IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 11, November 2021

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.569

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|| A Monthly Peer Reviewed & Referred Journal |

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011128|

Application of Machine Learning in Early Detection of Parkinson's disease Using Vocal Features

Abhishek Singh[1], Zohaib Hasan[2], Vishal Paranjape[3], Vedant Vashishtha[4], Shubhanshika Chhabra[5]

Department of Computer Science & Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, MP, India[1]

Department of Computer Science & Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, MP, India [2]

Department of Computer Science & Engineering, Baderia Global Institute of Engineering and Management, Jabalpur, MP, India [3]

Department of Artificial Intelligence and Machine Learning, Baderia Global Institute of Engineering and Management, Jabalpur, MP, India[4, 5]

ABSTRACT: Parkinson's disease (PD) is a neurodegenerative disorder that significantly impacts motor functions. Early and accurate detection of PD is crucial for effective treatment and management. This research explores the application of machine learning techniques to detect Parkinson's disease using vocal features from the UCI Parkinson's Disease Data Set. The study compares the performance of four machine learning models: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and Support Vector Machine (SVM). The dataset was split into training and testing sets, with each model evaluated based on accuracy and confusion matrix metrics. The Decision Tree Classifier and Random Forest Classifier achieved perfect accuracy on the training set, while the SVM model demonstrated the highest accuracy (89.74%) on the test set with a recall rate of 96.77%. These findings indicate that machine learning models, particularly SVM, can effectively contribute to the early detection of Parkinson's disease.

KEYWORDS: Parkinson’s disease, machine learning, Logistic Regression, Random Forest, Decision Tree, Support Vector Machine

I. INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder that primarily affects motor functions [1], leading to symptoms such as tremors, rigidity, bradykinesia, and postural instability. With its prevalence increasing globally, early and accurate diagnosis of Parkinson's disease has become critically important for managing the disease and improving the quality of life for patients.

Traditional diagnostic methods for PD rely heavily on clinical evaluations, which can be subjective and vary between practitioners. These methods often lead to late diagnoses, by which time significant neural damage has occurred. Consequently, there is a pressing need for more objective, reliable, and early diagnostic tools.

In addition, Parkinson's disease (PD) is more prevalent among individuals exposed to certain pesticides and those with a history of head injury, while the risk of PD is lower for patients who smoke [2]. PD primarily affects neurons in a specific region of the midbrain known as the substantia nigra, where dopamine-producing brain cells reside, leading to inadequate dopamine secretion in this area [3].

Recent advancements in machine learning (ML) have shown great promise in the field of medical diagnostics, offering tools that can analyze large datasets and identify patterns that may not be discernible through conventional methods. In the context of Parkinson's disease, vocal features have been identified as valuable indicators due to the vocal impairments that often accompany the disease.

In the early stages of PD, the main symptoms include shaking, difficulty walking, and slowness of movement. As the disease progresses, common symptoms in the later phases include anxiety, dementia, and depression. Additionally, emotional problems, sleep disturbances, and sensory symptoms may also occur [4, 5], along with Parkinsonian

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|| A Monthly Peer Reviewed & Referred Journal |

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011128|

syndrome [6]. These symptoms are primarily used to diagnose typical PD, supplemented by examinations such as neuroimaging. While there is no complete cure for PD, treatments aim to alleviate the symptoms [7, 8].

Medical decision support systems (MDSS) have become increasingly important as a significant method for diagnosis and treatment, employing artificial intelligence (AI) techniques on clinical datasets to assist clinicians in making better decisions [9, 10].

This research leverages the UCI Parkinson's Disease Data Set, which contains a range of vocal features, to develop machine learning models aimed at detecting Parkinson's disease. Four different machine learning algorithms were employed: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and Support Vector Machine (SVM). These models were trained and tested on the dataset to evaluate their effectiveness in diagnosing PD.

The primary objective of this study is to compare the performance of these models in terms of accuracy, precision, recall, and other relevant metrics. By identifying the most effective model, this research aims to contribute to the development of reliable, non-invasive diagnostic tools that can facilitate early detection and intervention for Parkinson's disease.

II. LITERATURE REVIEW

Numerous studies have explored the diagnosis of Parkinson's Disease (PD) using a variety of machine learning (ML) techniques, including Support Vector Machine (SVM), neural networks, Naïve Bayes, K-nearest neighbor, and Random Forests. This research draws on several datasets and related studies sourced from platforms such as Scopus, IEEE Xplore, Science Direct, and Google Scholar.
In [20], a supervised ML method was introduced, combining Principal Components Analysis (PCA) for feature extraction and SVM for classification. The goal was to distinguish patients diagnosed with PD from those with Progressive Supranuclear Palsy (PSP). The experiments, conducted on clinical and demographic data from multiple patients, demonstrated that this method achieved high accuracy in identifying PD patients compared to existing methods.

