[1]Nandan N
[2]Sanjay Pande M B
[3]Raveesh B N
[4]Rakesh

A Machine Learning Approach to Analyzing DATSCAN SBR Values for the Detection of Parkinson's Disease.

Journal of Electrical Systems

Abstract: - Parkinson's disease (PD) is a progressive neurodegenerative condition characterized by complex symptoms, making early diagnosis challenging. However, early detection is achievable through DaTSCAN, which evaluates brain function instead of focusing solely on anatomical details. This study aims to develop a machine-learning model to distinguish between PD and healthy controls (HC) while examining significant changes in biomarkers in PD patients. Our research focuses explicitly on the Striatal Binding Ratio (SBR) values of the putamen and caudate nucleus, located in the basal ganglia region in the brain. These regions are primarily responsible for cognition, motor skills, and executive functions. The significance of this research lies in its potential to improve early diagnosis of PD using a Random Forest algorithm, which yielded an impressive accuracy of 97%. Timely diagnosis can significantly enhance a patient's quality of life by facilitating early treatment interventions that may slow the progression of Parkinson's disease.

Keywords: DaTSCAN, Parkinson, Caudate and Putamen, Random Forest.

I. INTRODUCTION

Parkinson's disease (PD) is a progressive neurodegenerative disorder that primarily impacts the motor system. Key symptoms of this condition include tremors, rigidity, bradykinesia, and postural instability. In recent years, there has been a noticeable increase in the prevalence of Parkinson's disease, with projections estimating that approximately 10 million people worldwide will be affected by 2040[1]. Researchers have been working on more efficient and reliable diagnostic methods to enhance the accuracy of diagnosis and alleviate the burden on healthcare systems and society. Traditionally, Parkinson's disease is diagnosed primarily through clinical observations and neurological examinations, which can be subjective. This subjectivity may lead to delayed diagnoses, particularly in the early stages when symptoms are most pronounced. A significant challenge arises from the fact that typical motor symptoms usually do not appear until approximately 60-80% of dopaminergic neurons in the substantia nigra have degenerated, highlighting the critical need for early detection methods [2].

In recent decades, the global prevalence of Parkinson's disease has increased significantly. A systematic review and meta-analysis involving 83 studies from 37 countries examined how the prevalence of the disease has changed across various age groups, genders, and geographical and socioeconomic factors. This analysis also considered the Human Development Index (HDI) and the Sociodemographic Index (SDI). Over the past 20 years, there has been a notable rise in Parkinson's disease, with the overall prevalence estimated at 1.51 cases per 1,000 people. This marks a significant increase from 0.9 cases per 1,000 people in the 1980s to 3.81 cases per 1,000 people in recent years (2010-2023) [3].

Parkinson's disease primarily results from the degeneration of dopaminergic neurons in the substantia nigra, leading to various motor symptoms. Although current treatments mainly focus on managing these symptoms, there are currently no therapies available that can halt or slow the neurodegenerative process [4]. This review explores the molecular pathways involved in Parkinson's disease, including protein misfolding, oxidative stress, mitochondrial dysfunction, and the impact of genetic mutations (notably in the α-synuclein and parkin genes). Identifying mutations associated with familial Parkinson's disease has enabled the development of genetic models for investigating specific pathogenic pathways. These models indicate that protein aggregation and impaired protein degradation pathways play a crucial role in the pathogenesis of Parkinson's disease [5].

[1] *Corresponding author: Nandan N, Research Scholar, Department of CSE, GM Institute of Technology, Davanagere, Visvesveraya Technological University, Belagavi, Karnataka, India
[2] Sanjay Pande M B, Research Supervisor, GM Institute of Technology, Davanagere, Visvesveraya Technological University, Belagavi, Karnataka, India
[3] Raveesh B N, Department of Psychiatry, Mysore Medical College, Mysuru, Karnataka, India
[4] Rakesh, Dental surgeon, Private Practitioner, Mysuru, Karnataka, India

The caudate nucleus and putamen are essential structures within the brain, particularly in Parkinson's Disease (PD). The caudate nucleus is crucial for the planning and executing movements, regulating voluntary motor functions as part of the basal ganglia, which coordinates overall movement. Additionally, it is involved in various cognitive processes, including learning and memory. In PD, the caudate nucleus is impacted by the degeneration of dopamine-producing neurons. Notably, recent studies indicate that preserved dopamine levels in the caudate nucleus may correlate with the severity of tremors experienced by patients with PD [6].

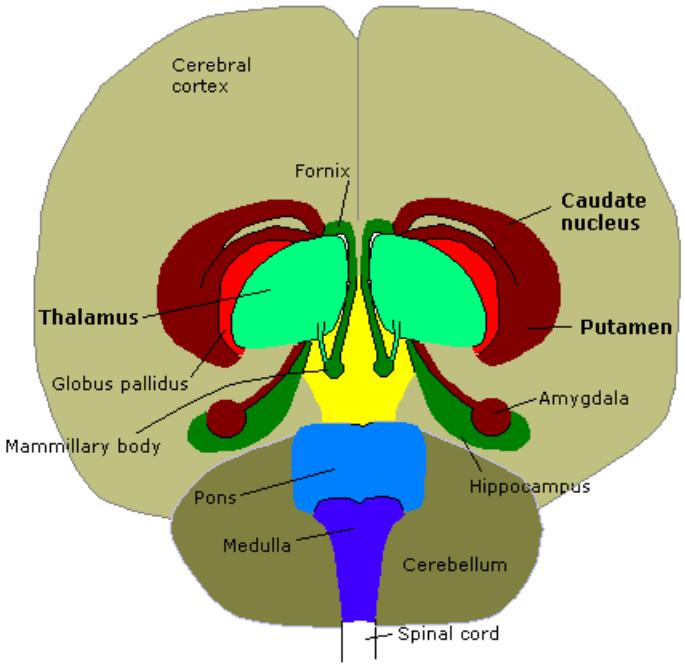

Figure 1: Coronal section of the brain

The putamen plays a crucial role in motor control and affects various types of movement. It works closely with the caudate nucleus and other regions of the basal ganglia. In Parkinson's disease (PD), the degeneration of dopaminergic neurons significantly impacts the putamen. The loss of dopamine in this area is a hallmark of PD and contributes to characteristic motor symptoms, including tremors, rigidity, and bradykinesia. Furthermore, the extent of dopamine loss in the putamen is directly related to the severity of motor symptoms experienced by PD patients [7].

