**Predicting Parkinson's Disease Progression: A Hybrid Approach Using AI and Statistical Methods**

**ABSTRACT**

## **1**.**Introduction**

Parkinson’s is the second most prevalent neurodegenerative disorder first described almost two centuries ago, characterized by the progressive death of dopamine-producing brain cells among elderly people from 50 to 70 years old, especially in countries with elderly populations. A United Nations research claims that nearly 1 billion people worldwide, or approximately one in six, suffer from neurological conditions like epilepsy, migraine, brain injuries, and neuro-infections like Alzheimer's, Parkinson’s Disease, stroke, and multiple sclerosis. Each year, 6.8 million of these sufferers were passing away [12]. Although the actual cause of Parkinson's disease is unknown, it is believed that a combination of genetic, gender, age, and environmental factors is responsible for it [13]. In the modern world, PD affects 2–3% of people who are at the age of 65 and older [14]. Parkinson's disease clinical manifestations and progression differ from person to person, and it is impossible to anticipate how quickly the disease may progress in any specific person as PD is highly heterogeneous. While some people may have only minor symptoms for years, others may do so quite fast as they progress to more severe problems. There is no cure for PD and its treatments include mainly medications and surgery. Early detection and prognosis are crucial for assisting patients to retain a good quality of life. Therefore, developing techniques that can provide an accurate and trustworthy prediction of Parkinson’s in subjects is of major significance for a society that cares about people’s well-being.

PD affects predominately dopamine-producing dopaminergic neurons in the substantia nigra, which is a specific area of the brain [16]. Dopamine is an organic substance produced by neurons that serves as a neurotransmitter in the brain, facilitating communication between neurons. Parkinson's disease results from impaired neuronal communication due to insufficient dopamine production in the brain [15]. Typical PD symptoms include bradykinesia, rigidity, slowed movement, sleep disorders, posture imbalance, depression, and rest tremors, which affect speech, hand coordination, gait, and balance [17].

In order to diagnose PD, currently various neuroimaging modalities, such as single photon emission computed tomography (SPECT) [18, 19, 20], magnetic resonance imaging (MRI) [21], and position emission tomography (PET) methods are used [22]. In fact, these methodologies can accurately detect PD. However, these methodologies are based on the detection of substantial losses of dopaminergic neurons, for example in the substantia nigra. SPECT scans from PD subjects show an asymmetrical, smaller, circular, and less clear striatum region whereas normal subjects show two relatively symmetric comma or crescent-shaped focal regions. Traditionally SPECT images are interpreted manually in clinics where the diagnosis result may be subject to the risk of human error. Thus, Semi-quantitative analysis of striatal FPCIT uptake, measured by the specific binding ratio (SBR), is commonly used to supplement visual interpretation [23]. The SBR value gives neutral measures of dopaminergic function but the problem is, that this semi-quantitative measure is affected by the variability of SPECT image characteristics caused by differences in acquisition and reconstruction protocols and by scan-specific variations such as head motion and varying radius of rotation of the camera heads. Furthermore, SBR results cannot reflect the striatal uptake shape information and particular patterns [23].

Deep Learning (DL) has widely been used to diagnose various diseases and conditions, often with results exceeding standard benchmarks [14, 24, 25, 26, 27]. To overcome the aforementioned problem with SBR, Computer-Aided Diagnosis (CAD) based on DL approaches has been developed for automatic PD diagnosis in SPECT DaTSCAN images. In medical imaging, the better availability of reliable and public datasets with ease of access in online platforms, and the advancement in computational processing power and storage have helped to increase the use of DL methods, especially Convolutional Neural Networks (CNN) [28]. CNN models have shown tremendous potential in the analysis and interpretation of medical images for diagnosing various diseases, including PD [29]. Through the use of DL, we can efficiently and accurately classify patients as to whether they have PD or not by detecting patterns in their SPECT scans, mainly around the putamen and caudate regions as they are relatively different compared to non-PD specimens.

## **2.Related Works**

Over the years, researchers worldwide have published multiple studies on Parkinson’s disease. To understand Parkinson’s disease progression, we first have to separate the PD patients from the healthy patients (HC). After identifying the markers, we can determine which features to keep an eye on and keep monitoring them. Researchers have used numerous AI methods to analyze PD classification, early detection, and progression. Amoroso, Nicola, et al. [1] proposed a novel approach where magnetic resonance imaging (MRI) scans were processed to obtain a network representation and with the help of Random Forest Feature Selection and Support Vector Machine accurate early diagnosis of PD was done. Khachnaoui, Hajer, et al. [2] proposed a Computer-Aided Diagnosis (CAD) system for automatic PD diagnosis using SPECT images, pre-trained CNN models, the Transfer Learning (TL) technique, and the Bilinear Pooling method. Junaid, Muhammad, et al. [3] proposed an ML pipeline based on time series data for predicting three-class and four-class based progression of PD that is both accurate and explainable. In 2019, Yagis, Ekin, Alba G. Seco De Herrera, and Luca Citi [4] used MRI images to assess the performance of the state-of-the-art CNN models on the classification of PD. Prashanth, R., et al. [5] in their research using ML models and the PPMI database suggested that a combination of non-motor, Cerebrospinal Fluid (CSF), and imaging markers may aid in the preclinical diagnosis of PD. In a research from 2022, Harvey, Joshua, et al. [6] tested the prediction of two cognitive outcome measures in newly diagnosed PD subjects within 8 years, using multiple variable subsets and ML algorithms which led to a discovery that combining both biological and clinical variables produced best-performing models, with a marginal improvement in predictive performance compared to models using clinical variables alone. Zhang, Xiaobo, et al. [7] recently proposed a sparsity-based multi-level graph learning framework for PD prediction which incorporates a fast graph construction method and works well on smaller datasets. Zhang, Xi, et al. [8] using the PPMI database also proposed a multi-view graph convolutional network method called MVGCN, which can directly take brain graphs from multiple views as inputs and do predictions on that. For PD diagnosis, Huang, Zhongwei, et al. [9] performed classification and clinical score prediction jointly by also proposing a novel adaptive unsupervised feature selection approach that exploits manifold learning from longitudinal multimodal data. Another recent paper from 2023 by Hajianfar, Ghasem, et al. [10] demonstrated that combining medical information with SPECT-based imaging features, and optimal utilization of Hybrid Machine Learning Systems (HMLS), can produce an excellent prediction of the presence of pathogenic gene variants in PD patients.

