Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Streamlit-based Web Application for Parkinson's Detection using Machine Learning

Revathy S.P.[1], Sindhuja M.[2], Jayashree R.[3]

[1]Assistant Professor,[2,3]Students, Department of Information Technology, Velammal Engineering College, India.

Email: [1]revathy.sp@velammal.edu.in, [2]sindhujaaam00@gmail.com, [3]jayashreeramalingam.05@gmail.com

Abstract

Parkinson's Disease (PD) is a progressive neurodegenerative disorder that impacts motor skills, including tremors, bradykinesia, and rigidity, affecting millions globally. Early diagnosis is essential for effective treatment yet remains challenging as the symptoms overlap with other conditions and the limitations of conventional diagnostic methods. This study presents a diagnostic tool utilizing machine learning that employs a Support Vector Machine (SVM) classifier for precise PD prediction through biomedical voice data. The system uses the UCI Parkinson's dataset, where pre-processing tasks like feature standardization, train-test split (80-20 ratio), and Recursive Feature Elimination (RFE)enhance model accuracy by identifying significant features. An easy-to-use Streamlit web application was developed to enable real-time predictions, permitting users to input voice parameters and receive instant diagnostic results. The SVM classifier achieved a precision rate of 92%, showcasing its capability and effectiveness in distinguishing PD from non-affected cases. By providing a scalable, cost-effective, and non-invasive approach, this tool bridges advanced computational techniques with real-world healthcare needs. Future enhancements will focus on integrating multimodal data, such as neuroimaging and wearable sensor data, as well as employing deep learning models to improve diagnostic accuracy and expand clinical applicability.

Keywords: Parkinson's Disease, Machine Learning, Voice Biomarkers, SVM Classifier, Diagnostic Tool, Streamlit Application.

DOI: https://doi.org/10.36548/jaicn.2024.4.006
Received: 10.11.2024, received in revised form: 08.12.2024, accepted: 22.12.2024, published: 02.01.2025

1. Introduction

Parkinson's Disease (PD) is a serious neurodegenerative condition that affects millions around the globe, significantly impairing motor abilities through symptoms such as tremors, bradykinesia, and rigidity. Early detection is essential for proper management and treatment, yet it continues to be challenging due to symptom similarities with other diseases and the limitations of traditional diagnostic methods. As a result, there is an urgent demand for innovative and reliable approaches to improve the precision and availability of PD diagnosis.

Recent advancements in machine learning have created possibilities for more precise diagnostic tools capable of processing complex biomedical data. This report focuses on the development of a diagnostic system based on machine learning that utilizes a Support Vector Machine (SVM) classifier to anticipate PD using biomedical voice data. By employing the UCI Parkinson's dataset, the research follows a structured process that includes data preprocessing, feature selection through Recursive Feature Elimination (RFE), and model training with an 80-20 train-test split. The system reaches an impressive 92% accuracy, highlighting its promise as a trustworthy diagnostic resource.

Additionally, the report emphasizes the creation of a user-friendly Streamlit web application designed for real-time predictions, allowing users to input voice parameters and receive prompt diagnostic feedback. This scalable, cost-effective, and non-invasive strategy bridges cutting-edge computational techniques with practical healthcare needs. Future directions involve integrating multi-modal data sources and exploring deep learning methodologies to further enhance diagnostic accuracy and expand clinical use.

2. Related Work

The domain of Parkinson's disease (PD) prediction and monitoring has experienced considerable progress with the incorporation of machine learning and feature selection methodologies. Mohammed Al-Sarem et al. [1] illustrated the efficacy of feature selection techniques such as Recursive Feature Elimination (RFE) in conjunction with classifiers like Support Vector Machines (SVM) to improve predictive precision. In a similar vein, Das [2] evaluated multiple classification algorithms, including Neural Networks and Decision Trees, to determine the model that performs best for PD diagnosis. Tsanas et al. [3] describes the usefulness of non-invasive speech assessments for the telemonitoring of disease progression,

utilizing voice characteristics such as jitter and shimmer to gauge symptom severity. Amit et al. [4] investigated the application of nonlinear dynamics and SVM to detect postural responses, accentuating the significance of biomechanical perspectives in comprehending motor symptoms.