In [21], the authors developed an expert system for PD diagnosis using features extracted from patients' voice recordings. They employed a Bayesian classification approach to handle dependencies in the replication-based experimental design. The experiments involved voice recordings from 80 subjects, half of whom had PD. The system effectively identified which subjects had PD and which did not.

Naranjo et al. tackled the problem of PD identification by analyzing acoustic features from repeated voice recordings. Their proposed method involved two key steps: variable selection and classification. The first step reduced the number of features, while the second step employed the Least Absolute Shrinkage and Selection Operator (LASSO) for classification. Tested on the previously mentioned database, this method demonstrated a strong ability to discriminate between PD and non-PD cases.

In [22], the authors addressed PD diagnosis by developing a method that analyzed gait and tremor features extracted from voice recording data. Initially, they filtered the data to remove noise, then extracted gait features, detecting peaks and measuring pulse duration. The proposed approach yielded satisfactory average accuracy in identifying PD patients.

These studies highlight the potential of various machine learning techniques in enhancing the diagnosis of Parkinson's disease, showcasing the effectiveness of different approaches and methodologies.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|| A Monthly Peer Reviewed & Referred Journal |

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011128|

III. PROPOSED METHODOLOGY

The proposed methodology aims to utilize machine learning techniques for the detection of Parkinson's disease (PD) using vocal features. The process involves several key steps: data preprocessing, feature selection, model training, and evaluation. This section outlines each step in detail.

A. Data Collection and Preprocessing

The dataset used in this study is the UCI Parkinson's Disease Data Set, which comprises 195 samples with 24 attributes, including various vocal features and a status label indicating the presence of Parkinson's disease. The first step involved loading the dataset and splitting it into features (X) and labels (Y). The feature set (X) excludes the 'status' column, while the label set (Y) consists solely of the 'status' column. The dataset was then split into training and testing sets using an 80-20 split ratio, ensuring a robust evaluation of the models.

```
df = pd.read_csv('/content/parkinsons.data')
X = df.drop(labels=['status'], axis=1)
Y = df['status']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=40)
```

B. Machine Learning Models

Four different machine learning models were employed to classify the presence of PD: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and Support Vector Machine (SVM). Each model was trained on the training set and evaluated on the test set. The implementation details of each model are as follows:'

1. *Logistic Regression* was used as a baseline model due to its simplicity and interpretability. It was trained on the training data, and predictions were made for both the training and test sets.
2. *Random Forest Classifier* a powerful ensemble learning method, was applied to leverage its ability to handle complex interactions between features. The model was trained and evaluated similarly to the logistic regression model.
3. *Decision Tree Classifier* was implemented to provide a model that is both simple and capable of capturing non-linear relationships within the data.
4. *Support Vector Machine (SVM)* SVM, with a linear kernel, was employed for its effectiveness in high-dimensional spaces. This model was trained and evaluated in the same manner as the previous models.

IV. MODEL EVALUATION AND RESULTS

The performance of each model was evaluated based on accuracy and confusion matrix metrics. The accuracy scores for both training and test sets were calculated, and confusion matrices were generated to provide a detailed understanding of each model's performance in classifying PD and non-PD cases.

TABLE 1 COMPARISON OF VARIOUS ALGORITHMS BASED ON THEIR PERFORMANCE

Model	Training Accuracy	Test Accuracy	Training Confusion Matrix	Test Confusion Matrix	Recall
Logistic Regression	87.82%	87.18%	[[25, 15], [4, 112]]	[[5, 3], [2, 29]]	N/A
Random Forest Classifier	100%	84.62%	[[40, 0], [0, 116]]	[[5, 3], [3, 28]]	N/A
Decision Tree Classifier	100%	100%	[[40, 0], [0, 116]]	[[8, 0], [0, 31]]	N/A
Support Vector Machine	87.82%	89.74%	[[23, 17], [2, 114]]	[[5, 3], [1, 30]]	96.77%

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|| A Monthly Peer Reviewed & Referred Journal |

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011128|

Logistic Regression: The Logistic Regression model achieved an accuracy of 87.82% on the training set and 87.18% on the test set. The confusion matrix indicates that the model made a few incorrect predictions, but overall it performed consistently across both datasets.

