A

MACHINE LEARNING PROJECT REPORT

On

**Parkinson’s Disease Detection**

Submittedby:

|  |  |
| --- | --- |
| Anirudh Dattu | 220328 |
| Akaash Gupta | 220332 |
| Satya Karthikeya Singidi | 220353 |
| Aryan Rongali | 220340 |

under mentorship of

**Dr. Hirdesh Kumar Pharasi**

(Professor)

****

Department of Computer Science Engineering

School of Engineering and Technology

BML MUNJAL UNIVERSITY, GURUGRAM (INDIA)

May 2024

**CANDIDATES DECLARATION**

We hereby declare that the work on **‘Parkinson’s Disease Detection’**, in partial fulfilment of requirements for the award of Degree of Bachelor of Technology in School of Engineering and Technology at BML Munjal University, having University Roll No.1232434, is an authentic record of my own work carried out during a period from July 2022 to January 2022 under the supervision of

SUPERVISOR NAME: Dr. Hirdesh Kumar Pharasi

Signature:

1. Anirudh Dattu –

2. Akaash Gupta –

3. Karthik Singidi –

4. Aryan Rongali –

**SUPERVISOR’S DECLARATION**

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Faculty Supervisor Name: **Dr. Hirdesh Kumar Pharasi**

Signature:

**ACKNOWLEDGEMENT**

We would like to express our gratitude to our course faculty **Dr. Hirdesh Kumar Pharasi** for his valuable guidance and support throughout the course.

**Dr. Hirdesh Kumar Pharasi** has provided great help in carrying out my work and is acknowledged with reverential thanks. Without wise counsel and able guidance, it would have been impossible to complete the training in this manner.

I would like to express thanks profusely to thank **Dr. Hirdesh Kumar Pharasi**, for stimulating me from time to time. I would also like to thank the entire team at BML Munjal University. I would also thank my friends who devoted their valuable time and helped me in all possible ways toward successful completion.

1. Anirudh Dattu

2. Akaash Gupta

3. Karthik Singidi

4. Aryan Rongali

**ABSTRACT**

Parkinson's disease (PD) is a progressive neurodegenerative disorder that affects millions of people worldwide. Early detection and accurate diagnosis of PD are crucial for timely intervention and effective management of the disease. In this project, we propose a comprehensive approach for Parkinson's disease detection using machine learning algorithms applied to two distinct types of data: voice recordings and spiral drawings.

The project consists of two main parts. The first part focuses on analysing voice recordings collected from individuals with and without PD. Various machine learning algorithms, including Support Vector Machines (SVM), Logistic Regression (LR), Random Forests (RF), and others, are trained and evaluated using features extracted from voice recordings to discriminate between PD and healthy individuals. Additionally, techniques such as feature engineering, normalization, and cross-validation are employed to enhance the performance of the classifiers.

In the second part of the project, we explore the use of spiral drawings as a supplementary modality for PD detection. Image processing techniques, including edge detection and Histogram of Oriented Gradients (HOG) feature extraction, are utilized to preprocess spiral drawings and extract discriminative features. Machine learning models such as Random Forest Classifier (RFC), Naive Bayes (NB), and Support Vector Machines (SVM) are trained on the extracted features to classify individuals into PD and healthy groups based on their spiral drawings.

The results from both parts of the project demonstrate promising performance in PD detection, with accuracies ranging from 70% to 80% depending on the machine learning algorithm and the type of data used. The proposed approach shows potential for non-invasive and cost-effective screening of Parkinson's disease, facilitating early detection and intervention to improve patient outcomes.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **Serial No.** | **Content** | **Page No.** |
|  | Candidates Declaration | 2 |
|  | Supervisors Declaration | 2 |
|  | Acknowledgement | 3 |
|  | Abstract | 4 |
|  | List of Figures | 5 |
|  | List of Abbreviations | 5 |
| 1 | Introduction | 6 |
| 1.1 | Overview | 6 |
| 1.2 | Existing Systems | 6 |
| 1.3 | Literature Review | 7 |
| 1.4 | Objective Of the Project | 7 |
| 2 | Data-Set Description | 8 |
| 3 | Methodology | 10 |
| 3.1 | Data Pre-processing | 10 |
| 3.2 | Exploratory Data Analysis and Visualization | 10 |
| 3.3 | Methods | 10 |
| 3.3.1 | Voice Analysis Method | 10 |
| 3.3.2 | Hand PD Method | 11 |
| 3.4 | Validation and testing | 11 |
| 4 | Results | 12 |
| 4.1 | Voice Recordings Analysis | 12 |
| 4.1.1 | Algorithm Performance | 12 |
| 4.1.2 | Optimized Model Performance | 12 |
| 4.1.3 | Feature Importance | 12 |
| 4.2 | Hand PD Analysis | 13 |
| 4.2.1 | Algorithm Performance | 13 |
| 4.2.2 | Optimized Model Performance | 13 |
| 4.2.3 | Feature Importance | 14 |
| 5 | Conclusion | 16 |
| 6 | Future Scope | 16 |
|  |  |  |