In addition to feature-centric methodologies, molecular and biomechanical research has also made significant contributions to the field. Dexter and Jenner [5] presented a comprehensive analysis of the molecular mechanisms that support PD, providing insights into potential biomarkers for predictive modeling. Ali et al. [6] introduced a hybrid framework that combines L1-regularized SVM with deep neural networks to refine feature sets and enhance accuracy. Rana et al. [7] concentrated on voice characteristics for PD diagnosis, demonstrating the effectiveness of algorithms such as Random Forest and Gradient Boosting in managing voice datasets. Gullapalli and Mittal [8] examined speech features and machine learning approaches, highlighting the promise of deep learning for early identification.

Further developments in transfer learning and ensemble methodologies have also been investigated. Li et al. [9] proposed a two-step sparse transfer learning algorithm for speech diagnosis, illustrating the adaptability of unsupervised learning across various datasets. Ali et al. [10] created a sample- and feature-dependent ensemble approach, emphasizing the reliability of merging multiple models for consistent predictions. Cnockaert et al. [11] evaluated low-frequency vocal modulations, stressing the potential of voice biomarkers for noninvasive diagnosis.

Clinical validation and mobile health initiatives have similarly gained prominence. Rizzo et al. [12] performed a systematic review and meta-analysis to assess the accuracy of clinical diagnoses, establishing a benchmark for the validation of machine learning models. Paragliola and Coronato [13] applied deep time-series methods to identify gait anomalies in PD patients, emphasizing the significance of movement patterns in early detection. Stamate et al. [14] devised a smartphone-based deep learning framework, demonstrating the practicality of mobile health solutions for PD monitoring. Finally, Tabashum et al. [15] reviewed machine learning frameworks for PD, highlighting the advancements and challenges in data preprocessing and model validation techniques. These investigations collectively reflect the increasing promise of machine learning in enhancing Parkinson’s disease prediction, with applications that use feature engineering and ensemble approaches to clinical validation and mobile health technologies.

3. Proposed Work

3.1 Dataset

The dataset used for this study was the UCI Parkinson's Dataset, which consists of 195 records from both Parkinson's Disease patients and healthy individuals. The dataset contains 22 features extracted from biomedical voice measurements, which are valuable indicators of vocal impairments commonly associated with PD. These features include jitter, shimmer, harmonics-to-noise ratio (HNR), and other voice-related attributes like fundamental frequency variation and speech instability. These parameters have been proven to capture subtle changes in speech that are often observed in PD patients, making them suitable for analysis using machine learning techniques. The dataset shown in Table 1 is labelled with binary outcomes (1 for PD-positive and 0 for healthy), which allows for the supervised learning approach used in the study.

Table 1. Sample Dataset

Name	MDVP:Fo(Hz)	MDVP:Fhi(Hz)	MDVP:Flo(Hz)	MDVP:Jitter(%)	MDVP:Jitter(Abs)	MDVP:RAP	MDVP:PPQ	Jitter:DDP	status
phon_R01_S01_1	119.992	157.302	74.997	0.00784	0.00007	0.0037	0.0055	0.01109	1
phon_R01_S01_2	122.4	148.65	113.819	0.00968	0.00008	0.00465	0.007	0.01394	1
phon_R01_S01_3	116.682	131.111	111.555	0.0105	0.00009	0.00544	0.0078	0.01633	1
phon_R01_S01_4	116.676	137.871	111.366	0.00997	0.00009	0.00502	0.007	0.01505	1
phon_R01_S01_5	116.014	141.781	110.655	0.01284	0.00011	0.00655	0.0091	0.01966	1
phon_R01_S01_6	120.552	131.162	113.787	0.00968	0.00008	0.00463	0.0075	0.01388	1
phon_R01_S02_1	120.267	137.244	114.82	0.00333	0.00003	0.00155	0.002	0.00466	1
phon_R01_S02_2	107.332	113.84	104.315	0.0029	0.00003	0.00144	0.0018	0.00431	1

Frequency Measures:

- MDVP:Fo(Hz): Fundamental frequency (Fo), representing the average pitch of the voice in Hertz.
- MDVP:Fhi(Hz): Maximum vocal fundamental frequency.
- MDVP:Flo(Hz): Minimum vocal fundamental frequency.

