www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 274-279

e-ISSN : 2583-1062
Impact Factor : 7.001

CURATING IDEAL DATASETS FOR PARKINSON'S DISEASE DETECTION WITH MACHINE LEARNING TECHNIQUES

N Sindhujaa[1], M K Saranya[2], Mr S Suseendran[3]

[1,2]PG Scholar, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam.

[3]Assistant Professor II, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam.

DOI: https://www.doi.org/10.58257/IJPREMS36646

ABSTRACT

A neurodegenerative disorder called Parkinson's disease primarily affects the dopamine-producing ("dopaminergic") neurons in a specific area of the brain called the substantia. The symptoms can be either movement-related (or "motor"), such as postural instability, trouble walking, speech problems, swallowing difficulties, etc., or they can be non-motor (or unconnected to movement). The onset of Parkinson's disease typically occurs at age 60 and naturally rises with age. While treatments and medications can lessen symptoms, there is no known cure. PD has doubled during the last 25 years, according to the WHO. Feed forward neural networks (FNN), a type of machine learning (ML) approach, have demonstrated significant promise in improving diagnostic accuracy. FNN performance is highly dependent on the caliber and applicability of the datasets employed. Effective feature extraction methods specific to PD datasets are evaluated in this study, with a view on significance of choosing the relevant features, such as tremor frequency, gait abnormalities, speech abnormalities, and patient demographics, as these factors have a direct influence on FNN learning ability. The study looks at data pre-processing methods such dimensionality reduction, feature scaling, and normalization to improve the model's effectiveness with 95% accuracy. We show through comparative analysis how well selected datasets can greatly increase the accuracy of PD identification. Our research aids researchers and medical professionals studying neurodegenerative illnesses create more trustworthy machine learning tools.

Keywords: Parkinson's disease, Machine Learning, Neural Networks, Neurodegenerative disorder, Motor Symptoms, Non Motor symptoms

1. INTRODUCTION

Parkinson's disease (PD), a common nervous disorder that significantly impairs quality of life, affects millions of individuals globally. Bradykinesia, rigidity, and tremors are some of the movement symptoms of PD, which affects about 1% of adults over 60. This number is expected to rise as the world's population ages (Goyal et al., 2020). Due to its diverse character, which can cause symptoms to vary widely from patient to patient and often overlap with other neurological conditions, Parkinson's disease is difficult to diagnose (Ayaz et al., 2022). Early PD detection is crucial because it allows for timely symptom management and treatment, perhaps slowing the progression of the disease (Ayaz et al., 2022). Traditional diagnostic methods rely on clinical examinations, which often take place only after significant neuronal loss has already taken place (Khaskhoussy & Ayed, 2022). This delay in diagnosis highlights the need for innovative approaches that leverage state-of-the-art technologies to enable earlier and more accurate detection (Sura et al., 2023).

Significant progress in AI and ML has made it feasible to design automated diagnostic systems and opened up new avenues for the study of complex datasets (Ayaz et al., 2022). (Ayaz et al., 2022). In this study, we aim to determine which datasets are best for using machine learning methods to diagnose Parkinson's disease. Our goal is to develop the precision and dependability of PD analysis models by applying methodical assessment and selection process to pertinent elements from various data sources. The results of employing curated datasets for Parkinson's disease diagnosis, feature selection methods, and several machine learning classifiers will all be covered in this paper. By supporting ongoing work to establish early diagnostic and intervention options for Parkinson's disease, we want to improve patient outcomes and quality of life (Govindu & Palwe et al 2023, Sura et al., 2023).