## **3.Methodology**

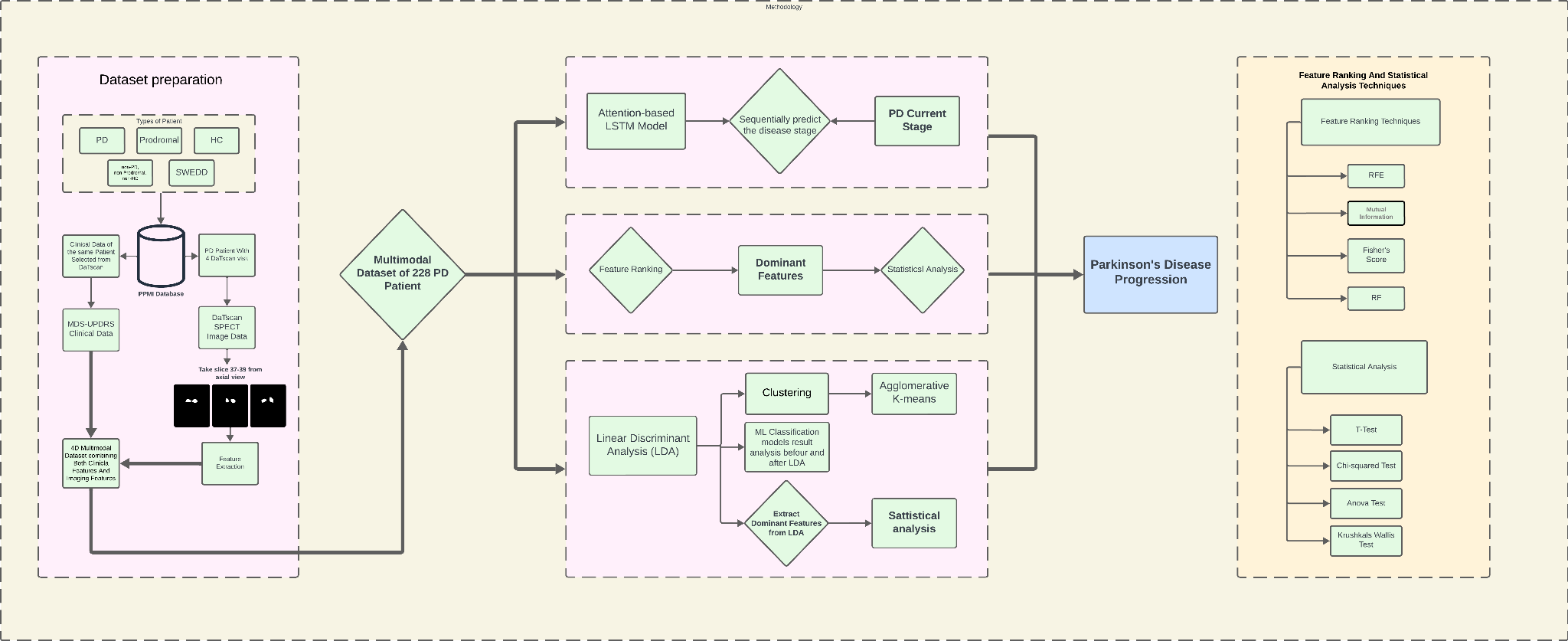
****

Figure 1 :Our detailed working pipeline

### *3.1 Dataset Description*

The Michael J. Fox Foundation and a core group of academic scientists and industry partners launched the Parkinson’s Progression Markers Initiative (PPMI) [30] to continue the study of PD biological markers and their progression. PPMI contains three main collections of data: 1. PPMI Clinical, 2. PPMI Remote, and 3. PPMI Online. In our study, we used Motor Assessments (MDS-UPDRS) clinical and Dopamine Transporter scans (DaTscan SPECT) image data from PPMI Clinical collection. The MDS-UPDRS data are divided into five distinct categories: 1. MDS-UPDRS Part I Patient Questionnaire, 2. MDS-UPDRS Part I Non-motor Aspects of Daily Living, 3. MDS-UPDRS Part II Patient Questionnaire: Motor Aspects of Daily Living, 4. MDS-UPDRS Part III Treatment Determination and Part III: Motor Examination, and 5. MDS-UPDRS Part IV Motor Complications. The data files we used had 2722 unique patient samples till April 3, 2024:

Table 1 : Dataset Demographic

|  |  |
| --- | --- |
| **Patient Type** | **No. of Patients** |
| Prodromal | 1233 |
| Parkinson's Disease | 1227 |
| Healthy Control | 249 |
| non-PD, non-Prodromal, non-HC (participants to be excluded) | 10 |
| SWEDD | 3 |

DaTscan SPECT images were acquired at PPMI imaging centres per the PPMI imaging protocol and sent to the imaging core lab at the Institute for Neurodegenerative Disorders (IND) for visual interpretation. The raw DaTscan SPECT images taken at PPMI-affiliated medical clinics had undergone some preprocessing which information can be found at <https://www.ppmi-info.org/>. The images were in DICOM (Digital Imaging and Communication in Medicine) [31] file format which is used for medical images. A collection of image-related metadata, including patient number (PAT NO), when the image was collected (EVENT\_ID, INFODT, ORIG\_ENTRY), scan type (MRI/DaTScan/PET), and other decimal data is included in DICOM files. The problem is, that it is difficult to use DICOM file format images when working with ML related tasks. Most of the ML libraries and frameworks do not support DICOM files, which is why these files need to be converted into PNG or JPG file format images before being used for image analysis.

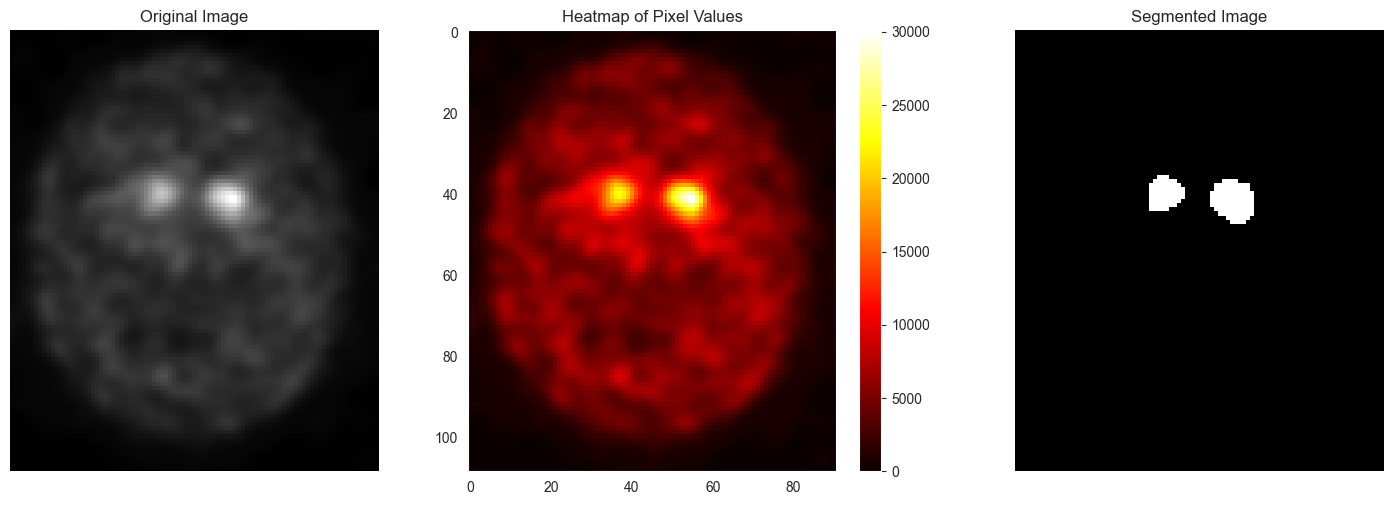
****

Figure 2 :SPECT DaTscan with putamen and caudate regions with its segmented view

### *3.2 Preprocessing*

In our study, we incorporated both the MDS-UPDRS data as well as DaTScan SPECT image data and combined them. For the tabular data, from five distinct MDS-UPDRS categories, we used four of them because MDS-UPDRS Part IV had a lot of missing data for the number of patients we were working with. Furthermore, out of five distinct categories of patients we only used PD (Parkinson’s Disease) subjects for our study. We selected 228 PD patients because from our analysis these selected patients contained all types of motor as well as clinical image data.

Some useful features from the selected MDS-UPDRS categories we used are given below:

|  |  |
| --- | --- |
| **Category** | **Features** |
| MDS-UPDRS Part I Patient Questionnaire | NP1COG, NP1HALL, NP1DPRS, NP1ANXS, NP1APAT, NP1DDS, |
| MDS-UPDRS Part I Non-motor Aspects of Daily Living | NP1CNST, NP1FATG, NP1LTHD, NP1PAIN, NP1SLPD, NP1SLPN, NP1URIN |
| MDS-UPDRS Part II Patient Questionnaire: Motor Aspects of Daily Living | NP2DRES, NP2EAT, NP2FREZ, NP2HOBB, NP2HWRT, NP2HYGN, NP2RISE, NP2SALV, NP2SPCH, NP2SWAL, NP2TRMR, NP2TURN, NP2WALK |
| MDS-UPDRS Part III Treatment Determination and Part III: Motor Examination | NP3BRADY, NP3FACXP, NP3FRZGT, NP3FTAPL, NP3FTAPR, NP3GAIT, NP3HMOVL, NP3HMOVR, NP3KTRML, NP3KTRMR, NP3LGAGL, NP3LGAGR, NP3POSTR, NP3PRSPL, NP3PRSPR, NP3PSTBL, NP3PTRML, NP3PTRMR, NP3RIGLL, NP3RIGLU, NP3RIGN, NP3RIGRL, NP3RIGRU, NP3RISNG, NP3RTALJ, NP3RTALL, NP3RTALU, NP3RTARL, NP3RTARU, NP3RTCON, NP3SPCH, NP3TTAPL, NP3TTAPR |

A detailed data dictionary and code list can be found on the <https://www.ppmi-info.org/> website.

On the image data, one of our key focuses was on analyzing the deep brain structures responsible for motor control specifically the caudate nucleus and putamen which together make up the striatum. These structures are critical for regulating movement and are particularly vulnerable to dopamine depletion, a hallmark of Parkinson’s disease. For the segmentation, we tried various algorithms like preprocessing-based segmentation and pixel-by-pixel analysis but both failed to extract the ROI because we can not generalize the algorithm for all DatScan’s. The segmentation was done by the K-Means clustering algorithm and we successfully extracted the Striatum from the axial slice of DaTscan. For each DaTscan we segment the axial slice no 37 to 39 and for each segmented slice, we extract some features. The extracted features are shown in [Table 1](#D2L_table_ref_Extracted feature from the segmented DaTscan.).