Dopamine is an essential neurotransmitter that regulates the brain's movement, emotional responses, and pleasure-reward systems. The caudate nucleus and putamen, critical basal ganglia components, work with dopamine to manage both movement and cognitive functions [8].

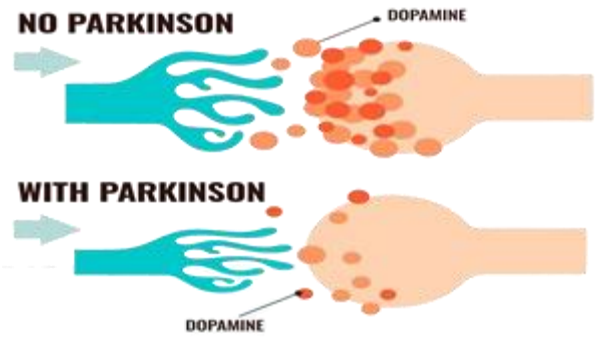

Figure 2: Dopamine Concentration

The structures involved are rich in dopamine receptors (D1 and D2), which play a crucial role in mediating the effects of dopamine released from the substantia nigra. This interaction facilitates smooth and coordinated movements and cognitive processes, including learning, habit formation, and reward processing. Dopamine

signaling in the caudate nucleus and putamen affects motor control and motivation, enabling individuals to adapt their behaviors based on rewards. This loss disrupts normal signaling pathways, resulting in impaired motor control characterized by symptoms such as rigidity, bradykinesia, and tremors. Additionally, reduced dopamine transmission impacts cognitive functions, making it challenging for individuals with PD to plan and execute movements and engage in reward-based learning. The relationship between dopamine deficiency and the functions of the caudate nucleus and putamen highlights the complex pathophysiology of Parkinson's disease and its effects on movement and cognition. [9].

Recent advancements in artificial intelligence (AI) and machine learning (ML) have created new opportunities for the early detection and diagnosis of Parkinson's disease (PD). ML algorithms can analyze complex patterns in various types of medical data, including neuroimaging, voice recordings, gait analysis, and genetic markers, to identify subtle indicators of PD before clinical symptoms become evident. The integration of multiple data sources and ML techniques has shown promising results in enhancing diagnostic accuracy and reducing the time required for diagnosis. Neuroimaging techniques, especially DaTSCAN imaging, have become valuable tools for diagnosing PD when combined with ML methods. Studies have demonstrated that deep learning models can automatically extract relevant features from DaTSCAN images and accurately differentiate PD patients from healthy individuals. [10]. Early diagnosis and prediction of disease progression are crucial. Biomarkers are essential for diagnosing and predicting PD. These include clinical, neuroimaging, and biofluid-based biomarkers. Studies show that combining different biomarkers and complex ML techniques can enhance diagnostic accuracy and predict disease progression more effectively [11].

PD diagnosis is often based on symptoms, medical history, and physical exams. DaTscan helps in cases where diagnosis is difficult. DaTscan uses a radiotracer to detect dopamine transporter activity in the brain. Reduced activity indicates PD. DaTscan helps differentiate PD from essential tremor, drug-induced parkinsonism, and psychogenic parkinsonism [12]. Despite significant advancements, challenges persist in developing reliable machine learning (ML) diagnostic tools for Parkinson's Disease (PD). Data quality, standardization, model interpretability, and clinical validation must be addressed. Additionally, the integration of ML systems into clinical practice requires careful consideration of ethical implications, regulatory requirements, and the training of healthcare providers. Looking toward the future, the successful detection of PD will depend on creating comprehensive ML systems that can integrate multiple data sources, adapt to individual patient characteristics, and provide interpretable results to support clinical decision-making. Ongoing research focuses on enhancing model performance, developing standardized data collection and analysis protocols, and validating ML-based diagnostic tools in large-scale clinical trials [13].

II. Literature Review

The study employed an artificial neural network (ANN) to differentiate Parkinson's Disease (PD) from similar conditions by analyzing brain imaging data obtained from DAT-SPECT. This research focused on a specific brain region known as the putamen. The dataset included images collected at the National Cheng Kung University Hospital between 2017 and 2019 for training and validation and a separate test set gathered in early 2020. Using the AlexNet architecture, the ANN achieved an impressive accuracy of 86% in identifying PD, demonstrating good sensitivity and specificity. However, the study encountered several challenges, such as the limited size of the dataset, potential difficulties in applying the model across various imaging sources, and the necessity for more advanced neural network architectures [14].

This study aimed to classify DaTscan SPECT images to differentiate between Parkinson's Disease (PD) cases and non-PD cases using a deep convolutional neural network. The dataset, sourced from the Parkinson's Progression Markers Initiative (PPMI), comprised 659 DaTscan images, with 449 labeled PD and 210 as non-PD. To prepare the images for analysis, preprocessing steps were performed to enhance regions with high dopaminergic activity, and the images were resized for compatibility with the InceptionV3 model, a neural network architecture pre-trained on the ImageNet database. Data augmentation techniques were employed to address the challenges posed by the relatively small dataset, along with a ten-fold cross-validation method that ensured balanced classes across the folds. The final model achieved an impressive accuracy of 98.48%. However, the study had several limitations. The small size of the dataset raises concerns about potential overfitting and the reliance on selective slices rather than full 3D images, which may overlook important information. Additionally, using human-labeled

data can introduce subjectivity, and the narrow age range within the dataset can limit the generalizability of the findings. [15].

This study presents an ensemble of deep learning models designed to detect Parkinson's Disease (PD) using DaTscan SPECT images sourced from the publicly available Parkinson's Progression Markers Initiative (PPMI) dataset, which comprises 645 images (432 PD cases and 213 non-PD cases). The research utilized four pre-trained Convolutional Neural Network (CNN) models—VGG16, ResNet50, Inception-V3, and Xception—as base classifiers. Each model processed images that were resized appropriately, and a Fuzzy Rank Level Fusion (FRLF) ensemble method was employed to integrate the individual models' outputs, aiming to enhance PD detection accuracy. The ensemble model attained an impressive accuracy rate of 98.45%. However, the study acknowledges limitations, including a lack of generalizability due to the dataset's size and the models' sensitivity to specific data characteristics, which may restrict their applicability in real-world settings. Future research could explore the use of complete 3D DaTscan volumes and evaluate the FRLF ensemble approach on various medical datasets to determine its broader effectiveness. [16].