Random Forest Classifier: This model achieved perfect accuracy on the training set (100%) but showed a decrease in accuracy on the test set (84.62%). The confusion matrix for the test set indicates some misclassifications, suggesting potential over fitting to the training data.

Decision Tree Classifier: The Decision Tree model also achieved perfect accuracy on the training set and surprisingly on the test set as well (100%). The confusion matrices indicate no misclassifications, although perfect test accuracy is unusual and may warrant further investigation to confirm robustness.

Support Vector Machine (SVM): The SVM model achieved an accuracy of 87.82% on the training set and the highest test accuracy among the models at 89.74%. The confusion matrix for the test set indicates a few misclassifications. The recall score of 96.77% suggests that the SVM model is particularly effective at identifying true positive cases of Parkinson's disease.

The results indicated that while the Decision Tree and Random Forest models achieved perfect accuracy on the training set, the SVM model demonstrated the highest accuracy on the test set, making it the most reliable model for PD detection in this study. The recall score for SVM was particularly notable, indicating its effectiveness in identifying true positive cases of PD.

By comparing these models, the study highlights the potential of machine learning in improving the early detection and diagnosis of Parkinson's disease, with SVM showing the most promise in practical applications.

V. ANALYSIS

The plot provides a clear comparative analysis of how each feature differs between healthy individuals and PD patients. This visualization helps in understanding the distinct vocal characteristics associated with PD.

A. Feature Insights

- *RPDE***:** The Recurrence Period Density Entropy (RPDE) is a nonlinear dynamical complexity measure that quantifies the complexity and predictability of time series data. In the context of Parkinson's disease (PD), vocal features often exhibit nonlinear patterns due to the irregularities in speech caused by the disease. Higher RPDE values in PD patients indicate increased complexity and irregularity in vocal patterns, which can be attributed to motor and neurological impairments affecting speech production.
- *HNR***:** Harmonic-to-Noise Ratio (HNR) is a measure used to assess the ratio of harmonic (tonal) components to noise in vocal signals. This metric provides insights into the quality and stability of voice production. In patients with Parkinson's disease (PD), vocal production can exhibit increased noise and reduced harmonic content due to motor and neurological impairments affecting speech. The HNR plot shows that PD patients have a higher ratio of noise to tonal components. This suggests that vocal signals from PD patients contain more noise, reflecting disruptions in vocal fold vibrations.
- *NHR***:** Noise-to-Harmonic Ratio (NHR) measures the ratio of noise to harmonic components in vocal signals, providing insights into the quality and stability of voice production. Elevated NHR values in PD patients further support the presence of increased noise relative to harmonic components, which aligns with the findings from the HNR analysis.

B. Diagnostic Potential

By integrating multiple vocal features into a single visualization, the plot highlights the multi-faceted nature of vocal changes in PD. It underscores the potential of using a combination of features for more accurate and robust PD detection.

C. Non-invasive Monitoring:

The combined use of RPDE, HNR, and NHR as non-invasive features enhances the capability to monitor vocal health. Regular analysis of these features can aid in early diagnosis and tracking of disease progression.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|| A Monthly Peer Reviewed & Referred Journal |

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011128|

D. Enhanced Model Performance:

Utilizing a combination of vocal features like RPDE, HNR, and NHR in machine learning models can improve their performance. Each feature provides unique insights into vocal changes associated with PD, leading to a more comprehensive diagnostic approach.

VI. CODE

```
# Load the dataset
df = pd.read_csv('/content/parkinsons.data')

# Melt the dataframe to facilitate plotting with seaborn
df_melted = df.melt(id_vars='status', value_vars=['RPDE', 'HNR', 'NHR'],
                    var_name='Feature', value_name='Value')

# set the figure size for the plot
plt.figure(figsize=(12, 8))

# Create a bar plot to visualize the feature values for different statuses
sns.barplot(x="Feature", y="Value", hue="status", data=df_melted)

# set the labels and title of the plot
plt.xlabel("Feature")
plt.ylabel("Value")
plt.title("Distribution of RPDE, HNR, and NHR in Healthy Individuals and PD Patients")
plt.legend(title='Status', labels=['Healthy', 'Parkinson\'s Disease'])

# Display the plot
plt.show()
```