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No** | **Figure Description** | **Page No** |
| 1. | Voice recording analysis accuracy results | 12 |
| 2. | Spiral Drawing accuracy results | 13 |
| 3. | Website to upload the hand-drawn images | 14 |
| 4. | Results of uploaded hand drawn images | 14, 15 |

**List of Abbreviations**

1. PD - Parkinson's Disease
2. ML - Machine Learning
3. SVM - Support Vector Machine
4. LR - Logistic Regression
5. RF - Random Forest
6. HOG - Histogram of Oriented Gradients
7. RFC - Random Forest Classifier
8. NB - Naive Bayes
9. MRI - Magnetic Resonance Imaging
10. PET - Positron Emission Tomography

**1. INTRODUCTION**

Parkinson's disease (PD) is a progressive neurodegenerative disorder that disrupts the central nervous system, leading to a decline in motor function. Characteristic symptoms include tremors, rigidity, slowness of movement (bradykinesia), and impaired balance and coordination. Early detection of PD is essential, as it allows for the prompt initiation of treatment strategies that can improve a patient's quality of life and slow disease progression. Traditional diagnosis of PD relies on a neurological examination and a patient's reported experience with symptoms. However, this approach can be subjective and may not capture the disease in its early stages.

This project proposes a novel approach to PD detection using machine learning (ML) algorithms. By leveraging non-invasive and readily obtainable data sources like voice recordings and spiral drawings, we aim to develop a cost-effective and objective method for early PD identification. This approach has the potential to revolutionize PD diagnosis by enabling earlier intervention, improved patient outcomes, and potentially even the discovery of new disease biomarkers through the analysis of voice and handwriting patterns.

**1.1 Overview**

Analysing voice recordings in this project will involve extracting features that can be subtly altered in Parkinson's patients. These features may include vocal tremor, which can manifest as a wavering in pitch or amplitude. We can also analyse speech prosody, which refers to the rhythm, intonation, and stress patterns of speech. Changes in prosody, such as a monotonous tone or reduced volume, can be indicative of PD.

Spiral drawing analysis, on the other hand, will focus on characteristics that reveal motor control impairments. Machine learning models can be trained to identify features like the size and shape of the spiral, the presence of tremor lines, and the level of detail and consistency within the drawing. These features can deviate significantly in Parkinson's patients due to rigidity, slowness of movement, and impaired hand-eye coordination. By analysing both voice and spiral drawing data, we aim to develop a comprehensive and objective approach to Parkinson's disease detection.

**1.2 Existing Systems**

* **Clinical Diagnosis:** Parkinson's disease diagnosis traditionally relies on clinical evaluation by neurologists, which can be subjective and may not detect early-stage PD reliably.
* **Diagnostic Tests:** Imaging studies like MRI and PET scans provide valuable information but are expensive and not suitable for routine screening.
* **Biological Markers:** Biomarkers in cerebrospinal fluid or blood show promise but are still under investigation and not widely available.
* **Voice-Based Systems:** Analysing speech recordings for changes in vocal characteristics associated with PD has shown promise.
* **Drawing Analysis:** Analysing drawings, such as spirals, has been explored for insights into motor function and coordination.

Despite advancements, existing systems have limitations like subjectivity, lack of sensitivity in early detection, and the need for specialized equipment. There is a need for non-invasive, cost-effective methods for PD screening, motivating the development of novel approaches using machine learning applied to voice recordings and spiral drawings.

**1.3 Literature Review**

The Parkinson's disease (PD) diagnosis has traditionally relied on clinical evaluation, neuroimaging, and biological markers. Clinical diagnosis by neurologists suffers from subjectivity, while neuroimaging techniques like MRI and PET scans are expensive and not ideal for routine screening. Emerging research focuses on digital biomarkers, including voice analysis and drawing analysis, as potential diagnostic tools. Machine learning techniques, particularly in the realm of voice and drawing analysis, show promise in aiding PD diagnosis.

Comparing existing diagnostic methods, machine learning approaches offer potential advantages in terms of cost-effectiveness, non-invasiveness, and scalability. Voice-based systems and drawing analysis, when coupled with machine learning algorithms, have demonstrated promising results in early PD detection compared to traditional methods. These techniques provide objective, quantitative measures that complement existing diagnostic modalities.