Table 2 :Extracted feature from the segmented DaTscan

|  |  |  |
| --- | --- | --- |
| **Features** | **Equation** | **Description** |
| Area |  | The area represents the total number of pixels in the segmented striatum, giving a quantitative measure of the region size. |
| PA Ratio |  | PA ratio measures the elongation of the segmented striatum. It indicates the structural change of a patient over time. |
| Solidity |  | Solidity is the ratio of the Area of the segmented striatum and the Convex hall area. This feature helps evaluate the integrity of the striatum structure. |
| Circularity |  | It measures how close the shape of the segmented striatum is to a perfect circle. It provides information on the regularity of the striatum's shape. |
| Equivalent Diameter |  | With this feature, we can extract the diameter of the striatum if the striatum is a circle. |
| Extent |  | It is the ratio of area and the bounding box of the striatum which may help to assess how compact or spread out the region is. |
| Mean Intensity |  | It is measured by the average pixel intensity within the striatum and may help to determine dopamine transporter activity. |
| Standard Deviation |  | This measures the variation in pixel intensities within the striatum. Higher values indicate more variability. |
| Skewness |  | Skewness is a statistical measure that describes the asymmetry of the distribution of pixel intensity values in the segmented region. |
| Kurtosis |  | Kurtosis measures the "peakedness" or "tailedness" of the intensity distribution. Higher kurtosis indicates more extreme intensity values, while lower kurtosis reflects a flatter distribution. |
| Shannon Entropy |  | We can measure the randomness of the intensity distribution in the striatum by Shannon Entropy. Higher entropy indicates greater complexity or heterogeneity in the region. |
| LBP Energy |  | The LBP Energy can quantify the uniformity of the texture in the striatum region. High energy indicates more regular or uniform texture patterns. |
| LBP Entropy |  | LBP entropy measures the complexity of the texture of segmented regions based on local binary patterns. Higher entropy suggests a more complex or irregular texture. |
| Gabor Energy |  | To capture the strength of texture patterns at specific frequencies and orientations of the striatum region, Gabor energy is used. High energy indicates strong texture patterns, while low energy suggests weaker or less defined textures. |
| Gabor Entropy |  | Randomness in texture pattern is extracted by Gabor entropy. Higher entropy suggests more complex or disordered textures. |
| Correlation | Correlation = | Correlation measures the linear relationship between neighboring pixel intensities in the segmented region. A high correlation indicates similar intensity values, while a low correlation suggests more variation |
| Dissimilarity |  | Dissimilarity measures how the pixel intensity varies between the naibours in the segmented region. |
| Homogeneity |  | Homogeneity measures the similarity of pixel intensities across the region. A higher homogeneity score suggests that the pixel values are more uniform, indicating a structurally intact and consistent region. |
| Contrast |  | It measures the intensity difference between the pixels in the striatum. |
| Energy |  | To measure if the intensity of pixels is uniform or not across the striatum, we could determine it by calculating energy. High energy indicates that the intensity values are consistent throughout the striatum. |
| Filled Area |  | The filled area is the complete area of the segmented striatum after filling in any gaps or holes. |
| Major & Minor Axis Length | if | Major and Minor axis length is the longest and shortest diameter of the ellipse that best fits the shape of the segmented striatum. It captures the most and least extended direction of the region. |

To understand the progression of PD patients we choose the patients with four periodical DaTscan visits: 1. Baseline, 2. Month 12, 3. Month 24, 4. Month 48. Similar to the tabular data, our previously selected 228 patients’s SPECT DaTscan was separated from the total 4433 scans. Each patient has 4 DaTscan imaging so the total number of DaTscan we are working on is 912. For each DaTscan we segment the striatum extract the feature and create a dataset. Both clinical assessment and extracted feature dataset were combined carefully based on the assessment date, event ID, and patient ID by precise human evaluation. In the 4D dataset, there were some missing values in the clinical analysis column and we fill the missing value with the mode for convenience of analysis.

### *3.3 Feature Selection and Analysis*

Feature selection is an important part of the machine learning process, as it lets us identify the most relevant and informative features from a dataset. In our study, we deployed four distinct feature selection techniques: Random Forest Feature Importance, Mutual Information Score, Fisher Score, and Recursive Feature Elimination. Each of these methods determines the significance of each feature by evaluating its impact on the model’s performance.

**Random Forest Feature Extraction (RFFE)** is a technique that uses a random forest algorithm which is an ensemble learning method that combines multiple decision trees to make a more accurate model to assess the importance of each feature. Feature importance in Random Forests is measured by how much each feature decreases the impurity (e.g., Gini impurity or entropy) across all the trees in the forest.

The Gini Impurity for a node is calculated as:

The feature importance FI of feature j is the average decrease in Gini impurity across all trees T in the forest:

Here is the feature importance is the total number of trees in the Random Forest, is the set of all nodes in the tree where is the features is used for splitting

**Mutual Information** measures how much the feature reduce uncertainty about the target class and it captures any relationship between features and the class. Features with higher MI scores are considered more important. It quantifies the amount of shared info between a feature and class. For our case, to stabilize the feature selection we average the score over multiple runs and selects the features with the highest average MI. The Mutual information score is calculated using this equation:

Here is the joint probability of and . and are the marginal probabilities of and .

**Fisher Score** is another feature selection technique that assesses the biased ability of features, based on the class distribution within a dataset. It takes the amount of ratio of the variance among the classes to the variance within classes. Features with a higher Fisher score are more likely to be relevant. Fisher Score is calculated using this equation:

Here is the number of classes and is the number or samples in class . is the mean of feature ​ for class . is the overall mean of feature. is the variance of a feature within class *k.*

**Recursive Feature Elimination (RFE)** is a wrapper-type feature selection algorithm. The goal of RFE is to select important features by recursively considering smaller and smaller sets of features discarding the least important ones and refitting the model. Features are scored by either any given ML model or by using a statistical method. In our case we use Logistic regression for feature selection using Recursive feature Elimination. For our case the RFE calculated by this equation:

Here is the feature value is the cofficient of and is the intercept and is the predicted probability. In RFE features with smaller is eliminated because they contribute less in the prediction of .

Beside these ML-based feature selection, we also use LDA (Linear Discriminant Analysis) to select important features. It is a supervised algorithm which is used for dimensionality reduction. It maximize the between-class variance and minimize the within-class variance and project the higher dimensional data into lower dimension. In feature selection it identify the most important features that maximize the separation between classes. The higher the score the more important the feature is for LDA. Linear Discriminant Analysis gives priority mostly to the image extracted feature which we can see in ***Table 3*** .On the other hand, feature extraction technique like RF, RFE, Mutual information, Fisher score gives importance mostly in the Clinical features ***Table 4.***

### *3.4 Dimensionality Reduction & Clustering*

This study used **Linear** **Discriminant** **Analysis** **(LDA)** as a dimensionality reduction technique to transfer the high-dimensional feature space into a lower-dimensional space that optimizes class separability. LDA is a supervised method that aims to project the data onto a subspace where the classes are most distinguishable.

LDA finds the linear combination of the original feature set which maximizes the ratio of between-class variance to within-class variance which tries to ensure that the projected data is as spread out as possible between classes while minimizing the spread within each class.

The following formula calculates the LDA:

The main objective is to find a transformation matrix that maximizes the ratio of the between-class scatter matrix to the within-class scatter matrix .

Here is the within-class scatter matrix. captures the variance within each class by summing the covariance of data points around their respective class mean.

where ​ is a data point in class , and ​ is the mean vector of class and is the number of classes.

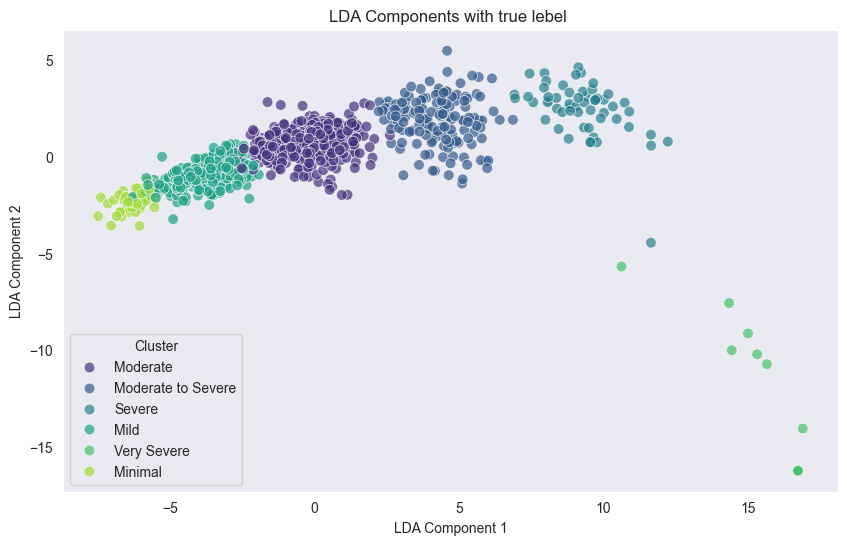
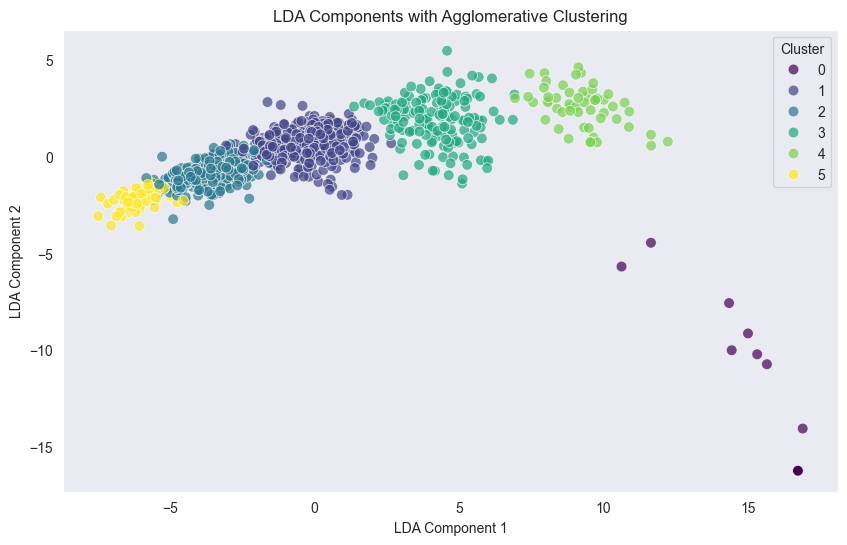
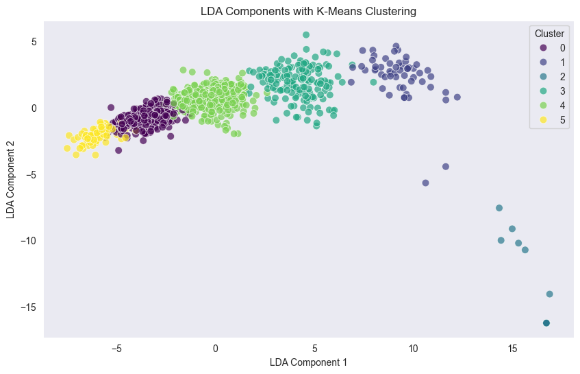
is the **between-class scatter matrix**, which measures the variance between class means:

where ​ is the mean vector of class , and is the overall mean vector of the dataset. ​ is the number of samples in class .

Once the class means and scatter matrix are calculated the next step in LDA involves finding the eigenvectors and eigenvalues of the matrix . These eigenvectors represent the direction in the feature space along with the data exhibits the most separation between classes.

In this equation is the matrix of eigenvectors (the LDA coefficients or scalings), is the diagonal matrix of eigenvalues. The eigenvectors derived from this eigenvalue problem represent the optimal directions of projection that maximize class separability.

Figure 3 :Clustering of the LDA components and compare the clusters with true disease severity levels.



A graph with a bar and a bar chart

Description automatically generated with medium confidence

We transfer the higher-dimensional dataset with 84 features to 2 linear discriminants, which cover the 96% variance of the total dataset in *Figure 4*. After transforming the higher-dimensional dataset into the lower-dimensional dataset, we cluster the LD(linear discriminant) using Agglomerative clustering and K-means clustering techniques to ensure that the LDs properly create distinct clusters among different severity levels on *Figure 3*.

Figure 4: Explained variance Of LD's

### *3.5 AI Models*

To predict the disease severity level sequentially we build an Attention-based LSTM model to handle the multimodal dataset which integrates data from both clinical assessment and image(DaTscan extracted feature) data. The clinical features include various symptomatic measurements which are recorded over multiple visits. Clinical data captures the dynamic of the disease as experienced by individual patients. In parallel the image-extracted features give quantitative metrics derived from the medical imaging. Combining both medical and imaging features enhances the predictive capabilities for assessing disease severity and enables a more holistic approach to patient evaluation and treatment planning. To train the model, a sequence was generated from the multimodal dataset. Each patient's data was organized into a sequence based on their visits. This strategy created structured input formats that are suitable for our model architecture. Each input sequence includes 4 visit data of a patient. This type of input sequence of visits enables the model to learn patterns over time about the disease severity. This approach

A blue and white grid with a blue square

Description automatically generated

ensures that the model can effectively analyze how symptoms evolve and correlate with imaging biomarkers, ultimately enhancing the ability to make informed clinical predictions.

**Input Layer**: The input layer accepts sequences of clinical and imaging features. The input shape is defined as a 3D tensor representing the number of time steps and the number of features (clinical and imaging features combined). This layer allows the model to process sequential data for each patient.

**First LSTM Layer:** The first layer in the model is an LSTM layer. This layer processes the input sequence and maintains temporal information, and dependencies across time steps. A **Batch Normalization** layer is also added after the first LSTM to standardize the activations and stabilize learning which prevents internal covariate shifts. A **Dropout** layer is also added after the LSTM layer to regularize the model and mitigate overfitting by randomly dropping 50% of the units during training, ensuring the model learns robust patterns.

**Bidirectional LSTM Layer**: The second layer of the model is a Bidirectional LSTM. It processes the sequence both forward and backward, the bidirectional LSTM captures temporal dependencies from both past and future time steps which allows the model to gain a more holistic understanding of the patient data. This layer is again followed by **Batch Normalization** and a **Dropout** layer to normalize the outputs which also improve convergence and generalization and reduce overfitting.

**Third LSTM Layer**: A third LSTM layer with the same units is added. This layer provides additional temporal modeling capabilities. The third layer allows the model to process even more complex patterns in the sequential data. As in the previous layers, **Batch Normalization** and **Dropout** layers are included to regularize the learning process and ensure stability.

**Attention Layer**: After the LSTM layers, an **Attention Layer** is applied. This attention mechanism enhances the model by focusing on the most relevant time steps in the sequence. The attention mechanism learns a set of weights, which are applied to the hidden states produced by the LSTM layers.

**Output Layer**: The final layer is a **Dense layer** with a softmax activation function, which gives a probability distribution over the possible classes (disease severity levels in this case). The number of units in this layer corresponds to the number of classes in the target variable, with the softmax activation ensuring that the model outputs valid probability values. A regularizer (l2) is applied to the weights of this layer to control for model complexity and reduce overfitting by penalizing large weight values.

To optimize the performance of the model to reduce the overfitting and underfitting possibilities and find the best hyperparameter use Grid Search. Grid search is a method that systematically evaluates a range of hyperparameter combinations to identify the most effective setting of the model. By using grid search we tune parameters such as the number of LSTM units, dropout rate, L2 regularization, and batch size. We split the total dataset into training and testing sets of 70% and 30%. Besides the hyperparameter tuning, we also applied Stratified K-Fold cross-validation. In this technique, the data is split into several folds (in this case, three), ensuring that each fold maintains the same proportion of target class labels, which is essential for balanced learning. The model is trained on a subset of the data and validated on the remaining fold, iterating over all splits to ensure comprehensive evaluation.

Besides the Attention-based LSTM model, we also use the Linear discriminant (LD1, LD2) to predict the disease severity using various classification models like Random Forest, Logistic Regression, AdaBoost, Decision Tree, Gradient Boosting, KNN, SVM, Naïve Bayes, XGB. We compare the results of these classification models between the original multimodal dataset and the linear discriminant of the original multimodal dataset in ***Table 6*** and ***Table 7***.

### *3.6 Statistical Methods*

Statistical analysis has a fundamental role in machine learning. It enables us to extract useful insights and meaning from large datasets, identify patterns, and make informed decisions. It can help to identify correlations and relationships between features which is essential for feature selection.

In our study, we performed a Chi-squared test, T-test, Kruskal-Wallis test and ANOVA test on our ranked feature to evaluate the correlations between features and disease severity.

**Chi-squared test** is a statistical test used to determine whether there is any significant association between the feature and the severity. It is only used for clinical features, extracted from feature ranking technique. Chi-squared is calculated by:

Here is the observed frequency in each category and E is the expected frequency if there is no association between the variables. The p-value is computed from the Chi-squared distribution. If the p-value is less than 0.05, we reject the null hypothesis and conclude that the feature is significantly associated with the severity class.

T-test is also a statistical hypothesis test used to determine if there is a significant difference between the meaning of two groups or not.

Analysis of Variance (ANOVA) is a statistical test used to identify difference between the means of more than two groups. ANOVA is a well-known statistical approach to rank features by calculating the ratio of variances between and within groups. [34]

The Kruskal-Wallis [32] test is a non-parametric statistical test used to compare two or more groups to see if there are any significant differences. It extends the Mann-Whitney U test and an alternative to the one-way ANOVA. Kruskal-Wallis method is computationally less expensive and very simple in use [33]. It tests if two or more classes have equal median gives the value of *P*. Features that have discriminative information are selected. If the value of P is close to “0” it means that the feature contains discriminative information; otherwise, it will not be selected.