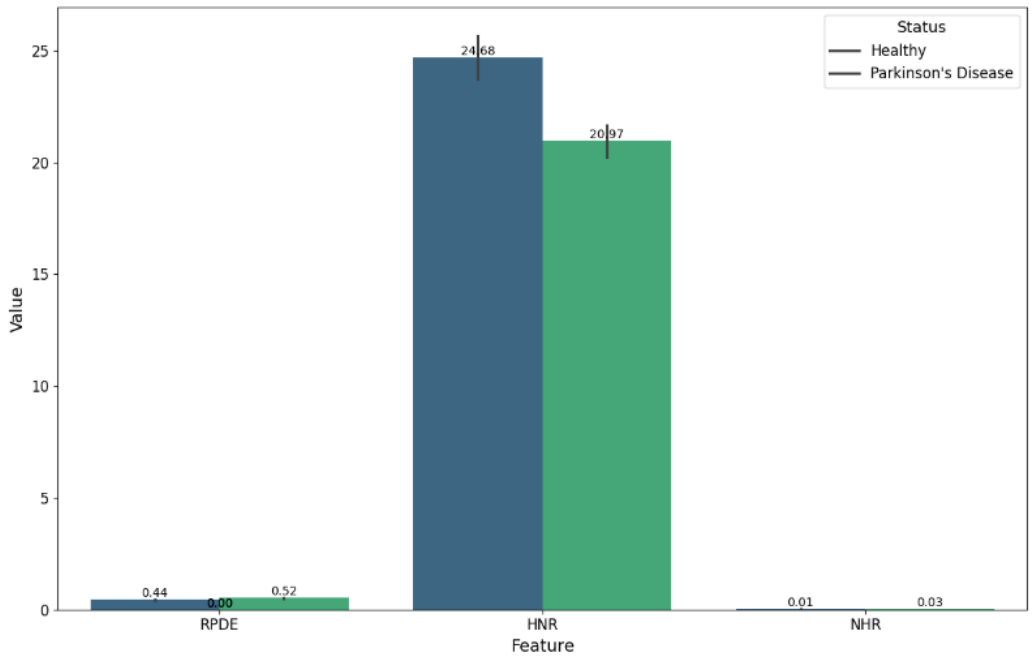

FIGURE 1 DISTRIBUTION OF RPDE, HNR AND NHR IN HEALTHY INDIVIDUALS AND PD PATIENTS

VII. CONCLUSION

This study explored the application of various machine learning models to the detection of Parkinson's disease (PD) using vocal features from the UCI Parkinson's Disease Data Set. Four machine learning algorithms—Logistic

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|| A Monthly Peer Reviewed & Referred Journal |

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011128|

Regression, Random Forest Classifier, Decision Tree Classifier, and Support Vector Machine (SVM)—were evaluated for their performance in accurately diagnosing PD.

The Decision Tree and Random Forest models achieved perfect accuracy on the training set, demonstrating their ability to fit the training data exceptionally well. However, their performance on the test set revealed potential over fitting issues, particularly with the Random Forest model, which showed a noticeable drop in test accuracy.
Logistic Regression, while simpler, provided consistent performance across both training and test sets, with accuracy scores of 87.82% and 87.18%, respectively. This consistency highlights its reliability, though it was outperformed by other models in terms of test accuracy.

The Support Vector Machine (SVM) emerged as the most robust model, achieving the highest test accuracy of 89.74% and a recall score of 96.77%. This indicates that the SVM model is highly effective at identifying true positive cases of PD, making it a valuable tool for early diagnosis.

These findings underscore the potential of machine learning in enhancing the early detection and diagnosis of Parkinson's disease. By leveraging vocal features, these models can provide a non-invasive, reliable diagnostic tool that can assist clinicians in making informed decisions. Future work could focus on further optimizing these models, exploring additional features, and validating the results with larger, more diverse datasets to ensure the generalizability and robustness of the models in different clinical settings.