Jitter Measures (Frequency Variability):

- MDVP:Jitter(%): Relative variation in fundamental frequency.
- MDVP:Jitter(Abs): Absolute variation in fundamental frequency.
- MDVP:RAP: Relative Average Perturbation, measuring short-term variation in pitch.
- MDVP:PPQ: Pitch Period Perturbation Quotient, another measure of pitch variation over longer periods.
- Jitter:DDP: A derivative of jitter measures, typically three times the RAP value.

Shimmer Measures (Amplitude Variability):

- MDVP:Shimmer: Variation in amplitude.
- MDVP:Shimmer(dB): Variation in amplitude expressed in decibels.
- Shimmer:APQ3: Three-point Amplitude Perturbation Quotient, measuring short-term amplitude variability.
- Shimmer:APQ5: Five-point Amplitude Perturbation Quotient.
- MDVP:APQ: Amplitude Perturbation Quotient for the entire sample.
- Shimmer:DDA: Average absolute difference of differences between the amplitude of consecutive periods, typically three times APQ3.

Noise and Harmonics Measures:

- NHR: Noise-to-Harmonics Ratio, indicating the amount of noise relative to tonal components in the voice.
- HNR: Harmonics-to-Noise Ratio, higher values indicate a cleaner voice signal.

Dynamical System Measures:

- RPDE (Recurrence Period Density Entropy): A nonlinear dynamical measure of signal complexity.
- DFA (Detrended Fluctuation Analysis): Measures self-similarity in voice signals over time.

Signal Measures:

- spread1 and spread2: Nonlinear measures representing the deviation of the voice from normal patterns.
- D2: Correlation dimension, measuring the complexity of the signal.
- PPE (Pitch Period Entropy.

3.2 Model Training

- **Algorithm Selection**: The Support Vector Machine (SVM) was chosen for this classification task due to its high effectiveness in binary classification problems, especially with high-dimensional data. The linear kernel was selected because it is simple and works well with the dataset, providing a good balance between performance and computational efficiency. SVM attempts to find the optimal hyperplane that separates the two classes (PD and healthy), maximizing the margin between them.

- **Implementation of SVM:** The Support Vector Machine (SVM) was selected as the primary algorithm for Parkinson's disease prediction due to its robust performance in binary classification tasks and its ability to handle high-dimensional datasets

effectively. SVM excels in finding the optimal hyperplane that maximizes the margin between different classes, making it well-suited for distinguishing Parkinson's patients from healthy individuals. This is particularly important in Parkinson's disease prediction, where subtle variations in biomedical features, such as jitter, shimmer, and harmonics-to-noise ratio (HNR), play a crucial role in classification. Additionally, SVM's flexibility in using kernel functions allows it to model non-linear relationships in the data, capturing complex patterns associated with the disease. Its resistance to overfitting, especially in small or imbalanced datasets, ensures reliable and generalized predictions. By utilizing these strengths, SVM provides an accurate method for identifying Parkinson's disease based on voice and other clinical measurements. The Figure 1 depicts the training and the testing curves of SVM.