## **4. Result Analysis**

From our analysis, we found these 20 features most important features based on each feature

ranking technique:

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature Selection Using Various Algorithms** | | | |
| **RFFE** | **RFE** | **Fisher Score** | **MI Score** |
| NP3RIGN | NP1ANXS | NP2DRES | NP2DRES |
| NP2HOBB | NP1PAIN | NP2HOBB | NP2HOBB |
| NP2DRES | NP1SLPD | NP2RISE | NP3BRADY |
| NP3BRADY | NP2DRES | NP2EAT | NP2EAT |
| NP3POSTR | NP2EAT | NP2TURN | NP2RISE |
| Circularity | NP2HOBB | NP3BRADY | NP2TURN |
| Solidity | NP2RISE | NP3RISNG | NP3HMOVL |
| NP3TTAPL | NP2SPCH | NP2FREZ | NP3TTAPL |
| NP2EAT | NP2WALK | NP3LGAGL | NP3TTAPR |
| Homogeneity | NP3FACXP | NP3HMOVL | NP3FACXP |
| Extent | NP3FTAPL | NP3TTAPR | NP3RISNG |
| LBP energy | NP3FTAPR | NP3LGAGR | NP3LGAGR |
| Gabor Energy | NP3LGAGR | NP3POSTR | NP3RIGN |
| Shannon Entropy | NP3POSTR | NP3TTAPL | NP3POSTR |
| Minor axis length | NP3PRSPL | NP3FACXP | NP3LGAGL |
| Major axis length | NP3PTRML | NP2WALK | NP2WALK |
| NP2RISE | NP3RIGRU | NP3RIGN | NP3FTAPL |
| NP3HMOVR | NP3RISNG | NP2SPCH | NP3SPCH |
| Convex Area | NP3RTCON | NP2HYGN | NP3FTAPR |
| correlation | NP3TTAPL | NP3SPCH | NP2SPCH |

Table 4: Feature importance by ML models

Using Linear Discriminant Analysis (LDA) we extracted 20 features that are important:

|  |  |  |  |
| --- | --- | --- | --- |
| **Extracted features and their scaling values from Linear Discriminant** | | | |
| **Top 20 Feature LD1** | **Scaling** | **Top 20 Feature LD2** | **Scaling** |
| Standard Deviation | -13.866 | Area | 122.3075 |
| Contrast | 13.61688 | Filled Area | -118.63 |
| Gabor Entropy | -12.2778 | homogeneity | 44.79036 |
| Shannon Entropy | -11.0716 | Standard Deviation | -41.9018 |
| dissimilarity | -10.2457 | Skewness | -35.7697 |
| Skewness | -8.23307 | Shannon Entropy | 30.08729 |
| homogeneity | -8.00484 | Contrast | 20.99671 |
| energy | -7.88104 | Energy | -20.3214 |
| Gabor Energy | 6.730667 | Kurtosis | 18.43916 |
| Mean | 6.301926 | Dissimilarity | -13.4105 |
| brightness | 6.301926 | Gabor Entropy | -8.93315 |
| Area | 5.5528 | LBP Energy | 8.501743 |
| Filled Area | -4.59068 | Mean | 7.296339 |
| Equivalent Diameter | -4.10329 | Brightness | 7.296339 |
| Major axis length | 3.709302 | LBP Entropy | 5.47781 |
| Kurtosis | 3.561322 | correlation | 3.45925 |
| Convex Area | -3.49746 | Equivalent Diameter | -3.39241 |
| Minor axis length | 2.912351 | Gabor Energy | 2.822265 |
| Correlation | 2.476936 | Solidity | 1.855058 |
| LBP Entropy | 1.915438 | Convex Area | 1.625698 |

*Table 3 : Top Priority features from LDA*

The result of our ANOVA test and Kruskal-Wallis tests are given below:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Image Features** | **ANOVA across severity** | | **ANOVA across Visit** | | **Kruskal-Wallis across severity** | | **Kruskal-Wallis across Visit** | |
|  | **F-statistic** | **p-value** | **F-statistic** | **p-value** | **H-statistic** | **p-value** | **H-statistic** | **p-value** |
| Shannon Entropy | 3.289974074 | 0.005964296 | 1.527026121 | 0.205950773 | 11.78443718 | 0.0378639 | 0.074156095 | 0.994747115 |
| LBP Energy | 2.847547808 | 0.014681947 | 1.921816976 | 0.124406595 | 10.89514366 | 0.053499071 | 1.387665089 | 0.708428633 |
| Gabor Energy | 7.441404863 | 7.46E-07 | 4.535383437 | 0.003649503 | 39.55053814 | 1.84E-07 | 23.06476811 | 3.91E-05 |
| Convex Area | 1.071937415 | 0.374364463 | 1.750757824 | 0.155041946 | 7.462174118 | 0.188473428 | 7.280892084 | 0.063463765 |
| energy | 4.707035829 | 0.000299311 | 4.693889944 | 0.002930752 | 16.96078171 | 0.004574779 | 15.4399323 | 0.0014768 |
| Kurtosis | 3.545991964 | 0.003510074 | 7.968742104 | 3.01E-05 | 18.47530543 | 0.002406081 | 25.33881993 | 1.31E-05 |
| brightness | 5.233854357 | 9.58E-05 | 5.6215073 | 0.000807286 | 19.91797015 | 0.001294807 | 26.7212639 | 6.74E-06 |
| correlation | 0.863529681 | 0.50507701 | 0.651475088 | 0.582134977 | 2.991536003 | 0.701290815 | 2.459020688 | 0.482742504 |
| Skewness | 4.132351469 | 0.001022618 | 8.020340797 | 2.80E-05 | 18.95991413 | 0.001955463 | 25.5990405 | 1.16E-05 |
| Major axis length | 2.086410312 | 0.064919518 | 1.128158509 | 0.336656711 | 12.63923869 | 0.027004539 | 3.620180568 | 0.305506337 |
| LBP Entropy | 2.598806168 | 0.024110622 | 1.459353537 | 0.224188638 | 10.76578505 | 0.056226698 | 0.931205025 | 0.817891528 |
| Standard Deviation | 5.14445779 | 0.000116315 | 7.048220707 | 0.000109711 | 19.91686404 | 0.001295426 | 27.05414588 | 5.74E-06 |
| homogeneity | 2.744922171 | 0.018034641 | 1.306271988 | 0.271064955 | 9.221286097 | 0.100556714 | 0.110750923 | 0.990516776 |
| Solidity | 0.764748836 | 0.575316397 | 0.787435755 | 0.501022752 | 2.194524342 | 0.821626499 | 1.69974854 | 0.636989703 |
| Mean | 5.233854357 | 9.58E-05 | 5.6215073 | 0.000807286 | 19.91797015 | 0.001294807 | 26.7212639 | 6.74E-06 |
| Filled Area | 1.342792594 | 0.243902877 | 2.249562087 | 0.081112113 | 8.404752537 | 0.135294674 | 9.188358017 | 0.026888608 |
| dissimilarity | 3.442828684 | 0.004349042 | 1.672587287 | 0.171307108 | 8.571812754 | 0.127411035 | 0.905498145 | 0.824100781 |
| Equivalent Diameter | 2.366983964 | 0.037963768 | 1.616276255 | 0.184005289 | 9.945099422 | 0.076805859 | 5.571379316 | 0.134431133 |
| Minor axis length | 0.648135323 | 0.663001767 | 1.844248253 | 0.137507148 | 1.924706723 | 0.859462155 | 8.227191452 | 0.04154244 |
| Gabor Entropy | 7.183546583 | 1.32E-06 | 5.180127452 | 0.001492587 | 22.99411312 | 0.000338442 | 20.1586069 | 0.000157367 |
| Area | 1.338373272 | 0.245685833 | 2.229822289 | 0.083244854 | 8.350043092 | 0.137970084 | 9.126199934 | 0.027659204 |
| contrast | 4.367601904 | 0.000619691 | 1.963547194 | 0.11786056 | 8.628739142 | 0.124820736 | 2.340446889 | 0.50481667 |
| Circularity | 1.427440354 | 0.211796445 | 2.703490192 | 0.044408847 | 15.07600847 | 0.010042455 | 4.144333355 | 0.246294602 |
| Extent | 1.830878704 | 0.104237948 | 0.52583289 | 0.664616696 | 7.783072536 | 0.168603119 | 1.776080037 | 0.62015397 |
| **Clinical Features** | **ANOVA across severity** | | **ANOVA across Visit** | | **Kruskal-Wallis across severity** | | **Kruskal-Wallis across Visit** | |
|  | **F-statistic** | **p-value** | **F-statistic** | **p-value** | **H-statistic** | **p-value** | **H-statistic** | **p-value** |
| NP1ANXS | 14.66915601 | 7.53E-14 | 1.891076419 | 0.129449303 | 56.98091138 | 5.10E-11 | 4.597999463 | 0.203713765 |
| NP1PAIN | 33.28530686 | 3.01E-31 | 7.737549164 | 4.17E-05 | 133.0741805 | 5.30E-27 | 24.37930203 | 2.08E-05 |
| NP1SLPD | 29.64708981 | 6.05E-28 | 17.42016345 | 5.29E-11 | 121.6237857 | 1.42E-24 | 50.03663027 | 7.85E-11 |
| NP2DRES | 107.6048471 | 3.26E-89 | 14.98397538 | 1.59E-09 | 313.6933923 | 1.14E-65 | 40.80454405 | 7.19E-09 |
| NP2EAT | 71.61071544 | 3.34E-63 | 12.81939664 | 3.29E-08 | 235.5715422 | 6.84E-49 | 33.82252909 | 2.16E-07 |
| NP2FREZ | 67.11653189 | 1.05E-59 | 8.725055037 | 1.04E-05 | 163.133199 | 2.13E-33 | 27.13507105 | 5.52E-06 |
| NP2HOBB | 88.40272036 | 9.26E-76 | 10.58218664 | 7.64E-07 | 276.1656275 | 1.33E-57 | 31.91000002 | 5.47E-07 |
| NP2HYGN | 45.09639092 | 1.24E-41 | 3.1432549 | 0.024572269 | 178.4545547 | 1.14E-36 | 9.898854511 | 0.019445768 |
| NP2RISE | 82.65157076 | 1.51E-71 | 12.9206463 | 2.86E-08 | 245.6119388 | 4.80E-51 | 32.5640813 | 3.98E-07 |
| NP2SPCH | 53.0654886 | 2.33E-48 | 7.790092767 | 3.87E-05 | 178.5776265 | 1.08E-36 | 19.1881139 | 0.000249973 |
| NP2TURN | 61.34141276 | 4.06E-55 | 19.8184343 | 1.89E-12 | 201.0969312 | 1.65E-41 | 51.68881977 | 3.49E-11 |
| NP2WALK | 62.24744848 | 7.62E-56 | 11.94124048 | 1.13E-07 | 194.6910009 | 3.88E-40 | 33.02202914 | 3.19E-07 |
| NP3BRADY | 77.30348448 | 1.51E-67 | 3.117419799 | 0.025446276 | 267.6506211 | 8.94E-56 | 9.190440553 | 0.026863158 |
| NP3FACXP | 55.25893067 | 3.58E-50 | 1.809167613 | 0.143853812 | 205.4643659 | 1.92E-42 | 4.598460756 | 0.203674166 |
| NP3FTAPL | 52.21135205 | 1.20E-47 | 3.887895952 | 0.008904255 | 197.4223313 | 1.01E-40 | 10.91493296 | 0.012194872 |
| NP3FTAPR | 43.85774228 | 1.44E-40 | 1.1334091 | 0.334537606 | 170.8219648 | 4.87E-35 | 3.106288339 | 0.375526148 |
| NP3HMOVL | 58.52892343 | 7.59E-53 | 4.510184352 | 0.003778882 | 215.4140816 | 1.43E-44 | 13.85043999 | 0.003115962 |
| NP3HMOVR | 46.02725364 | 1.98E-42 | 1.920675864 | 0.124590385 | 179.4267512 | 7.09E-37 | 6.702502163 | 0.082009456 |
| NP3LGAGL | 61.407269 | 3.60E-55 | 10.66745181 | 6.77E-07 | 189.5393989 | 4.90E-39 | 28.12160017 | 3.42E-06 |
| NP3LGAGR | 65.21947951 | 3.29E-58 | 6.288244908 | 0.000318115 | 188.3265728 | 8.90E-39 | 16.66464165 | 0.000828318 |
| NP3POSTR | 64.28164622 | 1.82E-57 | 6.663405088 | 0.000188148 | 242.8870895 | 1.85E-50 | 16.64631963 | 0.000835528 |
| NP3PRSPL | 33.70594057 | 1.26E-31 | 2.592978048 | 0.051469703 | 127.5469073 | 7.89E-26 | 8.242157601 | 0.041263374 |
| NP3PTRML | 9.174539929 | 1.58E-08 | 0.860430003 | 0.461174853 | 41.23806109 | 8.40E-08 | 2.05551412 | 0.560964849 |
| NP3RIGN | 53.80495035 | 5.68E-49 | 5.665557404 | 0.000759188 | 207.1286097 | 8.47E-43 | 16.19212967 | 0.00103563 |
| NP3RIGRU | 33.46490357 | 2.08E-31 | 1.933114693 | 0.122600869 | 135.9363484 | 1.31E-27 | 4.973333328 | 0.173760256 |
| NP3RISNG | 76.91586753 | 2.96E-67 | 4.743890279 | 0.002734649 | 238.3805348 | 1.71E-49 | 17.88868868 | 0.000463722 |
| NP3RTCON | 4.41665815 | 0.000558019 | 0.546731547 | 0.650440698 | 21.43927232 | 0.000668976 | 2.041543301 | 0.563829162 |
| NP3SPCH | 50.47527127 | 3.39E-46 | 7.172006248 | 9.22E-05 | 179.1040873 | 8.31E-37 | 20.39129607 | 0.00014082 |
| NP3TTAPL | 50.86687428 | 1.59E-46 | 10.11551944 | 1.47E-06 | 183.4703505 | 9.71E-38 | 25.87650528 | 1.01E-05 |
| NP3TTAPR | 60.35669827 | 2.52E-54 | 5.958501336 | 0.000504395 | 202.8660012 | 6.92E-42 | 16.91761648 | 0.000734856 |

