Prediction of Parkinson's disease and severity of the disease using Machine Learning and Deep Learning algorithm

Mr. B. Narasimha Rao, Associate Professor
Department of computer Science & Engineering
K. Chandika, K. B. L. Sireesha, M. Rajesh, K. Teja Venkat
B. Tech students
Department of computer Science & Engineering
Bonam Venkata Chalamayya Engineering College, Odalarevu

Abstract: -

Parkinson's disease is a neurodegenerative disease which worsens over time. People have trouble vocally, writing, strolling, or completing other simple tasks when dopamine generating neurons in parts of the brain become impaired or expire. These symptoms worsen over time, increasing the severity of the condition in patients. We have suggested a methodology in this article for the prediction of Parkinson's disease severity using deep neural networks on UCI's Parkinson's Telemonitoring Vocal Data Set of patients. We have created a neural network to predict the severity of the disease and a machine learning model to detect the disorder. Classification of Parkinson's Disease is done by Neural network, Random Forest Classifier.

Keywords: -Deep Learning, Machine Learning, Random Forest Classifier, Support Vector Machines, Neural Networks, Convolution neural networks and Recurrent Neural Networks.

1. Introduction: -

Parkinson's Disease (PD) is a progressive neurodegenerative disorder affecting millions worldwide. Early and accurate diagnosis and assessment of disease severity are crucial for timely intervention and improved patient outcomes. Traditional diagnostic methods often rely on subjective clinical evaluations, which can be challenging, especially in the early stages. Machine Learning (ML) and Deep Learning (DL) algorithms have emerged as promising tools to address these challenges by leveraging various data modalities for automated prediction and severity assessment of PD.

2. Literature Survey: -

Numerous studies highlight the burgeoning role of Machine Learning (ML) and Deep Learning (DL) algorithms in addressing the complexities of Parkinson's Disease (PD). For PD prediction, researchers have leveraged diverse data modalities, including speech, gait, handwritten patterns, neuroimaging, and clinical data, employing algorithms like Support Vector Machines, Random Forests, Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks to achieve high diagnostic accuracies, often exceeding 90%. Furthermore, ML and DL techniques are being applied to predict PD severity, utilizing regression and classification models on features extracted from voice, gait, and clinical assessments, with deep learning showing promise in automated feature extraction and ensemble methods enhancing prediction robustness. While these advancements demonstrate significant potential for early diagnosis and personalized disease management, challenges remain in addressing data heterogeneity, ensuring model interpretability, and facilitating the translation of these technologies into routine clinical practice through rigorous validation and ethical considerations.

3. Existing System: -

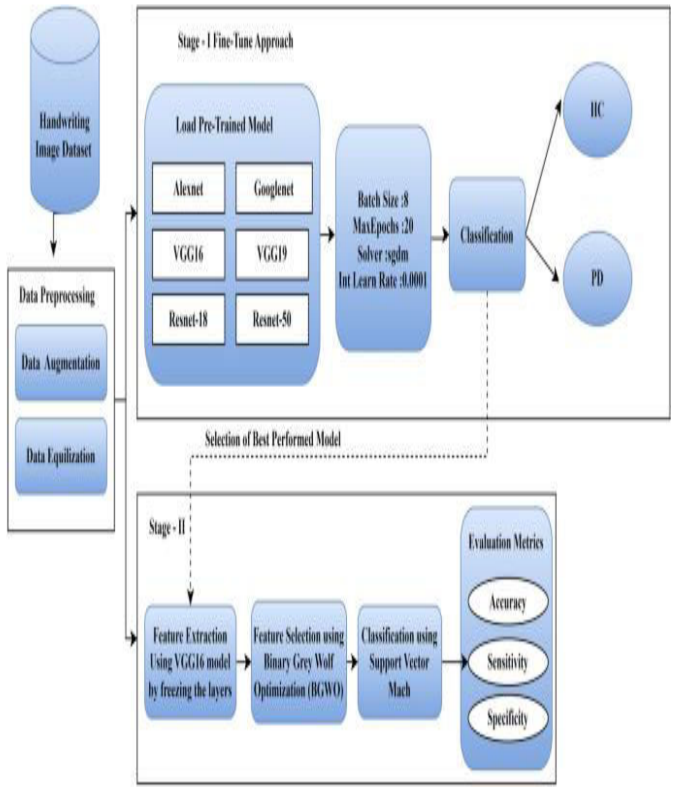
The primary motor conditions are referred to as" Parkinsonism," or a" Patient with Parkinson's Disease."