The study titled "Classifying Parkinson's Disease Identification within the SWEDD Group" investigates the application of machine learning (ML) algorithms to differentiate between patients with Parkinson's Disease (PD) and Healthy Controls (HC) within the SWEDD group. It utilized a dataset of 548 subjects sourced from the Parkinson's Progression Markers Initiative (PPMI) database, which included both clinical features and DaTSCAN SPECT imaging data. To reduce the number of features, the study employed Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), with LDA demonstrating superior performance. The clustering models utilized included Density-Based Spatial Clustering of Applications with Noise (DBSCAN), K-means, and Hierarchical Clustering, with Hierarchical Clustering achieving the highest accuracy, sensitivity, and specificity levels. The overall accuracy reached 64%. The findings highlight the potential of machine learning models for early diagnosis of Parkinson's Disease; however, the study also acknowledged limitations such as the high dimensionality of the data and the necessity for further validation. [17].

A machine learning model has been developed for the early diagnosis of Parkinson's Disease (PD) using DaTSCAN images. The dataset was obtained from the Parkinson's Progression Markers Initiative (PPMI) database. The methodology centered on utilizing transfer learning with the VGG16 Convolutional Neural Network (CNN) model to classify subjects based on their DaTSCAN scans, achieving an impressive accuracy of 95.2%. The Local Interpretable Model-Agnostic Explainer (LIME) framework enhanced the model's predictions' interpretability, highlighting the brain scans' most influential features. The results of this study suggest that the proposed system could effectively assist in the early diagnosis of PD. Nevertheless, limitations remain, including the necessity for further validation and the challenges associated with the model's interpretability [18].

The study aims to develop a fully automated CAD system to classify DaTSCAN SPECT images. It utilizes images from the Parkinson Progression Markers Initiative (PPMI). The images underwent normalization and were reduced using a mask, followed by a Gray Level Co-occurrence (GLC) matrix analysis to extract Haralick texture features. A Support Vector Machine (SVM) classifier was employed to identify patterns associated with Parkinson's disease (PD). The system achieved an impressive accuracy of up to 95.9% and a sensitivity of 97.3%, showcasing its robustness and potential effectiveness for clinical application. The limitations of the current methodologies include a dependence on the quality of images and the complexity of the computational processes involved. Future advancements in machine learning techniques are expected to enhance diagnostics accuracy significantly. [19].

A novel approach for early detection of Parkinson's Disease (PD) using deep learning models optimized with the Grey Wolf Optimization (GWO) algorithm. The study utilized T1, T2-weighted MRI, and SPECT DaTscan datasets from the Parkinson's Progression Markers Initiative (PPMI). The methodology involved preprocessing MRI images, removing empty tuples, and applying GWO to fine-tune the hyperparameters of four deep learning models: GWO-VGG16, GWO-DenseNet, GWO-DenseNet + LSTM, and GWO-InceptionV3. Additionally, a hybrid model combining GWO-VGG16 and InceptionV3 was proposed. The results showed that the hybrid model achieved an accuracy of 99.94% and AUC of 99.99% for the T1 and T2-weighted datasets and 100% accuracy and

99.92% AUC for the SPECT DaTscan dataset. However, limitations include testing the model with the larger datasets and further validation to ensure the robustness of the models [20].

The prediction of Parkinson's disease (PD) was conducted using DaTSCAN images derived from the Parkinson's Progressive Markers Initiative (PPMI) dataset. VGG-16 and AlexNet were employed for feature extraction, and classification was done using a Multi-Kernel Support Vector Machine (MSVM). The MSVM model demonstrated impressive accuracy, achieving 98.60% in classifying PD cases. While the model was trained on a smaller dataset, validating it on a larger dataset is crucial to ensure robustness [21].

Research is exploring using multi-modal features to improve the diagnostic accuracy of Parkinson's disease (PD). Early diagnosis is crucial for effective management. Two frameworks were examined: the feature-level framework analyzed a heterogeneous dataset with deep learning models. In contrast, the modal-level framework reduced MRI features using the ReliefF method before combining them with SPECT and CSF features. Due to an imbalance in the dataset (73 PD and 59 healthy subjects), performance metrics like F1-score, sensitivity, specificity, and accuracy were used. The convolutional neural network (CNN) achieved 93.33% accuracy in the feature-level framework and 92.38% in the modal-level framework, although the smaller dataset may limit real-time accuracy [22].

The paper presented a model developed using GenoML, an automated machine learning package, to predict the risk of Parkinson's disease (PD). It utilized data from the Parkinson’s Progression Marker Initiative (PPMI) and validated the model using the Parkinson’s Disease Biomarker Program (PDBP) dataset. The initial model achieved an area under the curve (AUC) of 89.72% for diagnosing PD, which was validated with an AUC of 85.03% on external data. There is potential for further enhancing the model's accuracy through hyper parameter tuning and multimodal approaches [23].

The study examines the effectiveness of parametric and non-parametric models in diagnosing Parkinson's Disease (PD). The dataset consists of 919 samples from the Parkinson's Progression Markers Initiative (PPMI), which includes 629 samples affected by PD and 290 healthy controls. Logistic regression is used for the parametric modeling, while K-Nearest Neighbors (KNN) is employed for the non-parametric approach. The results indicate that KNN, especially when optimized for the best value of 'k' and the choice of distance metrics, outperforms logistic regression in classification accuracy. Additionally, Analysis of Variance (ANOVA) is utilized to identify significant features. The study achieved an accuracy of 94.82% with logistic regression and 96.8% with KNN. However, there are limitations, including the need for larger datasets and further validation to ensure the robustness of the findings. [24].

This study presents an approach for the early detection of Parkinson's Disease (PD) through deep learning and machine learning techniques. Utilizing data from the Parkinson’s Progression Markers Initiative (PPMI), the research analyzed information from 183 healthy individuals and 401 early-stage PD patients. The methodology involved a comparative analysis of a deep learning model against twelve machine learning and ensemble learning methods, with a particular emphasis on premotor features such as Rapid Eye Movement (REM) sleep behavior disorder, olfactory loss, cerebrospinal fluid data, and dopaminergic imaging markers. The findings revealed that the deep learning model achieved the highest accuracy of 96.45%, surpassing the performance of the other methods. However, the study acknowledges limitations, including the relatively small dataset size [25].

III. PROPOSED METHODOLGY

This research aims to develop and evaluate a machine learning-based classification system that effectively distinguishes between patients with Parkinson's disease and healthy individuals. This classification will be based on the Striatal Binding Ratio (SBR) values extracted from DaTSCAN imaging, explicitly focusing on the caudate nucleus and putamen regions. The specific objectives of the study include:

- Analyzing the SBR values obtained from DaTSCAN to identify significant changes in biomarkers associated with Parkinson's disease in the caudate nucleus and putamen regions.
- Applying an efficient machine learning algorithm to classify individuals with Parkinson's disease and healthy controls.