**1.4 Objectives of the Project**

1. Develop a machine learning-based system for early Parkinson's disease detection using voice recordings and spiral drawings.
2. Implement robust algorithms for feature extraction and classification from voice recordings and drawing data.
3. Evaluate the performance of the developed system against existing diagnostic methods.
4. Provide a user-friendly interface for easy accessibility and usability by clinicians and patients.
5. Validate the system's effectiveness through rigorous testing and validation studies.

**2. DATASET DESCRIPTION**

This project utilizes two separate datasets to analyse voice recordings and spiral drawings for Parkinson's disease (PD) detection using machine learning algorithms.

**Dataset 1: Voice Recordings**

* **Source:** Department of Neurology, Cerrahpaşa Faculty of Medicine, Istanbul University
* **Participants:**
  + PD Group: 188 patients (107 men, 81 women)
  + Age Range: 33 - 87 years (mean: 65.1 years, SD: 10.9 years)
  + Control Group: 64 healthy individuals (23 men, 41 women)
  + Age Range: 41 - 82 years (mean: 61.1 years, SD: 8.9 years)
* **Recording Details:**
  + Microphone sampling rate: 44.1 kHz.
  + Speech task: Sustained phonation of vowel "/a/" (repeated three times)

**Dataset 2: Spiral Drawings**

* **Source:** NIATS of Federal University of Uberlândia (curated by Adriano de Oliveira Andrade and Joao Paulo Folado)
* **Description:** This dataset contains digital spiral drawings categorized as Parkinson's disease (PD) or control. Machine learning models will analyse features like shape deviations, line quality, and completion level to learn the motor control variations associated with PD, potentially aiding in early disease detection.
* **Number of Images:** 102
* **Data Split:**
  + Training Set: 72 images
  + Testing Set: 30 images

**3. METHODOLOGY**

This study employed a comprehensive machine learning approach to develop a Parkinson's disease prediction model. The methodology involved several key steps, including data preprocessing, exploratory data analysis, model training, and evaluation. We have used 2 different approaches to solve this problem, one is voice analysis and other is hand movement analysis. The following sections outline the detailed procedures undertaken.

**3.1 Data Preprocessing:**

The raw dataset underwent pre-processing to remove null values and missing values. Removing of these are essential in training of the model as they can interfere with the process and give inaccurate results which further hinders the models and the result.

**3.2 Exploratory Data Analysis and Visualizations:**

* Statistical Analysis: Descriptive statistics, such as mean, median, and standard deviation, were computed to summarize the distribution of features in the dataset. Additionally, hypothesis testing methods may be employed to compare feature distributions between different groups (Parkinson's vs. non-Parkinson's).
* Visualization Techniques: Various visualization techniques, including histograms, box plots, and scatter plots, were utilized to explore the relationships between features and identify potential patterns or trends.

**3.3 Methods**

* + 1. **Voice Analysis Method**
  + **Feature Extraction:** In the voice analysis method, we meticulously extracted pertinent features from the pre-processed voice data. These features were carefully selected to capture distinct patterns and characteristics indicative of Parkinson's disease.
  + **Model Implementation:** We employed three distinct machine learning models for the task: Support Vector Machine (SVM), Logistic Regression, and Random Forest. Each model was trained on the extracted voice features to learn the underlying patterns associated with Parkinson's disease.
  + **Evaluation:** To gauge the performance of each model, we conducted a comprehensive evaluation using a range of metrics, including accuracy, precision, recall, and F1-score. This thorough assessment provided valuable insights into the efficacy of each model in accurately detecting Parkinson's disease from voice data.
  + **Model Selection:** Following the evaluation, we identified Random Forest as the most promising model for voice-based PD detection. Random Forest exhibited superior performance compared to SVM and Logistic Regression, demonstrating its effectiveness in accurately identifying Parkinson's disease from voice samples.
    1. **Hand PD Detection Method**
  + **Feature Extraction:** In the hand PD detection method, we employed sophisticated techniques to extract features from hand images. Specifically, we utilized Histogram of Oriented Gradients (HOG) representations to capture the intricate patterns present in the hand images.
  + **Model Utilization:** We leveraged a diverse set of machine learning models for the task, including Logistic Regression, Linear Discriminant Analysis, K-Nearest Neighbors Classifier, Random Forest Classifier, Decision Tree Classifier, Gaussian Naive Bayes, and Support Vector Machine. Each model was trained on the extracted HOG features to discern the presence of Parkinson's disease.
  + **Performance Assessment:** Similar to the voice analysis method, we rigorously evaluated the performance of each model using standard evaluation metrics. This meticulous assessment provided a comprehensive understanding of the strengths and weaknesses of each model in accurately detecting Parkinson's disease from hand images.
  + **Model Selection:** Based on the evaluation results, Random Forest emerged as the optimal model for hand-based PD detection. Random Forest exhibited superior performance compared to other models, underscoring its efficacy in accurately identifying Parkinson's disease from hand images.
  1. **Validation and Testing**
* **Cross-validation:** To validate the chosen models and ensure their reliability, we employed robust techniques such as k-fold cross-validation. This validation method helped mitigate overfitting and provided assurance regarding the generalizability of the models.
* **Testing:** Following validation, we subjected the final models to rigorous testing using independent datasets. This testing phase was crucial in assessing the generalizability and robustness of the models in real-world scenarios, thereby validating their effectiveness in accurately detecting Parkinson's disease.