1.1 Parkinson's disease symptoms: When the muscles stay tense and constricted, it can cause pain or stiffness, which is one of the common symptoms of Parkinson's disease. Additional signs and symptoms include tremor, or shaking, which typically begins in the hands, feet, or jaws. A disease known as bradykinesia affects a person's mobility and makes daily tasks difficult. Additional symptoms include sadness, skin problems, sleep problems, and speech difficulties. Another indication of postural instability is feeling imbalanced, which raises the risk of falling.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 274-279

e-ISSN : 2583-1062
Impact Factor : 7.001

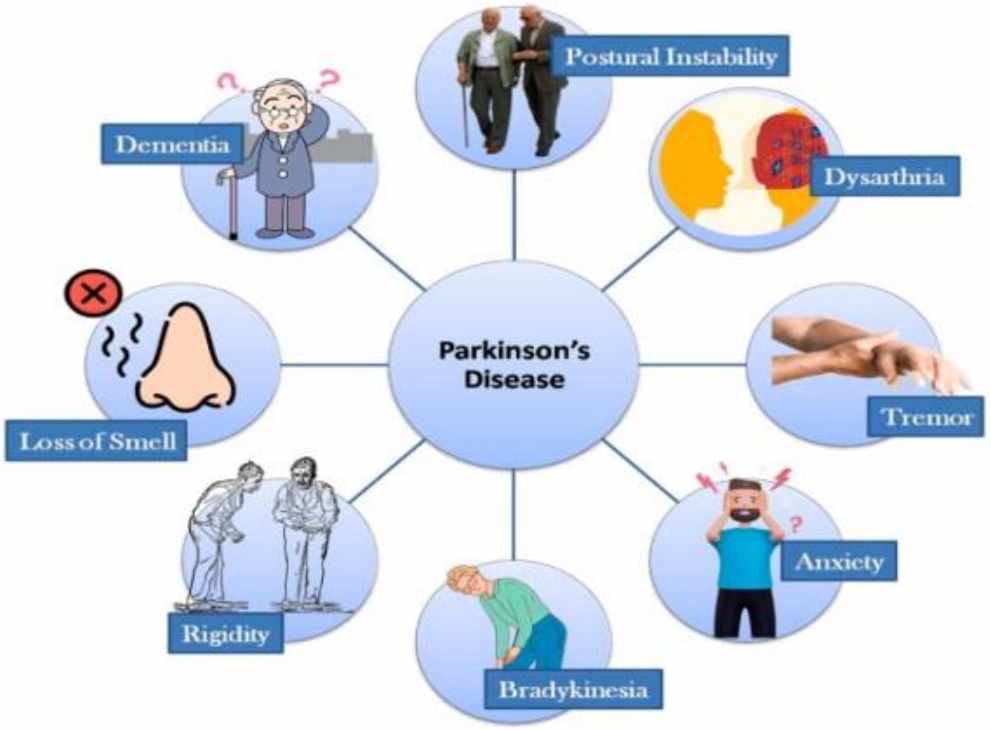

Figure.1 Parkinson's disease Symptoms

2. RELATED WORKS

2.1 Importance of Early Detection: Early diagnosis of Parkinson's disease (PD) is crucial for its efficient management and treatment. Voice issues, for instance, have been shown to manifest years before typical motor symptoms. The necessity for efficient, non-invasive diagnostic methods that can be applied in therapeutic contexts is thus highlighted. Global aging of the population further emphasizes the necessity for reliable early detection techniques (Govindu & Palwe, 2023; Goyal et al., 2020).

2.2 Datasets Utilized in PD Detection

Voice analysis is one potential technique for diagnosing Parkinson's disease. Voice recordings from PD patients and healthy controls make up the collection, which was created by Max Little of the University of Oxford. Research has shown that this dataset is suitable for training machine learning models and achieving high accuracy rates, with up to 95% accuracy reported in later studies using advanced feature extraction techniques (Ayaz et al., 2022; Goyal et al., 2020).

The effectiveness of audio as a non-invasive biomarker for Parkinson's disease (PD) identification was demonstrated by Aditi Govindu and Sushila Palwe (2023) using a Random Forest classifier trained on 22 variables of MDVP audio data (Govindu & Palwe, 2023).

This research demonstrates the potential of remote detection techniques, which allow patients to track audio using their mobile phones and provide new lease of life for those with mobility disabilities (Govindu & Palwe, 2023).

2.3 Machine Learning Techniques: Various ML techniques have been applied to aforementioned datasets, with study results showing varying performance levels. There has been widespread use of traditional classifiers such as Support Vector Machines (SVM), Random Forests, and K-Nearest Neighbors (KNN). Recent advancements in deep learning have also shown promise with feedforward neural networks (FNNs) being utilized for feature extraction and classification tasks (Ayaz et al., 2022; Goyal et al., 2020; Pahuja & Prasad, 2022).