A. Dataset Used

To prepare the data for our study, we utilized the Parkinson's Progression Markers Initiative (PPMI) database (http://www.ppmi-info.org/data). The PPMI is a landmark observational clinical study focused on identifying biomarkers for the progression of Parkinson's disease through advanced imaging techniques, biological samples, and clinical and behavioral assessments, thereby enhancing the scope of our research.

The dataset comprises 2,071 samples from the PPMI study, which we analyzed. Among these, 1,540 samples from individuals with Parkinson's disease (PD) and 531 control samples reveal significant degeneration in the midbrain regions of PD patients. Additionally, the dataset includes volume measurements of the Caudate and Putamen regions (both right and left sides). We observed a significant volumetric difference between the Putamen and Caudate regions in Parkinson's patients.

B. Dataset Preprocessing

The initial preprocessing phase of the dataset involves two critical steps: removing duplicates and detecting outliers. Duplicate data entries can significantly impact a model's performance by introducing bias and overrepresentation. Therefore, the first step is identifying and removing duplicate rows from the dataset using the pandas drop_duplicates() function. This process eliminated seven duplicate sample rows from a dataset containing 2,064 samples, where 1,534 are Parkinson's disease (PD) samples, and 530 are healthy control (HC) samples. This ensures that each unique case is represented only once in the dataset, maintaining data integrity and preventing potential overfitting in subsequent machine-learning models.

The second crucial preprocessing step utilizes the Isolation Forest (iForest) algorithm for outlier detection and removal. Isolation Forest is particularly well-suited for this task because it efficiently handles high-dimensional data and identifies anomalies based on isolation. The algorithm focuses on the Striatal Binding Ratio (SBR) values from four key regions: the right and left caudate and the right and left putamen, the primary biomarkers for detecting Parkinson's disease.

After implementing both preprocessing steps, the cleaned dataset provides a more reliable foundation for classification. Removing duplicates ensures a unique representation of cases, while the Isolation Forest algorithm effectively detects and removes anomalous SBR values that could skew the classification results. This preprocessing pipeline is essential for maintaining data quality and ensuring that subsequent machine learning models can effectively learn the underlying patterns distinguishing Parkinson's disease patients from healthy individuals based on their SBR values. The cleaned dataset undergoes validation through statistical analysis and visualization techniques to verify the effectiveness of the preprocessing steps and to ensure that the data is adequately prepared for the classification task.

Algorithm: iForest

```
Procedure IsolationForest(X, t, ψ):
    X: input dataset
    t: number of trees
    ψ: sub-sampling size
     Forest = []
     for i in range(t):
          # Select random subsample
          X' = Sample(X, ψ)
          # Create an isolation tree
          Tree = iTree(X', 0, l)
          Forest.append(Tree)
      return Forest
```

```
Procedure PathLength(x, Tree, e):
    x: instance
    Tree: isolation tree
    e: current path length
```

```
if Tree is ExNode:
    return e
a = Tree.splitAtt
if x[a] < Tree.splitValue:
    return PathLength(x, Tree.left, e+1)
else:
    return PathLength(x, Tree.right, e+1)
```

The dataset underwent careful preparation, particularly in encoding the target variable, which differentiates between samples of Parkinson's disease (PD) and Healthy Controls (HC). It initially encompasses 2064 samples, of which 1534 are designated PD and 530 as HC. The Isolation Forest algorithm detected potential anomalies within the remaining data points following a thorough review to identify and eliminate outlier samples. This sophisticated algorithm operates by training a model on what is considered inlier samples, enabling it to effectively isolate and identify deviations that significantly diverge from the overall distribution of most data.

The outcomes of this analysis led to the identification of a particular subset of samples that were categorized as outliers based on the unique features and patterns that emerged from the dataset. This provided invaluable insights into potential irregularities that might warrant further investigation. After successfully removing outliers, the final dataset consists of 1864 samples, including 1451 samples classified as PD and 413 as HC. To visually represent the relationships among the dataset's features, Figure 3 showcases a 3D scatter plot highlighting the interactions between the first three attributes. Additionally, box plots, also known as box-and-whisker plots, are presented to compare various characteristics between the two classes; these are distinctly represented in blue for HC and orange for PD, making it easier to discern and analyze the differences between the groups.

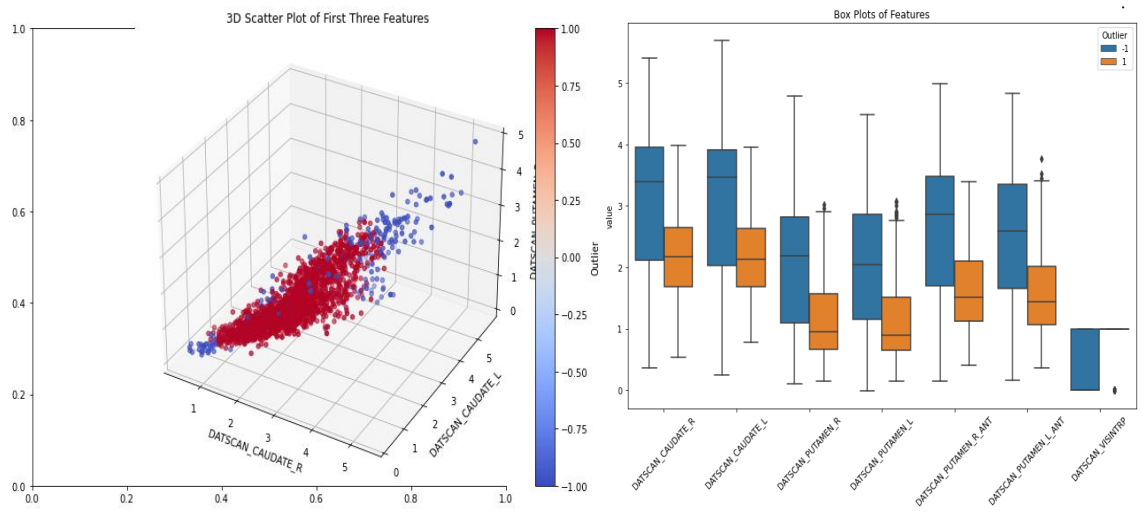

Figure 3: 3D scatter plot and Box plots of features

A pairwise scatter plot matrix, often called a pairs plot or scatterplot matrix, is a sophisticated visualization technique designed to facilitate concurrently examining relationships among multiple variables. This approach is beneficial in exploratory data analysis, allowing analysts to observe trends, correlations, and potential anomalies within a dataset.