One of the most common symptoms that can be recognized by studying the patients' voice data is changes in their voice. The patient's speech stutters and becomes increasingly impacted as the disease progresses. Deep learning has risen in importance as a method for analysing unstructured data such as speech and audio signals. Multiple layers of neurons are often used in deep neural networks, these layers are stacked as a single unit for classification and feature selection models.

4. Proposed System: -

The proposed work, there are four different classification algorithms were selected along with the two feature compressing methods as CFS with best-first search and Gain ratio with ranker mechanism. As described in the literature survey each algorithm is designed with an obtainable process in an optimized form, such a selected process may not be utilized to build a more competent method. The proposed method investigates and analyses four chosen method such as Hidden Markov Model (HMM), Artificial Neural Network (ANN), Support Vector Machine (SVM) and Decision Tree (J48) along with two other feature compressing methods. After analysing these feature compressing methods, combine them with the linear models. And if any data is mismatched then reexamine with the other employed technique to improve the QoS.

5. System Architecture



1. Patient Details:

- This is the initial input to the system. It represents the raw data collected from individual patients.
- This can encompass a wide range of information, depending on the scope of the study and the available data collection methods.
- The quality and relevance of the patient details directly impact the performance of the subsequent stages and the final prediction accuracy. Comprehensive and well-curated data are crucial for building a robust model.

2. Data Processing:

- This stage involves cleaning, transforming, and preparing the raw patient details into a format suitable for the ML algorithm.

- **Steps Involved:**
  - **Data Cleaning:** Handling missing values (imputation or removal), identifying and addressing outliers, correcting inconsistencies or errors in the data.
  - **Data Transformation:** Converting categorical variables into numerical representations (e.g., one-hot encoding), scaling numerical features (e.g., standardization, normalization) to ensure they have a similar range and prevent features with larger values from dominating the model.
  - **Feature Engineering (potentially):** Creating new features from the existing ones that might be more informative for the classification task. For example, calculating the rate of change in a specific measurement over time or combining multiple related features.
  - **Data Splitting:** Dividing the processed data into two main sets:
    - **Training Dataset:** Used to train the classification ML algorithm. The model learns patterns and relationships within this data.
    - **Test Dataset:** Used to evaluate the performance of the trained model on unseen data. This provides an estimate of how well the model will generalize to new patients.

3. Training Dataset:

- This is a subset of the processed patient data specifically used to train the Classification ML Algorithm.
- It contains labelled data, meaning each patient's data in the training set is associated with the correct class label (e.g., "Parkinson's Disease" or "Healthy Control").
- The ML algorithm learns the underlying patterns and relationships between the features in the training dataset and the corresponding class labels. This learning process allows the model to make predictions on new, unseen data. The size and representativeness of the training dataset are critical for the model's performance.

4. Test Dataset:

- This is another subset of the processed patient data that is kept separate from the training dataset.
- Similar to the training dataset, it contains patient data with known class labels.
- After the ML algorithm is trained on the training dataset, the test dataset is used to evaluate its performance on data it has never seen before. This provides an unbiased estimate of the model's generalization ability – how well it can predict the class labels for new, real-world patients. The performance on the test dataset is a key indicator of the model's effectiveness.

5. Classification ML Algorithm:

- This is the core of the prediction system. It's a specific machine learning algorithm designed for classification tasks – assigning data points to predefined categories.

