictions from the model on the test data. The binary classification matrix will consist of a table with dimensions 2X2.

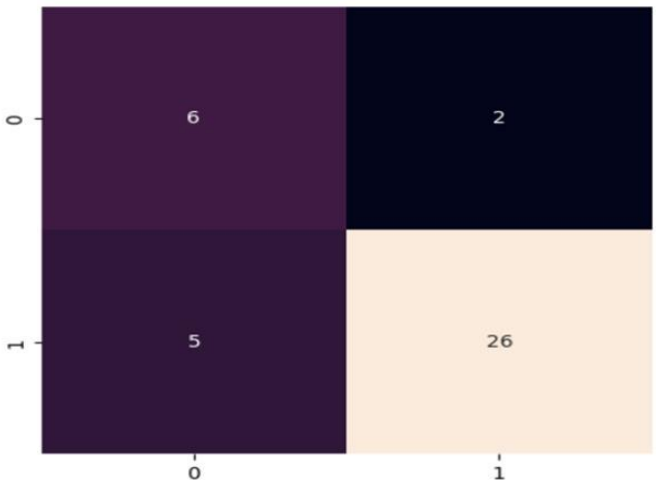

Figure 4. Confusion Matrix

The confusion matrix was generated to look at the accuracy of the SVM and KNN models in predicting Parkinson's disease. There are a total of 31 accurate negative results, 26 accurate positive results, 6 inaccurate negative results, and 2 inaccurate positive results. The level of precision is around 82%.

V.COMPARATIVE ANALYSIS

The stacking classifier using KNN and SVM used on the test dataset produces an accuracy of 82%, Matthews Correlation Coefficient of 56% and F1 score of 84%. This has led to improved performance of the model compared to individual base models. It also has made the model more robust to overfitting. These factors show significant improvement of the stacking classifier model when compared to KNN and SVM model when used alone.

VI.CONCLUSION

The main goal is to make use of machine learning methods, specifically SVM and KNN, in combination, to accurately detect Parkinson's disease thanks to a low rate of mistakes. The SVM model was trained and tested on 31 patients with 195 voice recordings in the study using the dataset. The model was found to have an 82 % accuracy rate in determining Parkinson's disease based on the results. The study additionally suggests that upcoming studies could concentrate on utilizing different sets of data and characteristics to categorize patients and detect the specific phases of Parkinson's disease. The research recognizes the drawbacks of employing a binary characteristic for categorizing patients and proposes that upcoming studies are able to incorporate more sophisticated elements to enhance the precision of the model. The suggested system is easy to understand, efficient, and can be used on any platform, increasing its effectiveness. The research suggests that employing machine learning methods to identify Parkinson's disease at an early stage shows great potential. This approach holds promise in enhancing patient results and offering more effective treatment choices for this incapacitating disorder.

REFERENCES

[1] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease

[2] Makarious, Mary B et al. “Multi-modality machine learning predicting Parkinson's disease.” *NPJ Parkinson's disease* vol. 8,1 35. 1 Apr. 2022, doi:10.1038/s41531-022-00288-w

[3] Boutet, Alexandre et al. “Predicting optimal deep brain stimulation parameters for Parkinson's disease using functional MRI and machine learning.” *Nature communications* vol. 12,1 3043. 24 May. 2021, doi:10.1038/s41467-021-23311-9

[4] Alshammri, Raya et al. “Machine learning approaches to identify Parkinson's disease using voice signal features.” *Frontiers in artificial intelligence* vol. 6 1084001. 28 Mar. 2023, doi:10.3389/frai.2023.1084001

[5] Grover, S., Bhartia, S., Yadav, A., & Seeja, K. R. (2018). Predicting severity of Parkinson’s disease using deep learning. *Procedia computer science*, *132*, 1788-1794.

[6] P. Raundale, C. Thosar and S. Rane, "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)*, Belagavi, India, 2021, pp. 1-5, doi: 10.1109/INCET51464.2021.9456292.

[7] Ma, A., Lau, K. K., & Thyagarajan, D. (2020). Voice changes in Parkinson's disease: what are they telling us? J. Clini.Neurosci. 72, 1–7. doi: 10.1016/j.jocn.2019.12.029

[8] Borzì, Luigi et al. “Prediction of Freezing of Gait in Parkinson's Disease Using Wearables and Machine Learning.” *Sensors (Basel, Switzerland)* vol. 21,2 614. 17 Jan. 2021, doi:10.3390/s21020614

[9] Saeed, Faisal & Al-Sarem, Mohammed & Al-Mohaimeed, Muhannad & Emara, Abdelhamid & Boulila, Wadii & Alasli, Mohammed & Ghabban, Drfahad. (2022). Enhancing Parkinson's Disease Prediction Using Machine Learning and Feature Selection Methods. Computers, Materials & Continua. 71. 5639-5658. 10.32604/cmc.2022.023124.

[10] Dadu, Anant et al. “Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts.” *NPJ Parkinson's disease* vol. 8,1 172. 16 Dec. 2022, doi:10.1038/s41531-022-00439-z

[11] Roheger, Mandy et al. “Progression of Cognitive Decline in Parkinson's Disease.” *Journal of Parkinson's disease* vol. 8, 2 (2018): 183-193. doi:10.3233/JPD-181306

[12] Kouba, Tomáš et al. “Study protocol for using a smartphone application to investigate speech biomarkers of Parkinson's disease and other synucleinopathies: SMARTSPEECH.” *BMJ open* vol. 12,6 e059871. 30 Jun. 2022, doi:10.1136/bmjopen-2021-059871

[13] Shafi, Muhtasim & Ahmed, Fizar. (2022). Parkinson’s Disease Detection Analysis through Machine Learning Approaches.

[14] M. Mamun, M. I. Mahmud, M. I. Hossain, A. M. Islam, M. S. Ahammed and M. M. Uddin, "Vocal Feature Guided Detection of Parkinson’s Disease Using Machine Learning Algorithms," *2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, NY, USA, 2022, pp. 0566-0572, doi: 10.1109/UEMCON54665.2022.9965732.

[15] Amit Kumar Patra *et al* 2019 *J. Phys.: Conf. Ser.* 1372 012041.

[16] Dadu, A., Satone, V., Kaur, R. *et al.* Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *npj Parkinsons Dis.* 8, 172 (2022). https://doi.org/10.1038/s41531-022-00439-z.

[17] P. Raundale, C. Thosar and S. Rane, "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)*, Belagavi, India, 2021, pp. 1-5, doi: 10.1109/INCET51464.2021.9456292.

[18] Kavita Bhatt, N. Jayanthi, Manjeet Kumar, High-resolution superlet transform based techniques for Parkinson's disease detection using speech signal, Applied Acoustics,Volume 214, 2023,109657,SSN 0003-682X, https://doi.org/10.1016/j.apacoust.2023.109657.