As illustrated in Figure 4, the matrix comprises a grid layout where each variable in the dataset is represented on both the horizontal and vertical axes. The result is a comprehensive visualization that presents all possible pairwise combinations of the variables. On the diagonal of the matrix, each element displays a histogram that reveals the distribution characteristics of the respective variable. These histograms can provide valuable insights into the variable's central tendency, variability, and overall distribution shape, such as whether it follows a normal distribution or exhibits skewness.

On the other hand, the off-diagonal elements showcase scatter plots that plot pairs of variables against each other. These scatter plots highlight the nature of the relationship between the two variables—be it linear, nonlinear,

or nonexistent. Observers can identify patterns such as clusters of data points, trends that indicate positive or negative correlations, and potential outliers that may warrant further investigation. Overall, the pairwise scatter plot matrix is an effective tool for gaining a multi-dimensional understanding of the interactions within a dataset, enabling a fine-grained analysis of variable relationships that might not be apparent through univariate analysis alone.

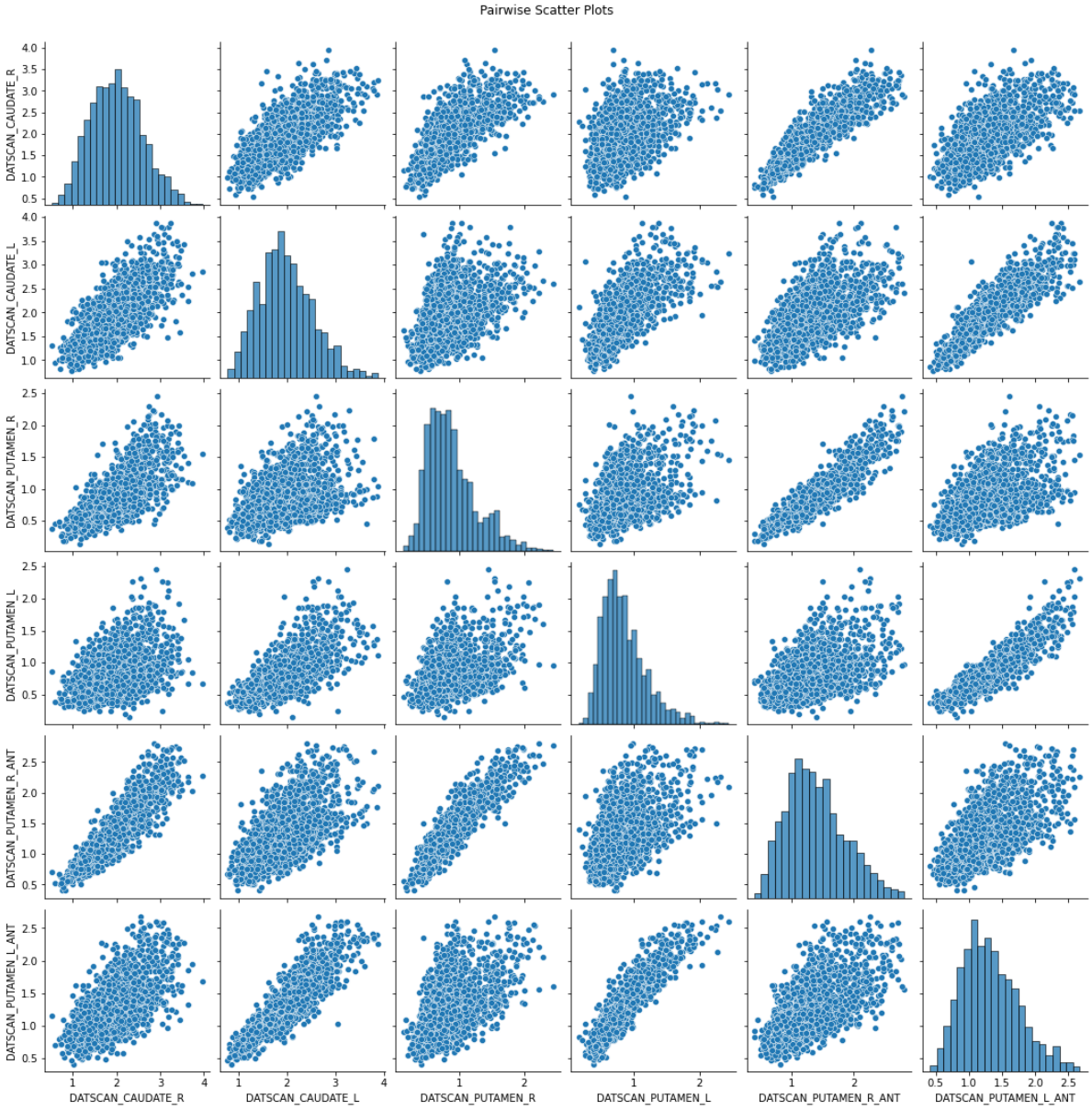

Figure 4: scatter plot of attributes

C. Proposed method

This study presents a comprehensive method for the automated detection of Parkinson's disease using Striatal Binding Ratio (SBR) values obtained from DaTSCAN imaging. Our approach analyzed SBR values from crucial brain regions, specifically the caudate nucleus and putamen. The dataset included 1,864 samples, consisting of 413 samples from healthy controls and 1,451 samples from patients with Parkinson's disease, reflecting a realistic class distribution commonly found in clinical settings. Our preprocessing pipeline, which included duplicate removal and the Isolation Forest algorithm for outlier detection, effectively enhanced the quality of this substantial dataset. This study presents an in-depth methodology for the automated detection of Parkinson's disease, utilizing Striatal Binding Ratio (SBR) values derived from DaTSCAN imaging. The SBR values were meticulously analyzed from two critical brain regions: the caudate nucleus and the putamen, which are essential for understanding the pathophysiology of Parkinson's disease.

The dataset used in this study comprised a total of 1,864 samples. Among these, 413 samples were collected from healthy control individuals, while 1,451 samples were obtained from patients diagnosed with Parkinson's. This distribution mirrors the real-world class distribution observed in clinical environments, enhancing our findings' relevance and applicability. To ensure the integrity and quality of the dataset, we implemented a comprehensive preprocessing pipeline. This involved the removal of duplicate entries to eliminate redundancy and potential biases. We also employed the Isolation Forest algorithm to detect and address outliers within the dataset effectively. This rigorous preprocessing was instrumental in refining the data quality, ultimately paving the way for more accurate and reliable results in the automated detection of Parkinson's disease.