*Table 8: ANOVA and Kruskal-Wallis test on DaTscan image data and MDS-UPDRS clinical data*

The results of our performed Chi-squared test for feature importance in Disease Severity and Across Visits are below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Clinical Features** | **Chi-squared Test across Severity** | | **Chi-squared Test across Visit** | |
|  | **chi2** | **p-value** | **chi2** | **p-value** |
| NP1ANXS | 86.45042961 | 4.53E-12 | 11.98541779 | 0.21413263 |
| NP1PAIN | 178.5002549 | 1.91E-27 | 39.57785008 | 8.45E-05 |
| NP1SLPD | 192.4037643 | 3.56E-30 | 55.06753376 | 1.76E-07 |
| NP2DRES | 511.3494862 | 2.24E-99 | 53.13093232 | 2.76E-08 |
| NP2EAT | 404.8813595 | 6.52E-77 | 45.95675464 | 6.13E-07 |
| NP2FREZ | 330.2399202 | 5.18E-58 | 33.986991 | 0.000677735 |
| NP2HOBB | 379.6877161 | 3.30E-68 | 40.2048137 | 6.65E-05 |
| NP2HYGN | 240.4822133 | 9.42E-40 | 16.92746178 | 0.152346391 |
| NP2RISE | 411.4742337 | 8.49E-75 | 46.88674639 | 4.88E-06 |
| NP2SPCH | 270.1605269 | 8.58E-49 | 26.35404241 | 0.001787413 |
| NP2TURN | 394.3893439 | 1.04E-74 | 59.77029512 | 1.48E-09 |
| NP2WALK | 357.1877354 | 1.47E-63 | 37.10866779 | 0.00021452 |
| NP3BRADY | 371.5739003 | 1.57E-66 | 16.83507573 | 0.155907239 |
| NP3FACXP | 316.7670408 | 3.00E-55 | 15.61045747 | 0.209736133 |
| NP3FTAPL | 259.6574594 | 1.28E-43 | 15.76773987 | 0.202107091 |
| NP3FTAPR | 202.2449482 | 4.05E-32 | 6.440485605 | 0.892271524 |
| NP3HMOVL | 294.4088516 | 1.12E-50 | 19.08470308 | 0.08650472 |
| NP3HMOVR | 207.7895593 | 3.22E-33 | 13.77804355 | 0.315106086 |
| NP3LGAGL | 390.2480999 | 2.15E-70 | 35.04746037 | 0.000460228 |
| NP3LGAGR | 337.9947383 | 1.32E-59 | 23.43495459 | 0.024252243 |
| NP3POSTR | 290.088499 | 6.39E-53 | 25.59235017 | 0.002381262 |
| NP3PRSPL | 185.6616988 | 7.54E-29 | 16.16624224 | 0.183739974 |
| NP3PTRML | 52.1104481 | 5.42E-06 | 9.581481532 | 0.385415815 |
| NP3RIGN | 315.078133 | 6.66E-55 | 28.67357728 | 0.004403991 |
| NP3RIGRU | 191.2372729 | 1.27E-32 | 14.07719298 | 0.119605698 |
| NP3RISNG | 354.1512999 | 2.84E-66 | 20.72575189 | 0.013925384 |
| NP3RTCON | 35.93837269 | 0.015639609 | 22.23165697 | 0.035004038 |
| NP3SPCH | 423.2609167 | 8.86E-81 | 22.91801532 | 0.006383092 |
| NP3TTAPL | 292.2581087 | 3.07E-50 | 38.3168913 | 0.000136273 |
| NP3TTAPR | 403.6707116 | 3.54E-73 | 25.15145332 | 0.014121267 |