According to Doneti Sowmya and Dodla Kavya (2022), for instance, a machine learning-based approach to diagnosing Parkinson's disease (PD) combines a variety of symptoms, including handwriting samples, tremors, and gait. According to Ayaz et al. (2022), their work focused on preparing and combining data from many datasets, selecting 30 features relevant to PD diagnosis, and utilizing supervised classifiers like SVM in order to improve accuracy and reliability.

3. Proposed System: The proposed approach develops a framework for selecting datasets that improve the use of machine learning techniques for Parkinson's disease identification. In this part, we outline the methodology we employed in our study to identify Parkinson's disease using the FeedforwardNeural Network (FNN) algorithm.

www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 274-279

e-ISSN : 2583-1062
Impact Factor : 7.001

3. METHOD

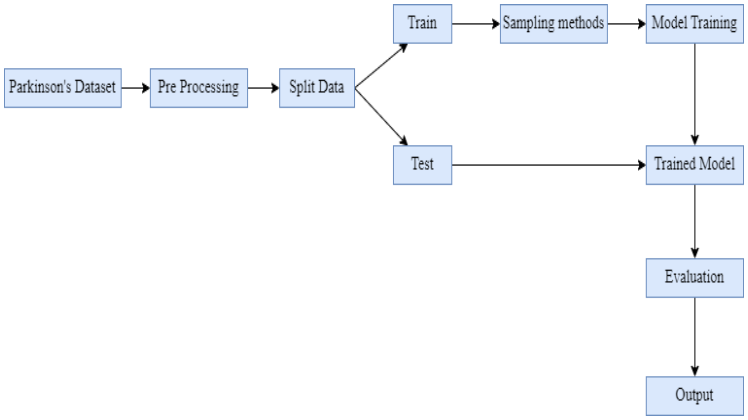

Figure2.Flow diagram for proposed system

3.1.1 Data Collection:

In order to guarantee consistency across all assessments, the study will employ the same group of PD patients and healthy controls. We will record voice samples in a calm environment using high-end audio equipment. After reading a typical text extract, participants will be asked to engage in spontaneous conversations in order to capture real speech patterns. Every recording will be between one and three minutes long in order to ensure that a variety of speech features are captured for consideration.

Table 1. Voice Data Attributes

Attribute	Purpose
Name	Data is stored in ASCII CSV format where patient name and recording number is stored
MDVP: Fo (Hz)	Fundamental frequency of pitch period
MDVP: Fhi (Hz)	Upper limit of fundamental frequency or maximum threshold of voice modulation
MDVP: Flo (Hz)	Lower limit or minimal vocal fundamental frequency
MDVP: Jitter, Abs, RAP, PPQ, DDP	These are various Kay Pentax's multi-dimensional voice program (MDVP) measures. MDVP is a traditional measure of frequency of vibrations in vocal folds at pitch period to vibrations at start of next cycle called pitch mark [25]
Jitter and Shimmer	Measures of absolute difference between frequencies of each cycle, after normalizing the average
NHR and HNR	Signal to noise and tonalratio measures, that indicate robustness of environment to noise
Status	0 indicates healthy person while 1 indicates PWP.
D2	Correlation dimension is used to identify dysphonia in speech using fractalobjects. It is a nonlinear, dynamic attribute.
RPDE	Recurrence Period Density Entropy quantifies the extent to which signal is periodic
DFA	Detrended Fluctuation Analysis or DFA measures the extent of stochastic self-similarity of noise in speech signals.
PPE	Pitch Period entropy is used to assess abnormal variations in speech on a logarithmic scale
Spread1, spread2	Analysis of extent or range of variations in speech with respect to MDVP: Fo(Hz)

3.1.2 Data Preprocessing

Data preprocessing is necessary to ensure the quality and usability of the datasets.