Algorithm: Random_Forest

```
Procedure RandomForest(D, T, m):
    D: Training dataset
    T: Number of trees
    m: Number of features to consider at each split
    forest = []
    for t in range(T):
        # Bootstrap sample
        D_t = Bootstrap(D)
        # Build a decision tree
        tree = BuildDecisionTree(D_t, m)
        forest.append(tree)
    return forest
```

```
Procedure BuildDecisionTree(D, m):
    # Build a single decision tree with random feature selection
    if stopping_criteria_met():
        return Leaf(majority_class)
    # Select m random features
    features = RandomSelect(all_features, m)
    # Find the best split using the Gini index
    best_split = None
    best_gini = float(info)
    for feature in features:
        gini = CalculateGini(D, feature)
        if gini < best_gini:
            best_gini = gini
            best_split = feature
    return Node(best_split,
                left=BuildDecisionTree(left_split),
                right=BuildDecisionTree(right_split))
```

When utilized on the meticulously preprocessed dataset, the Random Forest classifier demonstrated a remarkable accuracy rate of 97% in differentiating between individuals diagnosed with Parkinson's disease and those categorized as healthy controls. This high level of accuracy is particularly noteworthy, especially considering the imbalanced composition of our dataset, which comprises 78% cases of Parkinson's disease contrasted with only 22% healthy individuals. This distribution reflects the real-world patient demographics commonly encountered in specialized movement disorder clinics and poses a challenge for many classification algorithms.

The effectiveness of our Random Forest model can be attributed to several factors. Firstly, the preprocessing stage involved rigorous techniques to ensure data quality and relevance, optimizing the dataset for the model's analysis. Additionally, the Random Forest algorithm is inherently designed to accommodate and manage the complexities associated with imbalanced datasets. Its ensemble learning approach allows it to construct multiple decision trees based on random subsets of the data, thereby improving its robustness and ensuring that it maintains high classification accuracy, even when faced with uneven class distributions. This combination of careful data

handling and the algorithm's strengths contributed significantly to the exceptional performance of our model in accurately classifying patients.

IV. Results and Discussion

The Random Forest classifier applied to the preprocessed dataset achieved an impressive 97% accuracy in distinguishing individuals with Parkinson's disease from healthy subjects. This high level of precision is particularly noteworthy given the uneven distribution of patients, which accurately reflects the typical composition found in specialized movement disorder clinics. The exceptional performance of our model can be attributed to meticulous data preprocessing and the Random Forest algorithm's capacity to manage imbalanced datasets while maintaining excellent classification outcomes effectively. The model also exhibits outstanding performance with a nearly 98% F1 score, signifying a remarkable balance between precision (98.59%) and recall (96.89%). Additionally, an analysis utilizing a confusion matrix— a two-dimensional matrix commonly employed in classification experiments— helps evaluate the performance of the classification system by determining the number of data sets that were accurately classified versus those that were misclassified, allowing for the identification of which data sets tend to be misclassified most frequently.

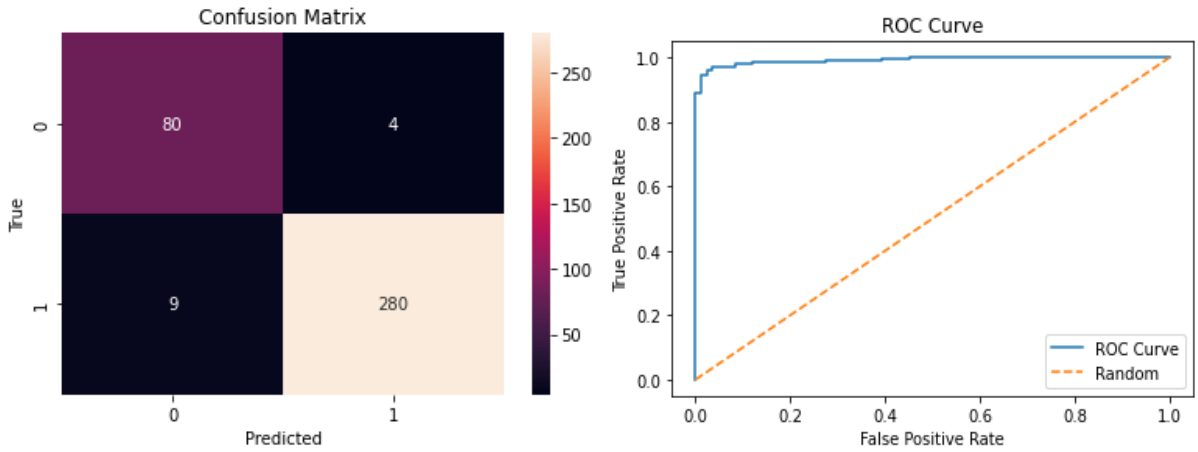

Figure 5: Confusion matrix and ROC Curve

In a receiver operating characteristic (ROC) curve in Figure 5, you can see how well a classification model performs under all kinds of classification thresholds based on the value of the curve. The graph uses two axes: the y-axis displays the true positive rate (TPR), and the x-axis shows the false positive rate (FPR).

Performance metrics: The performance of the proposed model is evaluated through a set of comprehensive metrics designed to assess its effectiveness. These metrics include precision, which indicates the accuracy of the optimistic predictions; recall, which measures the model's ability to identify all relevant instances; F1-score, which provides a balanced assessment of precision and recall; and sensitivity, also known as valid positive rate, which evaluates the proportion of actual positives that are correctly identified. Each of these metrics offers valuable insights into the model's performance, allowing for a thorough analysis of its predictive capabilities.

Metrics	Score
Accuracy	0.9700000000000000
F1 Score:	0.9773123909249564
Precision:	0.9859154929577465
Recall:	0.9688581314878892
Sensitivity:	0.9688581314878892

Proposed model performance comparison: In this study, we aim to evaluate the performance of our proposed model by comparing it with existing research on the specific striatal binding ratios (SBR) obtained through DaTSCAN imaging in the putamen and caudate regions of the brain. Our analysis focuses exclusively on data from the Parkinson's Progression Markers Initiative (PPMI) database, which provides a robust and comprehensive set of information on Parkinson's disease progression. By utilizing this dataset, we seek to assess the effectiveness and

accuracy of our model in predicting and understanding the alterations in these critical brain regions associated with Parkinson's disease, using the established benchmarks set by previous studies.