**4. RESULTS**

**Part 1: Voice Recordings Analysis**

The results obtained from the analysis of voice recordings for early Parkinson's disease detection are summarized as follows:

**4.1.1 Algorithm Performance:**

* Various machine learning algorithms, including Logistic Regression, Random Forest Classifier, Support Vector Machine, etc., were evaluated for their performance in classifying voice recordings as indicative of Parkinson's disease or not.
* The Random Forest Classifier achieved the highest accuracy of 86.8% on the validation dataset.

**4.1.2 Optimized Model Performance:**

* Grid search was employed to optimize hyperparameters for the Random Forest Classifier.
* The best parameters obtained were a maximum depth of 31 and the number of estimators set to 14.

**4.1.3 Feature Importance:**

* Feature importance analysis revealed that the 'PPE' (Pitch Period Entropy) feature contributed significantly to the classification of Parkinson's disease in voice recordings.

A black and white screen with numbers

Description automatically generated

Fig(1)

**A graph with different colored lines

Description automatically generated**

Fig(2)

**Part 2: Spiral Drawings Analysis**

The results obtained from the analysis of spiral drawings for early Parkinson's disease detection are summarized as follows:

**4.2.1 Algorithm Performance:**

* Various machine learning algorithms, including Logistic Regression, Linear Discriminant Analysis, K-Neighbour’s Classifier, Random Forest Classifier, Decision Tree Classifier, Gaussian NB, SVM etc., were evaluated for their performance in classifying spiral drawings as indicative of Parkinson's disease or not.
* Random Forest Classifier achieved the highest accuracy of 86.0% on average.

**4.2.2 Optimized Model Performance:**

* The Random Forest Classifier, after optimization using grid search, achieved the highest accuracy of 86.6% on the test dataset.
* The confusion matrix indicated balanced performance in classifying healthy and Parkinson's drawings.

**4.2.3 Grid Search Results:**

* Grid search was performed to optimize hyperparameters for the Support Vector Machine and Random Forest Classifier models.
* The best parameters for the Support Vector Machine were {'C': 1, 'kernel': 'rbf'}.
* For the Random Forest Classifier, the optimized parameters were {'max\_depth': None, 'n\_estimators': 100}.

A screenshot of a computer screen

Description automatically generated

Fig(3)

A screenshot of a computer

Description automatically generated

Fig(4)

A screenshot of a drawing

Description automatically generated

Fig(4)

Overall, both parts of the project demonstrate the effectiveness of machine learning-based approaches in accurately detecting early signs of Parkinson's disease using voice recordings and spiral drawings, respectively. The models developed for both parts offer promising results in diagnosing Parkinson's disease based on non-invasive data sources.

**5. CONCLUSION**

In conclusion, this project presents a comprehensive approach for Parkinson's disease (PD) detection using machine learning algorithms applied to voice recordings and spiral drawings. Through meticulous data preprocessing, feature extraction, model training, and evaluation, we have successfully developed models capable of accurately identifying early signs of PD. The results obtained from both voice analysis and hand movement analysis demonstrate promising performance, with accuracies ranging from 70% to 80% depending on the algorithm and data type. The optimized Random Forest Classifier emerged as the most effective model for both voice-based and hand-based PD detection, showcasing its robustness and versatility in discerning patterns indicative of Parkinson's disease.

**6. FUTURE SCOPE**

1. **Integration of Additional Data Sources:** Future enhancements could involve incorporating additional data sources, such as gait analysis or smartphone sensor data, to further improve the accuracy and reliability of PD detection models.
2. **Exploration of Deep Learning Techniques:** Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), could be explored to leverage the inherent complexity and non-linearity of PD-related patterns in voice recordings and hand movements.
3. **Longitudinal Studies:** Conducting longitudinal studies to track changes in voice characteristics and hand movements over time could provide valuable insights into disease progression and treatment response, facilitating personalized management strategies for PD patients.
4. **Development of User-friendly Applications:** Building user-friendly applications or web platforms that leverage the developed models for PD detection could facilitate widespread adoption and enable easy access for clinicians and patients, thereby promoting early diagnosis and intervention.