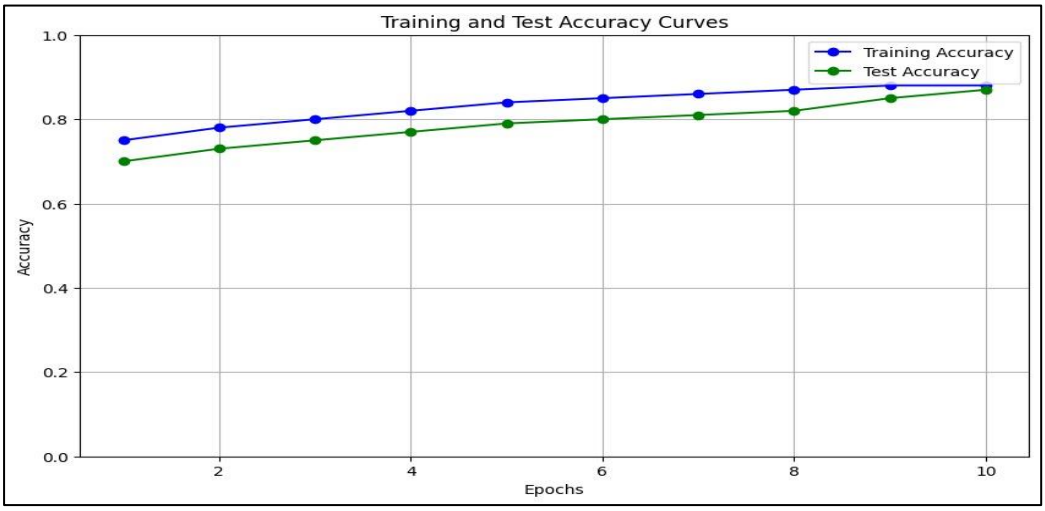

Figure 1. SVM Training and Test Accuracy Curves

3.3 Application Interface

To make the model accessible and user-friendly, an interactive web-based application was developed using Streamlit, a Python library designed for creating simple, fast, and interactive data applications. The following features were incorporated into the interface:

- **Data Input**: The application allows users to input the required biomedical measurements through clearly labelled fields. These fields correspond to voice parameters like jitter, shimmer, and HNR, which can be easily entered without requiring any technical expertise.

- **Prediction Display**: Once the data is entered, the model processes the inputs in real time and provides an output indicating the likelihood of Parkinson's Disease. The result is displayed immediately, allowing for quick assessments. The prediction is accompanied by a visual indicator, which simplifies the interpretation for users.

- **Export Options**: For further clinical use, the application offers the functionality to export the prediction results into a downloadable report. This report can be saved in PDF format and shared with patients, doctors, or other healthcare professionals for review and follow-up treatment planning.

3.4 Software Used

Python is a versatile programming language that acts as the foundation of the Parkinson's Disease prediction system. Its broad library ecosystem, which includes NumPy for numerical calculations, Pandas for data handling, and Scikit-learn for machine learning, simplifies the whole process from data preprocessing to model creation. Python's easy-to-understand syntax and strong libraries allow for effective management of large datasets, feature extraction, and the training of predictive models with simplicity. Python's efficiency and adaptability render it an essential tool for creating advanced healthcare applications.

3.4.1 Libraries Used

- **NumPy**: A library for numerical computing in Python, used for handling arrays and performing mathematical operations. it converts the input data into NumPy arrays and does efficient mathematical computations.

- **Pandas**: A versatile library for data manipulation and analysis, providing data structures like dataFrames. It is used for oading, processing datasets, and manipulating data.

- **Scikit-learn (sklearn)**: A widely used machine learning library for building models, data preprocessing, and evaluating performance.

- **Streamlit**: An open-source framework for creating interactive web applications using Python . Used in Building a user interface for Parkinson's Disease prediction, accepting user inputs, and displaying prediction results.

3.4.2 Modules Used

- **train_test_split**: Splits the dataset into training and testing sets.
- **StandardScaler**: Standardizes features by removing the mean and scaling to unit variance.
- **svm (Support Vector Machine)**: The algorithm used for model training and prediction.
- **accuracy_score**: Evaluates the accuracy of predictions.

3.5 Algorithm

Step 1: Initialize the process.

Step 2: Collect patient data, such as voice samples (frequency, amplitude, jitter, shimmer), motor symptoms (tremors, rigidity, bradykinesia), non-motor symptoms (sleep disturbances, mood changes), and demographic information (age, gender).

Step 3: Perform the following preprocessing steps: handle missing or erroneous values (data cleaning), scale features to a uniform range (normalization), and increase the dataset size using synthetic data if needed (data augmentation).

Step 4: Extract relevant features from the dataset, such as voice frequency metrics, Unified Parkinson's Disease Rating Scale (UPDRS) scores, and gait and posture abnormalities.

Step 5: Select the most important features using Recursive Feature Elimination (RFE).

Step 6: Choose a suitable machine learning model, Support Vector Machine (SVM).

Step 7: Split the dataset into training and testing sets (e.g., 80-20 split). Train the selected model using the training data and optimize hyperparameters to improve model performance.

Step 8: Evaluate the model using the testing dataset. Calculate metrics such as accuracy, precision, recall, and F1-score.

Step 9: Run new patient data through the trained model and predict whether the patient has Parkinson's disease or not.

Step 10: Cross-check the model prediction with clinical evaluations or expert diagnosis.