Audio preprocessing will be done using programs like Audacity, which will smooth out the recordings by removing background noise and adjusting volume levels. Programs like Praat or Librosa will then be used to extract important acoustic features from voice samples, including pitch, tone, and intensity. In addition to speech rate and pause duration, other metrics such as shimmer (amplitude variation) and jitter (frequency variation) will be computed. It will be simpler to identify speech patterns that may indicate Parkinson's disease thanks to these traits.

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 274-279
www.ijprems.com
editor@ijprems.com
e-ISSN : 2583-1062
Impact Factor : 7.001

3.1.3 Feature Extraction

The project aims to extract different characteristics from voice data. Vocal analysis will focus on acoustic metrics such as mean pitch and pitch range, as well as temporal features such as speech rate and pause duration, to identify patterns associated with Parkinson's disease.

3.1.4 Model Training

The model has been trained using the FNN approach. To facilitate training, the dataset is divided into segments. To assess how well the model performs, the dataset is split into training (80%) and testing (20%) sets. The FNN's architecture dictates the number of layers, the number of neurons in each layer, the activation functions (such as softmax for the output layer and ReLU for hidden levels), and dropout layers to avoid overfitting.

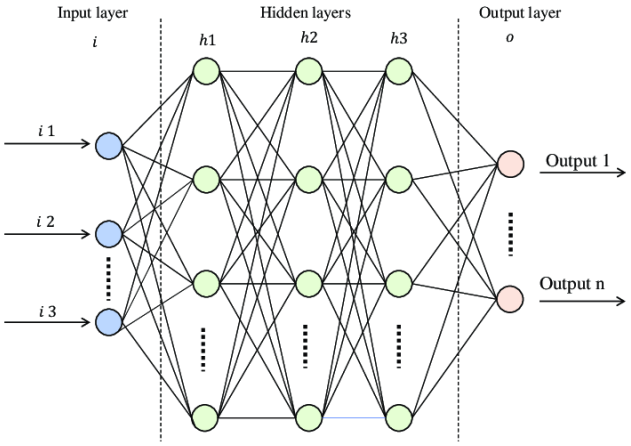

Figure 3. FNN Architecture

3.1.5 Training and Validation

The data will be divided into three sets for the training phase: 70% for training, 15% for validation, and 15% for testing. Cross-validation methods will be applied to optimize model parameters and prevent overfitting. Important metrics like accuracy, sensitivity, specificity, and the area under the AUC-ROC curve will be utilized to assess the model's performance in order to provide a comprehensive assessment of its ability to correctly classify and differentiate between Parkinson's patients and healthy controls.

3.1.7 Evaluation Metrics

When working with medical data, the metrics provide useful information regarding the model's effectiveness and reliability, as wrong categorization can have significant consequences. Models can be analyzed using metrics like F1-score, recall, accuracy, and precision. For samples that have been satisfactorily categorized, TP (True Positive) and TN (True Negative) are used. In contrast, cases that were incorrectly classified are denoted by FP (False Positive) and FN (False Negative).

Accuracy: Accuracy is a crucial measure that is defined as the ratio of correctly identified examples to all instances. The following formula is used to compute:

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP} \quad (1)$$

Precision: Precision measures how many accurate positive predictions there are out of all the model's positive predictions. It's explained as:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

Recall: Shows the percentage of actual positive cases that the model picked up correctly.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

F1 Score: In order to balance recall and precision, this score takes the harmonic mean of the two measurements. The following formula is used to compute it:

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Confusion Matrix: The model's performance across different classes (PD vs. healthy controls, for instance) is displayed using confusion matrix. To provide a broad overview of a classification model's performance, a confusion matrix is a table that displays the percentages of false positives, false negatives, true positives, and true negatives.

ROC Curve and AUC: The Receiver Operating Characteristic (ROC) curve is plotted, and the Area Under the Curve (AUC) is calculated, to evaluate the model's ability to distinguish between classes.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 11, November 2024, pp : 274-279
e-ISSN : 2583-1062
Impact Factor : 7.001

4. RESULT AND DISCUSSION

The Feedforward Neural Network (FNN) algorithm, an efficient supervised learning technique, was utilized in this study to identify Parkinson's disease. Through training the FNN on these multimodal datasets, we were able to create a model that can identify intricate patterns in various domains, which enhances the model's accuracy by 95% in diagnosing Parkinson's disease.