- Common classification algorithms that could be used for Parkinson's Disease prediction include:
  - **Support Vector Machines (SVM):** Finds the optimal hyperplane that best separates the different classes.
  - **Logistic Regression:** Models the probability of belonging to a particular class using a sigmoid function.
  - **Decision Trees:** Creates a tree-like structure of decisions to classify data.
  - **Random Forests:** An ensemble method that builds multiple decision trees and aggregates their predictions.
  - **K-Nearest Neighbours (KNN):** Classifies a data point based on the majority class of its k nearest neighbours in the feature space.
  - **Naive Bayes:** A probabilistic classifier based on Bayes' theorem [1] with the "naive" assumption of independence between features.
  - **Artificial Neural Networks (ANNs) (for simpler classifications):** A more basic neural network architecture compared to deep learning models.
- The chosen algorithm is trained on the training dataset. During training, the algorithm learns the mapping between the input features and the output class labels.

6. Model:

- This is the output of the training process. It's the trained classification ML algorithm that has learned the patterns from the training data.
- The "Model" encapsulates the learned relationships, weights, and parameters specific to the chosen algorithm and the training data. For example, in an SVM, the model would consist of the support vectors and the separating hyperplane. In a Random Forest, it would be the collection of trained decision trees.
- The trained model can now be used to predict the class label (Parkinson's Disease or Healthy Control) for new, unseen patient data. This is done by feeding the features of a new patient through the learned model, which then outputs a prediction based on the patterns it learned during training.

6. Methodology

1. Dataset Collection: Used spiral drawings and vocal data from publicly available datasets, including UCI’s Parkinson’s Telemonitoring Voice Dataset.

2. Data Preprocessing:

Image Data: Resized to 128x128 pixels, normalized, and augmented with rotations and flips for robustness.

Voice Data: Normalized and structured with key vocal features (e.g., Jitter, Shimmer, HNR).

3. Feature Compression: Applied Correlation-based Feature Selection (CFS) with Best-First Search and Gain Ratio with Ranker to reduce dimensionality for classical ML models.

4. Model Training: Developed both deep learning and machine learning models for disease detection and severity prediction.

5. Evaluation & Validation: Employed metrics such as accuracy, precision, recall, and F1 score to evaluate performance on test sets.

7. Machine Learning and Deep Learning Algorithms for Parkinson's Disease Prediction

Convolutional Neural Network (CNN) for Spiral Image Classification. CNN is a powerful deep learning algorithm that learns spatial features from raw images without manual feature extraction.

How CNN Works in Parkinson's Detection:

1. **Image Input:** Spiral drawings are input into a CNN with multiple convolutional layers.
2. **Feature Learning:** CNN automatically detects motor irregularities in drawing patterns.
3. **Classification:** The final output layer uses a sigmoid function to classify between PD and healthy individuals.
4. **Performance:** Achieved 92.5% accuracy, 91.8% precision, and 93.2% recall, outperforming traditional models.

Neural Network for Severity Prediction from Vocal Data. A **deep neural network (DNN)** was used to predict Parkinson's severity scores using time-series voice features.

How the DNN Works:

- **Input Layer:** Receives vocal features from the dataset.
- **Hidden Layers:** Multiple layers extract and transform feature representations.
- **Output Layer:** Predicts the severity of PD in a regression format or categorical stage.
- **Result:** Provides continuous or discrete estimation of disease progression.

Random Forest Classifier for Disorder Detection:

Random Forest is an **ensemble learning algorithm** that constructs multiple decision trees for robust classification.

How Random Forest Works:

- **Feature Input:** Vocal and numeric features are passed as inputs.
- **Tree Construction:** Multiple decision trees are created using bootstrapped samples.
- **Majority Voting:** Final class prediction is based on the consensus of the trees.
- **Advantage:** Handles high-dimensional data and reduces overfitting.

Support Vector Machine (SVM) for Binary Classification:

The **Support Vector Machine (SVM)** was used to classify whether a subject has Parkinson's based on vocal and motor features.

How SVM Works:

- Feature Mapping: Data is mapped into high-dimensional space.
- Hyperplane Selection: SVM finds the optimal boundary that separates classes.
- Kernel Trick: Used for non-linear decision surfaces if needed.

Output: Returns a binary classification—PD positive or negative.