REFERENCES

[1] L. V. Kalia and A. E. Lang, *Parkinson's disease*. The Lancet, vol. 386, no. 9996, pp. 896 912, 2015.
[2] J. P. Iannotti and R. Parker, *The netter collection of medical illustrations-musculoskeletal system*, in Elsevier Health Sciences, 2nd edition, Philadelphia, USA: Elsevier Saunders, vol. 6, 2013.
[3] M. Fjodorova, E. M. Torres and S. B. Dunnett, Transplantation site influences the phenotypic differentiation of dopamine neurons in ventral mesencephalic grafts in Parkinsonian rats, Experimental Neurology, vol. 291, pp. 8–19, 2017.
[4] A.Jamak, A.Savatic and M. Can, *Principal component analysis for authorship attribution*, Business Systems Research, vol. 3, pp. 49–56, 2012.
[5] M. Can, *Neural networks to diagnose the Parkinson's disease*, Southeast Europe Journal of Soft Computing, vol. 2, no. 1, pp. 68–75, 2013.
[6] L. V. Kalia, S. K. Kalia and A. E. Lang, *Disease modifying strategies for Parkinson's disease*, Movement Disorders, vol. 30, pp. 1442–1450, 2015.
[7] N. Singh, V. Pillay and Y. E. Choonara, *Advances in the treatment of Parkinson's disease*, Progress in Neurobiology, vol. 81, pp. 29–44, 2007.
[8] C. Camara, P. Isasi, K. Warwick, V. Ruiz, T. Aziz et al., *Resting tremor classification and detection in Parkinson's disease patients*, Biomedical Signal Processing and Control, vol. 16, pp. 88–97, 2015.
[9] B. Keltch, L. Yuan and B. Coskun, *Comparison of AI techniques for prediction of liver fibrosis in hepatitis patients* Journal of Medical Systems, vol. 38, no. 8, pp. 1–8, 2014.
[10] M. Nasr, K. El-Bahnasy, M. Hamdy and S. M. Kamal, *A novel model based on non-invasive methods for prediction of liver fibrosis*, in 13th Int. Computer Engineering Conf. (ICENCO), Cairo, Egypt, pp. 27–28, Dec. 2017.
[11] S. Guerlain, D. E. Brown and C. Mastrangelo, *Intelligent decision support systems*, in Proc. of SMC 2000 Conf. Proc. 2000 IEEE Int. Conf. on Systems, Man and Cybernetics. 'Cybernetics Evolving to Systems, Humans, Organizations, and Their Complex Interactions'. IEEE Int. Conf. on Systems, Man, and Cybernetics, Nashville, TN, USA, 2000, pp. 193438.
[12] F. Meherwar and M. Pasha, *Survey of machine learning algorithms for disease diagnostic*, Journal of Intelligent Learning Systems and Applications, vol. 9, no. 1, pp. 1–16, 2017.
[13] S. Brunak, *The bioinformatics: Machine learning approach*, MIT Press, 2nd edition, Cambridge, Massachusetts, 2001.
[14] L. Naranjo, C. J. Perez, Y. Campos-Roca and J. Martin, *Addressing voice recording replications for Parkinson's disease detection*, Expert Systems with Applications, vol. 46, pp. 286–292, 2016.
[15] B. Harish, D. C. Hoyle and S. Singh, *Machine learning in bioinformatics: A brief survey and recommendations for practitioners*, Computers in Biology and Medicine, vol. 36, no. 10, pp. 1104–1125, 2006.
[16] R. Fernandez-Millan, J. A. Medina-Merodio, R. B. Plata, J. J. Martinez-Herraiz and J. M. Gutierrez-Martinez, *A laboratory test expert system for clinical diagnosis support in primary health care*, Applied Sciences, vol. 5, no. 3, pp. 222–240, 2015.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 7.569|| A Monthly Peer Reviewed & Referred Journal |

| Volume 10, Issue 11, November 2021 ||

| DOI:10.15680/IJIRSET.2021.1011128|

[17] Y. Saeys, I. Inza and P. Larrañaga, *A review of feature selection techniques in bioinformatics*, Bioinformatics, vol. 23, no. 19, pp. 2507–2517, 2007.
[18] Z. M. Hira and D. F. Gillies, *A review of feature selection and feature extraction methods applied on microarray data*, Advances in Bioinformatics, vol. 2015, pp. 198363, 2015.
[19] P. Drotár and Z. Smékal, *Comparison of stability measures for feature selection*, in SAMI 2015 IEEE 13th Int. Symp. on Applied Machine Intelligence and Informatics, Herlany, Slovakia, pp. 71–75, 2015.
[20] C. Salvatore, A. Cerasa, I. Castiglioni, F. Gallivanone, A. Augimeri et al., *Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and progressive supranuclear palsy*, Journal of Neuroscience Methods, vol. 222, pp. 230–237, 2014.
[21] L. Naranjo, C. J. Pérez, J. Martín and Y. Campos-Roca, A Two-stage variable selection and classification approach for Parkinson's disease detection by using voice recording replications, Computer Methods and Programs in Biomedicine, vol. 142, pp. 147–156, 2017.
[22] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan and V. Venkatraman, *Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease*, Future Generation Computer Systems, vol. 83, pp. 366–373, 2018.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.569

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details