On the other hand, we performed a T-test between each visit to see which features changed after each visit:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Rank | V1 vs V2 | V1 vs V3 | V1 vs V4 | V2 vs V3 | V2 vs V4 | V3 vs V4 |
| 1 | NP1COG | NP3FACXP | NP3PTRML | NP3FTAPR | NP3RTALU | NP2TRMR |
| 2 | NP1ANXS | NP3PTRMR | NP3RTCON | NP1PAIN | NP3RTARL | NP3PRSPR |
| 3 | NP3LGAGR | NP2HYGN | NP3RTARU | NP3HMOVR | NP3FTAPR | NP3RTALU |
| 4 | NP3PTRMR | NP3RTALJ | NP3RTARL | NP3PRSPR | NP3PRSPR | NP3FTAPR |
| 5 | NP3FACXP | NP3RTALL | NP2TRMR | NP3PTRMR | NP3RIGLU | NP3RTARL |
| 6 | NP3RTALL | NP3RTCON | NP3PTRMR | NP3KTRMR | NP2TRMR | NP1APAT |
| 7 | NP3KTRML | NP3PTRML | NP3RTALL | NP3RIGRL | NP3PTRMR | NP3RTALL |
| 8 | NP1FATG | NP3FRZGT | NP1ANXS | NP3HMOVL | NP3HMOVR | NP1DPRS |
| 9 | NP3RTALJ | NP3KTRML | NP3RTALJ | NP3RTALL | NP3RTALL | NP3PTRMR |
| 10 | NP3RTARL | NP1ANXS | NP3RTALU | NP3SPCH | NP3RTALJ | NP3RTARU |
| 11 | NP1DDS | NP1COG | NP3FRZGT | NP3RIGRU | NP3RTCON | NP3RTCON |
| 12 | NP3FRZGT | NP3RTARU | NP3KTRML | NP3FRZGT | NP3RIGRU | NP3PTRML |
| 13 | NP3PSTBL | NP1FATG | NP3FTAPR | NP3FACXP | NP3KTRMR | NP3KTRMR |
| 14 | NP3RTCON | NP2TRMR | NP3PRSPR | NP2RISE | NP1DPRS | NP3HMOVR |
| 15 | NP1HALL | NP3LGAGR | NP1DPRS | NP3TTAPL | NP1ANXS | NP3BRADY |
| 16 | NP1DPRS | NP3RIGRL | NP3FACXP | NP3PRSPL | NP3PTRML | NP1DDS |
| 17 | NP1LTHD | NP1DPRS | NP3KTRMR | NP2DRES | NP3FRZGT | NP3FRZGT |
| 18 | NP3BRADY | NP3RTARL | NP1DDS | NP1DPRS | NP3HMOVL | NP3KTRML |
| 19 | NP2HYGN | NP1DDS | NP2HYGN | NP1URIN | NP3RTARU | NP3RIGRU |
| 20 | NP1APAT | NP3KTRMR | NP3HMOVR | NP3POSTR | NP2HYGN | NP3RIGLL |

The result analysis of the sequential prediction model is in ***Table 5***.

Table 5

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Attention-based LSTM Report** | | | | | | |
| **Class** | **precision** | | **recall** | | **f1-score** | **support** |
| Minimal | 1.0 | 0.33 | | | 0.50 | 3 |
| Mild | 0.90 | 0.93 | | | 0.95 | 19 |
| Moderate | 0.97 | 1.00 | | | 0.98 | 29 |
| Moderate To Severe | 1.00 | 0.90 | | | 0.95 | 10 |
| Severe | 1.00 | 1.00 | | | 1.00 | 5 |
| Very Severe | 1.00 | 1.00 | | | 1.00 | 3 |
| accuracy | – | – | | | 0.96 | 69 |
| macro avg | 0.98 | 0.87 | | | 0.90 | 69 |
| weighted avg | 0.96 | 0.96 | | | 0.95 | 69 |
| **Training Accuracy** | | | | **0.98** | | |
| **Testing Accuracy** | | | | **0.96** | | |

The comparison of results among the ML-based classification models before and after applying LDA to our multi-modal dataset are in ***Table 6*** and ***Table 7***.

Table 6

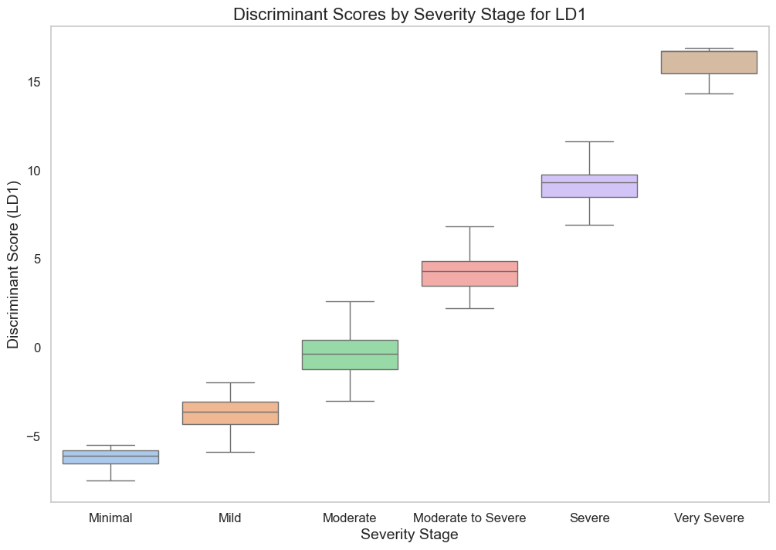
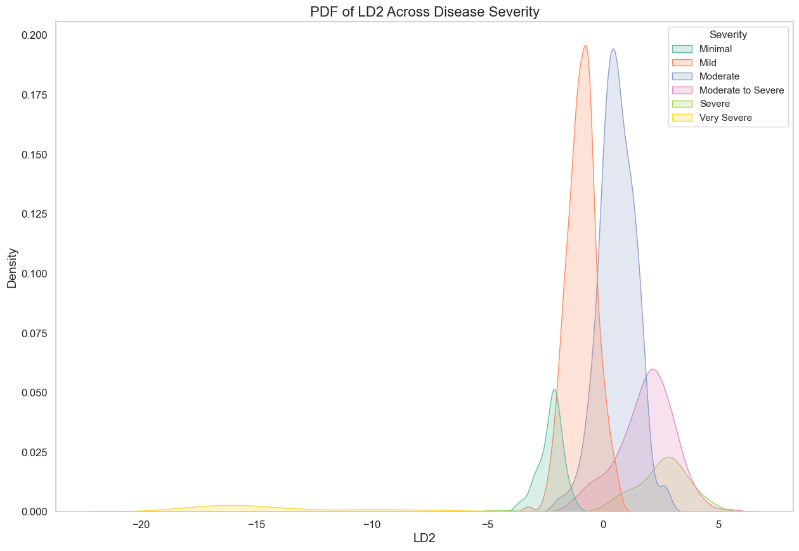
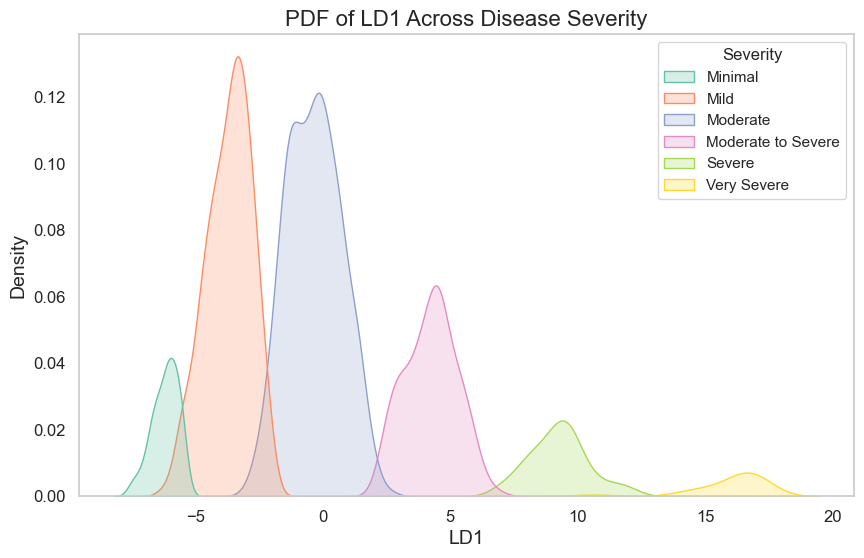
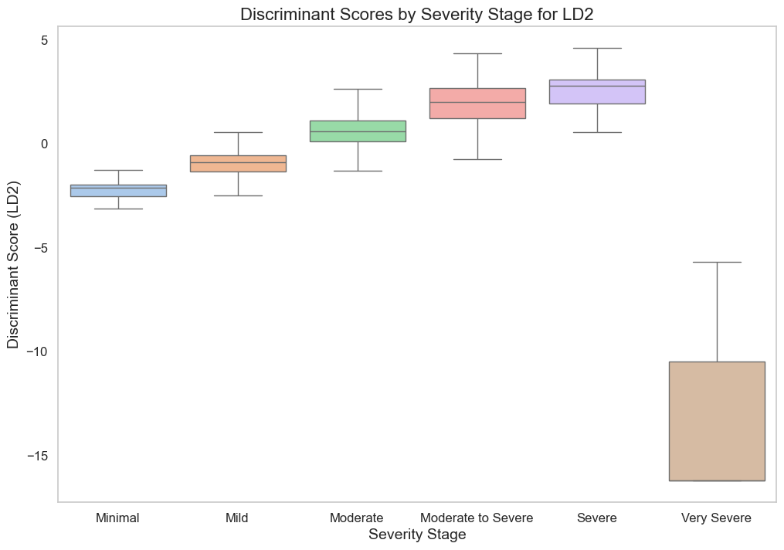
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification algorithm result After LDA with LD1 and LD2** | | | | | |
| **Algorithms** | **Training** | **Testing** | **Precision** | **Recall** | **F1-score** |
| Random Forest | 1 | 0.91 | 0.92 | 0.91 | 0.91 |
| Logistic Regression | 0.96 | 0.90 | 0.90 | 0.90 | 0.90 |
| Decision Tree | 1 | 0.90 | 0.90 | 0.90 | 0.90 |
| Gradient Boosting | 1 | 0.90 | 0.90 | 0.90 | 0.90 |
| KNN | 0.96 | 0.92 | 0.92 | 0.92 | 0.92 |
| SVM | 0.96 | 0.90 | 0.90 | 0.90 | 0.90 |
| Naïve Bayes | 0.96 | 0.90 | 0.91 | 0.90 | 0.90 |
| XGB | 0.99 | 0.91 | 0.91 | 0.90 | 0.90 |