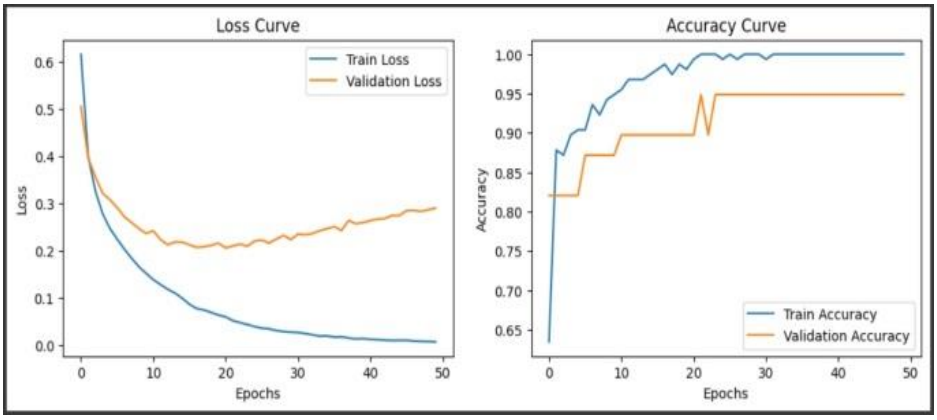

Figure.4 Accuracy and Loss curve for Voice data

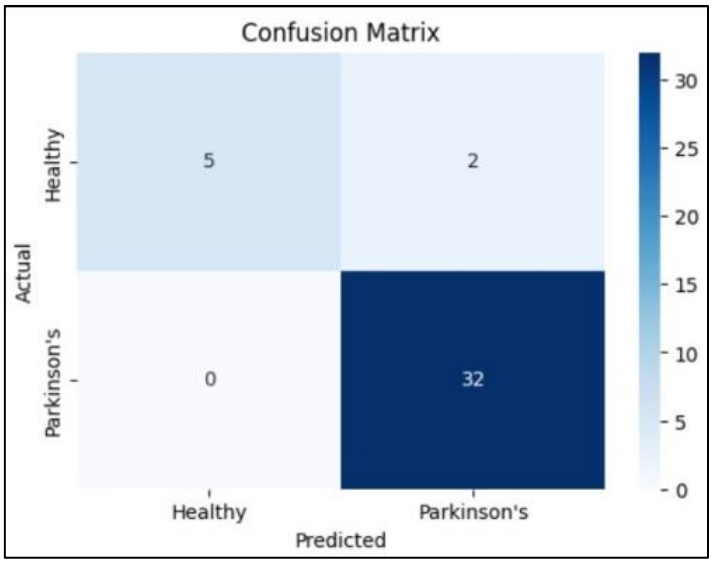

Figure.5 Confusion Matrix for Voice data

4.1 Discussion: The voice dataset performs better, which is explained by the obvious connection between vocal characteristics and motor function impairments in Parkinson's disease. Voice analysis picks up on subtle differences in speech patterns that may not be apparent with other datas. Also, the richness of the voice dataset, which includes a range of vocal characteristics, likely contributed to the model's high accuracy.

5. CONCLUSION

The importance of choosing the optimal datasets for machine learning-based Parkinson's disease detection is highlighted by this study. By demonstrating that voice data outperforms in terms of accuracy, the findings underscore the need for targeted data gathering efforts in subsequent studies. This approach improves patient outcomes, increases the accuracy of the diagnosis, and expands our understanding of the underlying causes of the illness.

5.1 Future Work

The work can be carried out by using the preferred dataset for early prediction. This process can also be done with more number of clinical data. This work can also be implemented with new deep learning methods.

6. REFERENCES

[1] Goyal, Jinee, PadmavatiKhandnor, and Trilok Chand Aseri. "A Comparative Analysis of Machine Learning classifiers for Dysphonia-based classification of Parkinson's Disease." International Journal of Data Science and Analytics 11 Springer (2021): 69-83.

[2] Ayaz, Zainab, et al. "Automated methods for diagnosis of Parkinson's disease and predicting severity level." Neural Computing and Applications 35.20 (2023): 14499-14534.