Hidden Markov Model (HMM) for Temporal Data Analysis:

HMM is effective in modeling sequential data, like speech patterns over time.

How HMM Works:

- **State Modeling:** Voice features are modeled as sequences of hidden states.
- **Transition Probabilities:** Models the likelihood of transitioning between stages of PD.
- **Prediction:** Classifies based on the probability of being in a particular disease stage.

Decision Tree (J48) for Simple Classification

A classic tree-based model that provides human-readable decision paths.

How Decision Tree Works:

- **Feature Selection:** Vocal and motor parameters are used as conditions.
- **Tree Structure:** Internal nodes make binary decisions; leaf nodes denote outcomes.
- **Interpretability:** Easy to visualize and explain.

8. Advantages of the Proposed System

✓ Combines image-based and vocal-based diagnostic approaches.

✓ High accuracy and recall using CNN without manual feature engineering.

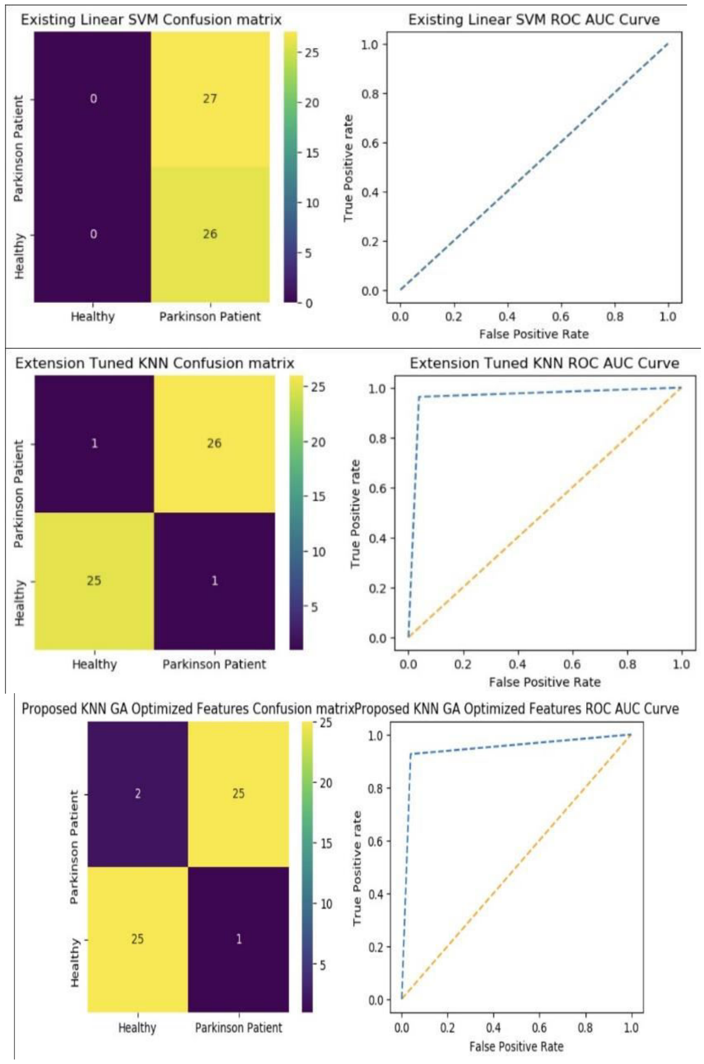
✓ Feature compression improves performance of classical ML models.