Step 11: Provide the result to the user: Positive (Parkinson's disease detected) or Negative (Parkinson's disease not detected).

Step 12: Terminate the process

3.6 Flowchart

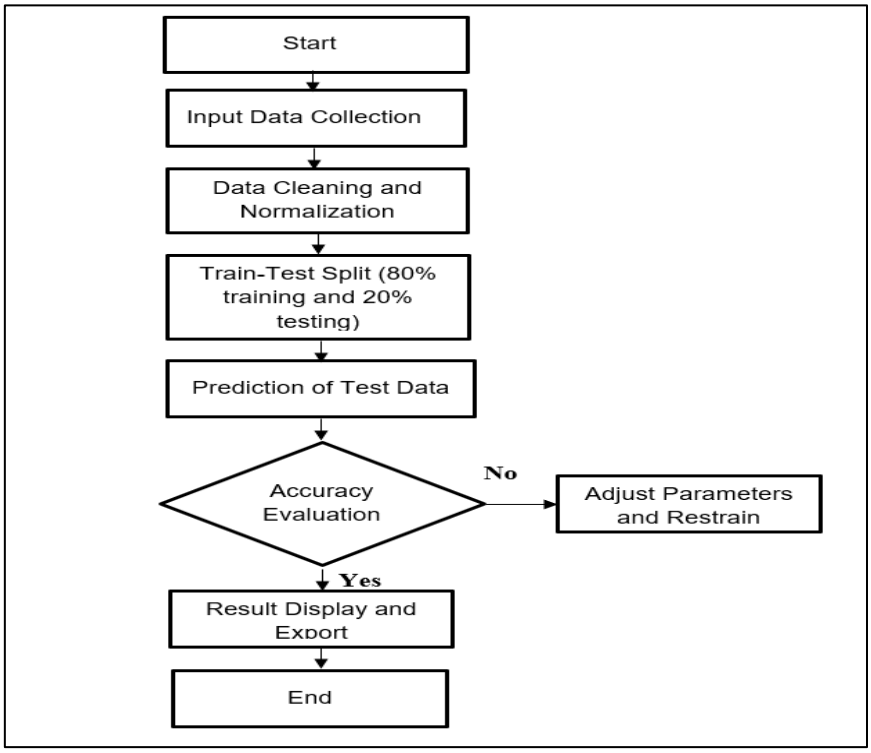

Figure 2. Flowchart

4. Results and Discussion

4.1 Model Performance

The SVM classifier achieved an accuracy of 92% on the test data, demonstrating its robustness and reliability for PD prediction.

4.2 Application Workflow

The system workflow, detailed in Figure 2, illustrates the end-to-end process:

1. **Prediction**: The SVM model processes the data to classify the likelihood of PD.
2. **Result Display**: Real-time outcomes and diagnostic insights are shown in Figure 3and 4