[3] [3] Pahuja, Gunjan, and Bhanu Prasad. "Deep learning architectures for Parkinson's disease detection by using multi-modal features." Computers in Biology and Medicine 146 (2022): 105610.

[4] [4] Quan, Changqin, Kang Ren, and ZhiweiLuo. "A deep learning based method for Parkinson's disease detection using dynamic features of speech." IEEE Access 9 (2021): 10239-10252.

[5] [5] Abdullah, SuraMahmood, et al. "Deep transfer learning based parkinson's disease detection using optimized feature selection." IEEE Access 11 (2023): 3511-3524.

[6] [6] Govindu, Aditi, and SushilaPalwe. "Early detection of Parkinson's disease using machine learning."Procedia Computer Science 218 Elsevier(2023): 249-261.

[7] [7] Khaskhoussy, Rania, and Yassine Ben Ayed. "Speech processing for early Parkinson's disease diagnosis: machine learning and deep learning-based approach." Social Network Analysis and Mining 12.1 Springer (2022): 73.

[8] [8] Braga, Diogo, et al. "Automatic detection of Parkinson's disease based on acoustic analysis of speech." Engineering Applications of Artificial Intelligence 77 Elsevier (2019): 148-158.

[9] [9] Aghav-Palwe S, Mishra D (2020) "Statistical tree-based feature vector for content-based image retrieval." Int J Comput Sci Eng 2020 https://doi.org/10.1504/IJCSE.2020.106868

[10] [10] Aghav-Palwe, S., Mishra, D. (2019). "Color Image Retrieval Using Statistically Compacted Features of DFT Transformed Color Images. In:Bhatia, S., Tiwari, S., Mishra, K., Trivedi, M. (eds) Advances in Computer Communication and Computational Sciences. Advances in Intelligent Systems and Computing," vol 760. Springer, Singapore. https://doi.org/10.1007/978-981-13-0344-9_29

[11] [11] D. Yadav and I. Jain (2022), "Comparative Analysis of Machine Learning Algorithms for Parkinson's Disease Prediction," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1334-1339, doi:10.1109/ICICCS53718.2022.9788354

[12] [12] D. V. Rao, Y. Sucharitha, D. Venkatesh, K. Mahamthy and S. M. Yasin (2022), "Diagnosis of Parkinson's Disease using Principal Component Analysis and Machine Learning algorithms with Vocal Features," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), 2022, pp. 200-206, doi: 10.1109/ICSCDS53736.2022.9760962.

[13] [13] Y. Guan (2021), "Application of logistic regression algorithm in the diagnosis of expression disorder in Parkinson's dis ease," 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), 2021, pp. 1117-1120,

[14] [14] 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.

[15] [15] W. Wang, J. Lee, F. Harrou and Y. Sun, "Early Detection of Parkinson's Disease Using Deep Learning and Machin e Learning," in IEEE Access, vol. 8, pp. 147635-147646, 2020, doi: 10.1109/ACCESS.2020.3016062.Aditi Govindu et al. / Procedia Computer Science 218 (2023) 249–261 261 Author name / Procedia Computer Science 00 (2019) 000–000 13

[16] [16] 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:

[17] 10.1109/ITNEC52019.2021.9586915.

[18] [17] P. Raundale, C. Thosar and S. Rane (2021), "Prediction of Parkinson's disease and severity of the disease using Machine Learn ing and Deep Learning algorithm," 2021 2nd International Conference for Emerging Technology (INCET), pp. 1-5, doi:10.1109/INCET51464.2021.9456292.

[19] [18] 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.

[20] [19]Aditi Govindu, Sushila Palwe (2023), "Early detection of Parkinson's disease using machine learning", Procedia Computer Science 218 (2023) 249–261 261,https://creativecommons.org/licenses/by-nc-nd/4.0/

[21] [20] Chakraborty, Sabyasachi, et al. "Parkinson's disease detection from spiral and wave drawings using convolutional neural networks: A multistage classifier approach." 2020 22nd International Conference on Advanced Communication Technology (ICACT). IEEE, 2020.