Author & Year	Dataset Source & Size	Methodology	Results
WuWang et al. (2020) [25]	PPMI database 584 samples, (183 HC and 401 PD)	Deep learning model	96.45% accuracy
Madhusudhana G K et al. (2021) [24]	PPMI database 919 samples, (290 HC and 629 PD)	Logistic Regression and KNN	94.82% LR and 96.8% KNN
Proposed Model	PPMI database 2071 samples, (531 HC and 1540 PD)	Random Forest	97% accuracy

The brain biomarkers selected for this study reveal a distinctive pattern of abnormal development, starting with the left putamen. This area undergoes abnormal changes earlier than other brain regions, which suggests it may play a crucial role in the early stages of disease progression. Following the left putamen, significant abnormal development is observed in the right, indicating a bilateral involvement in the pathological process. Subsequently, the left caudate also exhibits abnormal developmental patterns, with the right caudate showing similar changes last in this sequence indicated in Figure 6.

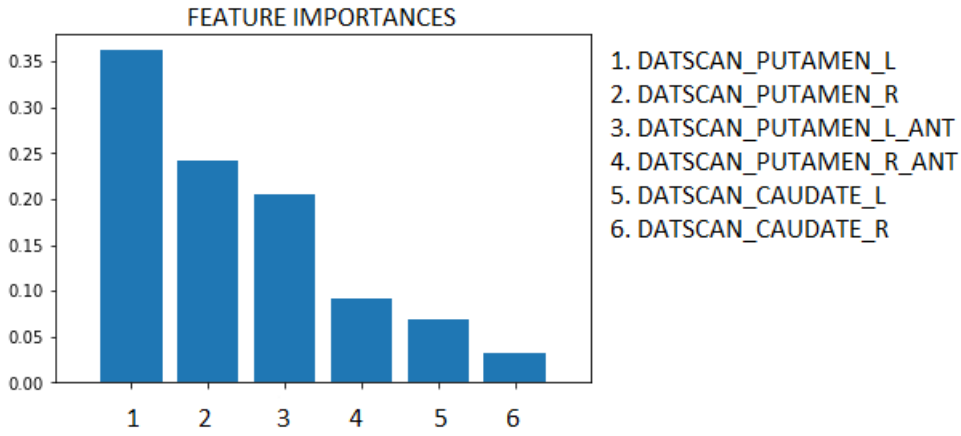

Figure 6: Feature Importance

Furthermore, the study highlights noteworthy differences in estimated disease stages between individuals diagnosed with Parkinson's disease (PD) and those who do not have the disease. These differences underscore the impact of PD on brain structure and function, providing insights into how the disease progresses over time. To enhance identification accuracy, researchers applied a Random Forest algorithm, a robust machine-learning technique that successfully distinguishes between patients with PD and healthy individuals. This approach demonstrates the potential for using advanced computational methods to improve diagnostic capabilities in neurological disorders.

V. CONCLUSION

This study presents a comprehensive approach for the automated detection of Parkinson's disease utilizing Striatal Binding Ratio (SBR) values derived from DaTSCAN SBR measurements. Our methodology analyzed SBR values from crucial brain regions—the caudate nucleus and putamen—significantly impacted by Parkinson's disease. We implemented a preprocessing pipeline featuring duplicate removal and the Isolation Forest algorithm for outlier detection, thereby enhancing the quality of the dataset. Among the brain biomarkers selected for this study, the left putamen exhibited abnormal development the earliest, followed by the right putamen, left caudate, and right caudate. The Random Forest classifier demonstrated an impressive accuracy of 97% in differentiating between Parkinson's disease patients and healthy controls, underscoring the robustness of our approach. The model

achieved an outstanding performance with a nearly 98% F1 score, reflecting an exceptional balance between precision (98.59%) and recall (96.89%). Our method serves as a valuable computer-aided diagnostic tool, assisting clinicians in making more accurate and objective diagnoses, particularly in the early stages of the disease when clinical symptoms may be subtle.

ACKNOWLEDGMENT

We would like to express our sincere gratitude to the PMMI organization for providing the valuable dataset that was essential for our research study. The availability of high-quality data significantly contributed to the success of our project and enabled us to develop an effective machine learning model for the detection of Parkinson's disease. We appreciate your support and commitment to advancing research in this critical area. Thank you for your invaluable contribution to our work.

REFERENCES

[1] Islam MA, Hasan Majumder MZ, Hussein MA, Hossain KM, Miah MS. A review of machine learning and deep learning algorithms for Parkinson's disease detection using handwriting and voice datasets. Heliyon. 2024 Feb 5;10(3):e25469. doi: 10.1016/j.heliyon. 2024.e25469. PMID: 38356538; PMCID: PMC10865258.

[2] Shirin Akbari, Mohammad Reza Deevband, Amin Asgharzadeh Alvar, Emadodin Fatemi Zadeh, Hashem Rafie Tabar, Patrick Kelley, Meysam Tavakoli, Brain network analysis in Parkinson's disease patients based on graph theory, Neuroscience Informatics, Volume 4, Issue 4, 2024, 100173, ISSN 2772-5286, https://doi.org/10.1016/j.neuri.2024.100173.

[3] Jinqiao Zhu, Yusha Cui, Junjiao Zhang, Rui Yan, Dongning Su, Dong Zhao, Anxin Wang, Tao Feng, Temporal trends in the prevalence of Parkinson's disease from 1980 to 2023: a systematic review and meta-analysis, The Lancet Healthy Longevity, Volume 5, Issue 7, 2024, Pages e464-e479, ISSN 2666-7568, https://doi.org/10.1016/S2666-7568(24)00094-1.

[4] S. Priyadharshini, K. Ramkumar, Subramaniyaswamy Vairavasundaram, K. Narasimhan, S. Venkatesh, Rengarajan Amirtharajan, Ketan Kotecha, A Comprehensive framework for Parkinson's disease diagnosis using explainable artificial intelligence empowered machine learning techniques, Alexandria Engineering Journal, Volume 107, 2024, Pages 568-582, ISSN 1110-0168, https://doi.org/10.1016/j.aej.2024.07.106.