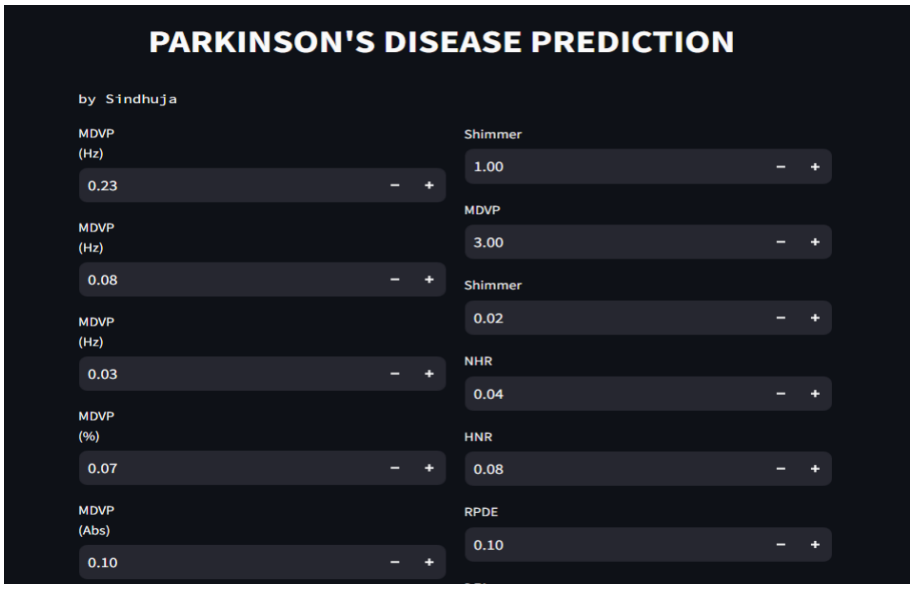

Figure 3. Input of Sample Data

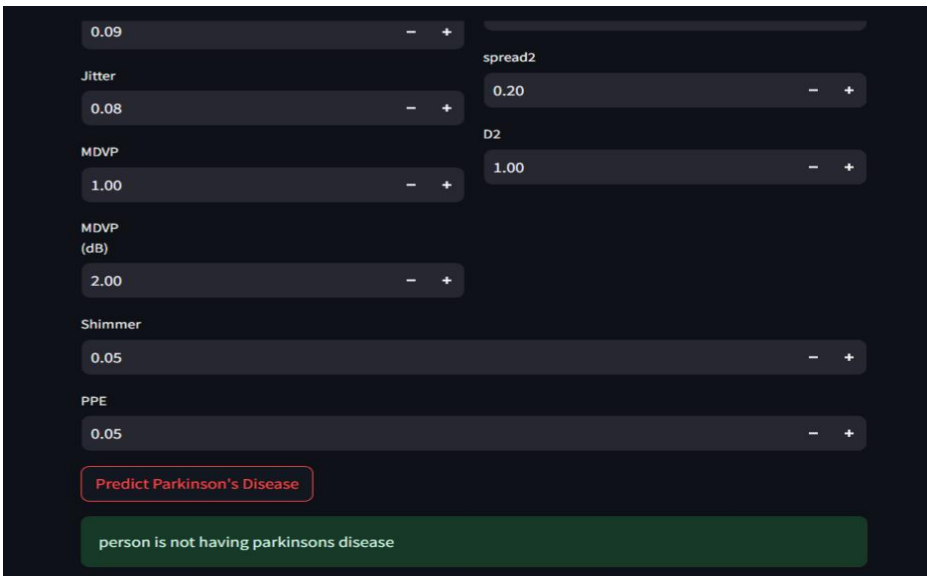

Figure 4. Prediction Display

The output of the Parkinson's Disease prediction tested for the given sample inputs is illustrated in Figure.3 and Figure.4.

4.3 Future Enhancements

Future improvements for predicting Parkinson's disease can greatly benefit from sophisticated machine learning strategies and the integration of multi-modal data. By utilizing deep learning methods, scientists can explore intricate patterns within information from various origins, including neuroimaging, genetic tests, and wearable technologies. Moreover, combining these different data types—like clinical evaluations, lifestyle factors, and physiological data—can offer a more thorough understanding of a person's risk profile, increasing the precision of forecasting models. Another vital aspect for improvement includes biomarker identification and longitudinal studies. Researchers can concentrate on discovering new biomarkers through genomic and proteomic research, possibly creating non-invasive blood tests to identify early indications of neurodegeneration. Additionally, performing extended studies with ongoing patient monitoring can generate significant insights into disease advancement, supporting the refinement of predictive models and enabling adaptable risk assessments that evolve as additional data is gathered over time. Finally, encouraging community and family participation, along with collaboration with healthcare professionals, can enhance the efficacy of predictive instruments. Informing families and communities about the initial signs of Parkinson's disease promotes prompt diagnosis and intervention. Furthermore, incorporating predictive tools into clinical practices and offering training for healthcare staff will improve their capability to effectively apply these models in patient care. This cooperative strategy focuses on establishing a nurturing atmosphere for at-risk individuals, advancing awareness and proactive management of Parkinson's disease.

5. Conclusion

In conclusion, enhancing the prediction of Parkinson's disease is vital for improving early diagnosis and treatment results. By utilizing cutting-edge machine learning methods and integrating various data sources including neuroimaging, genetic data, and wearable sensor information, researchers can create more precise predictive models. Furthermore, identifying new biomarkers and conducting longitudinal research will deepen the understanding of the disease's advancement and enable timely interventions. Promoting community awareness and collaborating with healthcare providers is essential for effectively applying these predictive tools, ensuring that at-risk individuals receive prompt support and management. Ultimately, a comprehensive approach that merges technological advancements with community and clinical

involvement will greatly enhance the quality of life for those impacted by Parkinson's Disease, resulting in better health outcomes and more customized care strategies.