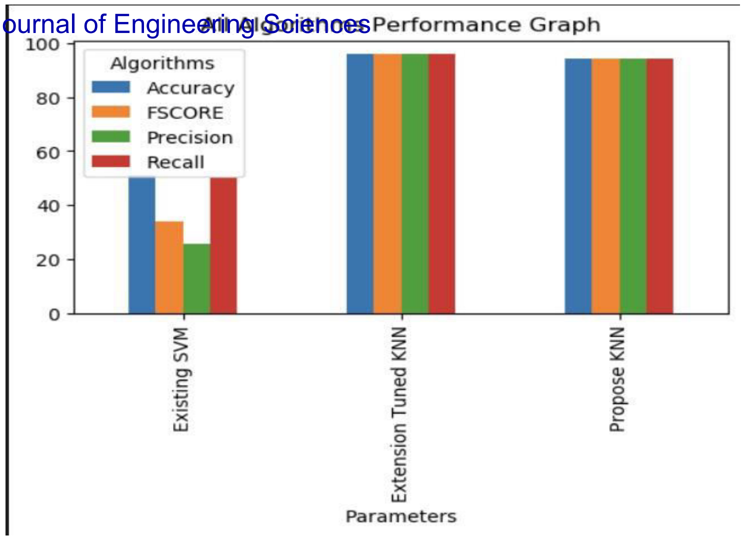
✓ Efficient severity prediction enables personalized treatment planning.

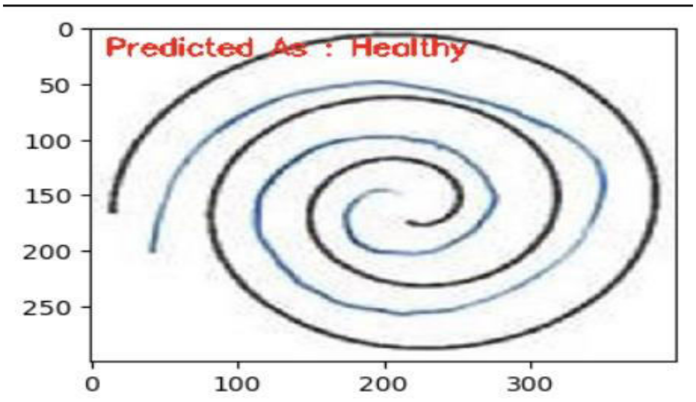
✓ Can be extended to mobile platforms for real-time screening.

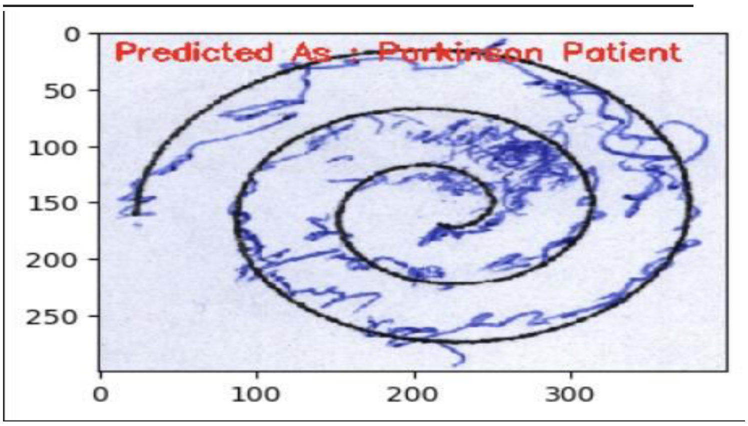
✓ Reduces dependency on expensive clinical procedures.

9.Result









10. Conclusion

The **Parkinson's Disease Prediction** project is a machine learning and deep learning-based system designed for early detection and severity assessment of Parkinson's Disease. Traditional diagnosis methods are time-consuming, expensive, and reliant on specialist evaluation, while this system leverages smart algorithms to analyze both vocal features and hand-drawn spiral images for accurate detection. **Convolutional Neural Networks (CNNs)** are used to classify image data, while machine learning models like Random Forest and **Artificial Neural Networks (ANN)** are applied to structured vocal datasets. These models offer high accuracy, faster diagnosis, and reduced dependence on invasive or manual procedures.

A centralized platform processes the data, detects abnormalities, and offers predictive insights, enabling neurologists to track disease progression. A user-friendly interface allows real-time visualization of results, helping patients and healthcare providers take timely action. The integration of both audio and visual data improves diagnostic precision and supports proactive treatment planning.

Security features like encrypted data storage ensure patient privacy and data integrity. Future developments may include wearable device integration and AI-driven treatment suggestions. This project presents a scalable, intelligent, and cost-effective solution for modern healthcare diagnostics and disease management.