[5] Wen-Sheng Huang, Shinn-Zong Lin, Jiann-Chyun Lin, Shiaw-Pyng Wey, Gann Ting, Ren-Shyan Liu "Evaluation of Early-Stage Parkinson's Disease with 99mTc-TRODAT-1 Imaging" Journal of Nuclear Medicine Sep 2001, 42 (9) 1303-1308;

[6] Abedelahi A, Hasanzadeh H, Hadizadeh H, Joghataie MT. Morphometric and volumetric study of caudate and putamen nuclei in normal individuals by MRI: Effect of normal aging, gender, and hemispheric differences. Pol J Radiol. 2013 Jul;78(3):7-14. doi: 10.12659/PJR.889364. PMID: 24115954; PMCID: PMC3789937.

[7] Dauer W, Przedborski S. Parkinson's disease: mechanisms and models. Neuron. 2003 Sep 11;39(6):889-909. doi: 10.1016/s0896-6273(03)00568-3. PMID: 12971891.

[8] Tsang, K., Walker, R. Dopamine transporter single photon emission computed tomography (DaT-SPECT) use in the diagnosis and clinical management of parkinsonism: an 8-year retrospective study. J Neurol 270, 2550–2558 (2023). https://doi.org/10.1007/s00415-023-11563-y

[9] Nandan, N., Pande, M.S., Raveesh, B.N. et al. Sensitive Two-Dimensional Photonic Crystal Biosensor for The Detection of Parkinson's Disease. J Opt (2024). https://doi.org/10.1007/s12596-024-02306-x

[10] Seifert KD, Wiener JI. The impact of DaTscan on the diagnosis and management of movement disorders: A retrospective study. Am J Neurodegener Dis. 2013;2(1):29-34. Epub 2013 Mar 8. PMID: 23515233; PMCID: PMC3601468.

[11] Arash Yaghoobi, Homa Seyedmirzaei, Marzie Jamaat, Moein Ala, The role of AI and machine learning in the diagnosis of Parkinson's disease and atypical parkinsonisms, Parkinsonism and Related Disorders, Volume 126, 2024, 106986, ISSN 1353-8020, https://doi.org/10.1016/j.parkreldis.2024.106986.

[12] Bega D, Kuo PH, Chalkidou A, Grzeda MT, Macmillan T, Brand C, Sheikh ZH, Antonini A. Clinical utility of DaTscan in patients with suspected Parkinsonian syndrome: a systematic review and meta-analysis. NPJ Parkinsons Dis. 2021 May 24;7(1):43. doi: 10.1038/s41531-021-00185-8. PMID: 34031400; PMCID: PMC8144619.

[13] Sourabarna Roy, Tannistha Pal, Swapan Debbarma, A Comparative Analysis of Advanced Machine Learning Algorithms to diagnose Parkinson's Disease, Procedia Computer Science, Volume 235, 2024, Pages 122-131, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2024.04.015.

[14] Chien CY, Hsu SW, Lee TL, Sung PS, Lin CC. Using Artificial Neural Network to Discriminate Parkinson's Disease from Other Parkinsonisms by Focusing on Putamen of Dopamine Transporter SPECT Images. Biomedicines. 2020 Dec 24;9(1):12. doi: 10.3390/biomedicines9010012. PMID: 33374377; PMCID: PMC7823797.

[15] Justin Quan, Lin Xu, Rene Xu, Tyrael Tong, and Jean Su, DaTscan SPECT Image Classification for Parkinson's Disease, ArXiv 2019, abs/1909.04142

[16] Kurmi A, Biswas S, Sen S, Sinitca A, Kaplun D, Sarkar R. An Ensemble of CNN Models for Parkinson's Disease Detection Using DaTscan Images. Diagnostics (Basel). 2022 May 8;12(5):1173. doi: 10.3390/diagnostics12051173. PMID: 35626328; PMCID: PMC9139649.

[17] Khachnaoui H, Khlifa N, Mabrouk R. Machine Learning for Early Parkinson's Disease Identification within SWEDD Group Using Clinical and DaTSCAN SPECT Imaging Features. J Imaging. 2022 Apr 2;8(4):97. doi: 10.3390/jimaging8040097. PMID: 35448224; PMCID: PMC9032319.

[18] Pavan Rajkumar Magesh, Richard Delwin Myloth, Rijo Jackson Tom, An Explainable Machine Learning Model for Early Detection of Parkinson's Disease using LIME on DaTSCAN Imagery, Computers in Biology and Medicine, Volume 126, 2020, 104041, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2020.104041.

[19] Martínez-Murcia, F.J., Górriz, J.M., Ramírez, J., Illán, I.A., Puntonet, C.G. (2013). Texture Features Based Detection of Parkinson's Disease on DaTSCAN Images. In: Ferrández Vicente, J.M., Álvarez Sánchez, J.R., de la Paz López, F., Toledo Moreo, F.J. (eds) Natural and Artificial Computation in Engineering and Medical Applications. IWINAC 2013. Lecture Notes in Computer Science, vol 7931. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-38622-0_28

[20] Majhi B, Kashyap A, Mohanty SS, Dash S, Mallik S, Li A, Zhao Z. An improved method for diagnosis of Parkinson's disease using deep learning models enhanced with metaheuristic algorithm. BMC Med Imaging. 2024 Jun 24;24(1):156. doi: 10.1186/s12880-024-01335-z. PMID: 38910241; PMCID: PMC11194992.

[21] Kavitha Paranjothi, Fathima Ghouse, Revathi Vaithiyanathan, Detection of Parkinson's Disease on DaTSCAN Image Using Multi-kernel Support Vector Machine, International Journal of Intelligent Engineering and Systems, Vol.17, No.5, 2024 DOI: 10.22266/ijies2024.1031.05

[22] Babita Majhi and Bhanu Prasad, Deep learning architectures for Parkinson's disease detection by using multi-modal features. Comput. Biol. Med. 146, C (Jul 2022). https://doi.org/10.1016/j.compbiomed.2022.105610

[23] Makarious, M.B., Leonard, H.L., Vitale, D. et al. Multi-modality machine learning predicting Parkinson's disease. npj Parkinsons Dis. 8, 35 (2022). https://doi.org/10.1038/s41531-022-00288-w

[24] Madhusudhana G K, Sanjaypande M B, Raveesh B N, Prediction of Parkinson's disease using the Parametric and Non-Parametric Machine Learning Models, December 2021 | IJIRT | Volume 8 Issue 7 | ISSN: 2349-6002

[25] Wuwang, Junho Lee, Fouzi Harrou And Ying Sun Early Detection of Parkinson's Disease Using Deep Learning and Machine Learning, August 12, 2020, IEEE Access Digital Object Identifier 10.1109/ACCESS.2020.3016062