References

[1] Mohammed Al-Sarem, F., Saeed, F., Al-Mohaimeed, M., & Ghabban, D. (2022). Enhancing Parkinson's Disease Prediction Using Machine Learning and Feature Selection Methods. Computers, Materials & Continua, 71(3), 5639–5658.

[2] R. Das, (2010) “A Comparison of multiple classification methods for diagnosis of Parkinson disease.” Expert Systems with Applications, vol. 37, no. 2, pp. 1568-1572.

[3] Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, (2010) “Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests.” Biomedical Engineering, IEEE Transactions, vol. 57, no. 4, pp. 884-893.

[4] S. Amit, M. Ashutosh, A. Bhattacharya and F. Revilla, (2014) "Understanding Postural Response of Parkinson's Subjects Using Nonlinear Dynamics and Support Vector Machines", Austin J. Biomed Eng., vol. 1, no. 1, pp. id1005.

[5] D. T. Dexter and P. Jenner, (2013)"Parkinson disease: from pathology to molecular disease mechanisms", Free Radical Biology and Medicine, vol. 62, pp. 132-144.

[6] Liaqat Ali, Ashir Javeed, Adeeb Noor, Hafiz Tayyab Rauf, Seifedine Kadry, Amir H. Gandomi,(2024) "Parkinson’s disease detection based on features refinement through L1 regularized SVM and deep neural network", Scientific Reports, vol.14, no.1, 2024.

[7] Arti Rana, Ankur Dumka, Rajesh Singh, Mamoon Rashid, Nazir Ahmad, Manoj Kumar Panda,(2022) "An Efficient Machine Learning Approach for Diagnosing Parkinson’s Disease by Utilizing Voice Features", Electronics, vol.11, no.22, 3782.

[8] Ajay Sankar Gullapalli, Vinay Kumar Mittal, (2022)"Early Detection of Parkinson’s Disease Through Speech Features and Machine Learning: A Review", ICT with Intelligent Applications, vol.248, 203.

[9] Yongming Li, Xinyue Zhang, Pin Wang, Xiaoheng Zhang, Yuchuan Liu, (2021) "Insight into an unsupervised two-step sparse transfer learning algorithm for speech

diagnosis of Parkinson’s disease", Neural Computing and Applications, vol.33, no.15, 9733.

[10] Liaqat Ali, Chinmay Chakraborty, Zhiquan He, Wenming Cao, Yakubu Imrana, Joel J. P. C. Rodrigues,(2022) "A novel sample and feature dependent ensemble approach for Parkinson?s disease detection", Neural Computing and Applications, 2022.

[11] Cnockaert, Laurence, Jean Schoentgen, Pascal Auzou, Canan Ozsancak, L. Defebvre, and Francis Grenez. "Low-frequency vocal modulations in vowels produced by Parkinsonian subjects." Speech communication 50, no. 4 (2008): 288-300.

[12] Giovanni Rizzo, Massimiliano Copetti, Simona Arcuti, Davide Martino, Andrea Fontana and Giancarlo Logroscino,(2016) "Accuracy of clinical diagnosis of parkinson disease: a systematic review and meta-analysis", Neurology, vol. 86, no. 6, pp. 566-576.

[13] Giovanni Paragliola and Antonio Coronato, (2018) "Gait anomaly detection of subjects with parkinson’s disease using a deep time series-based approach", IEEE Access, vol. 6, 73280-73292.

[14] Cosmin Stamate, George D Magoulas, Stefan Küppers, Effrosyni Nomikou, Ioannis Daskalopoulos, Marco U Luchini, et al., (2017) "Deep learning parkinson’s from smartphone data", 2017 IEEE International Conference on Pervasive Computing and Communications.,Kona, HI, USA 31-40.

[15] Tabashum, T., Snyder, R. C., O'Brien, M. K., & Albert, M. V. (2024). Machine Learning Models for Parkinson Disease: Systematic Review. Journal of Medical Internet Research, 12.