Table 7

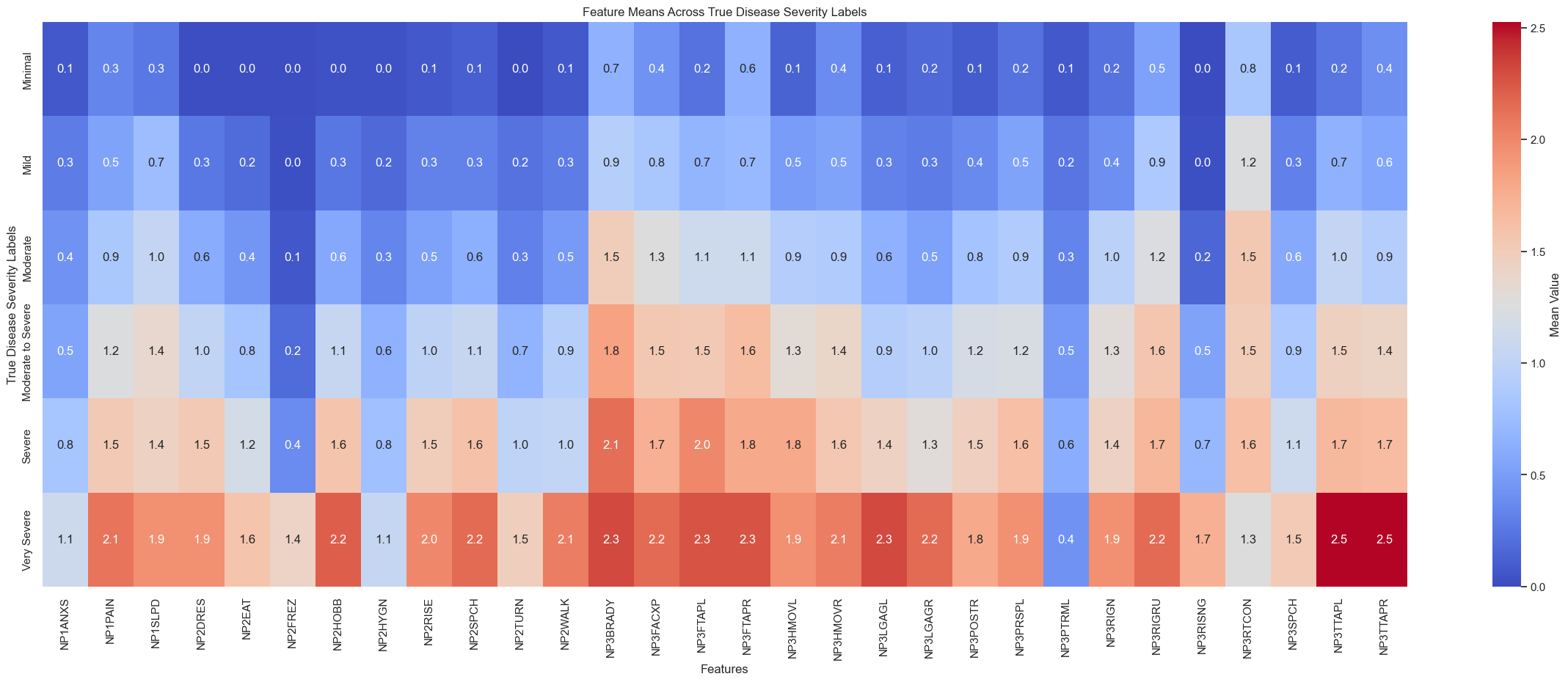
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification algorithm result Before LDA, with the actual multimodal Dataset** | | | | | |
| **Algorithms** | **Training** | **Testing** | **Precision** | **Recall** | **F1-score** |
| Random Forest | 1 | 0.64 | 0.68 | 0.64 | 0.61 |
| Logistic Regression | 0.99 | 0.83 | 0.84 | 0.83 | 0.82 |
| Decision Tree | 1 | 0.42 | 0.44 | 0.42 | 0.42 |
| Gradient Boosting | 1 | 0.64 | 0.65 | 0.64 | 0.62 |
| KNN | 0.72 | 0.57 | 0.61 | 0.57 | 0.55 |
| SVM | 0.97 | 0.82 | 0.86 | 0.82 | 0.81 |
| Naïve Bayes | 0.54 | 0.48 | 0.66 | 0.48 | 0.48 |
| XGB | 1.00 | 0.62 | 0.62 | 0.62 | 0.60 |

PPMI training data have 183 patients diagnosed with Parkinson's disease (PD) and the PPMI testing dataset contains 45 PD patients each contributing longitudinal data from four DaTscan visits. This approach allowed us to create a robust 4D dataset of 84 features combining clinical assessment MDS-UPDRS and DaTscan features which facilitate a comprehensive examination of the progression of PD over time. We employed an attention-based Long Short-Term Memory (LSTM) model to sequentially predict the disease's progression.

In our analysis dimensionality reduction technique Linear Discriminant Analysis (LDA) played a pivotal role, enabling us to reduce our 91 features to two linear discriminants (LDs) that accounted for 96% of the variance in the dataset. Notably, LD1 captured 93% of the variance alone, indicating its significance in distinguishing between patient disease states. Both LD1 and LD2 have a negative correlation with each other, as shown in Figure 3. By analyzing the linear discriminants (LDs), we extracted the top-ranking 20 features based on their contribution to each LD in *Figure 4*. These features are considered the most discriminative for differentiating between disease severity levels. The top features, identified through their coefficients in the LDs, provided insights into the underlying patterns of Parkinson's disease progression. Key features contributing to LD1 included Shannon Entropy and Contrast with respective scaling values of -12.63 and 12.24, while Area, Filled Area notably influenced LD2 with scaling values of 107.80



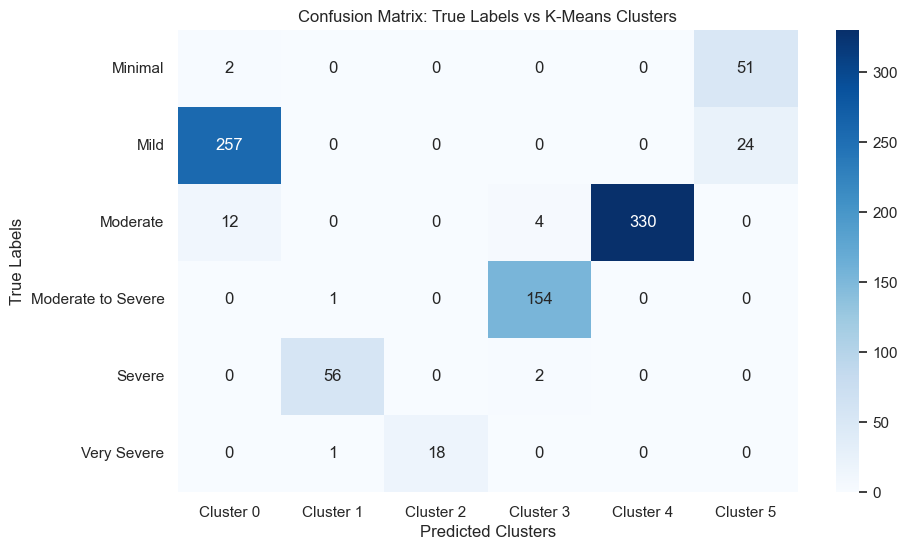
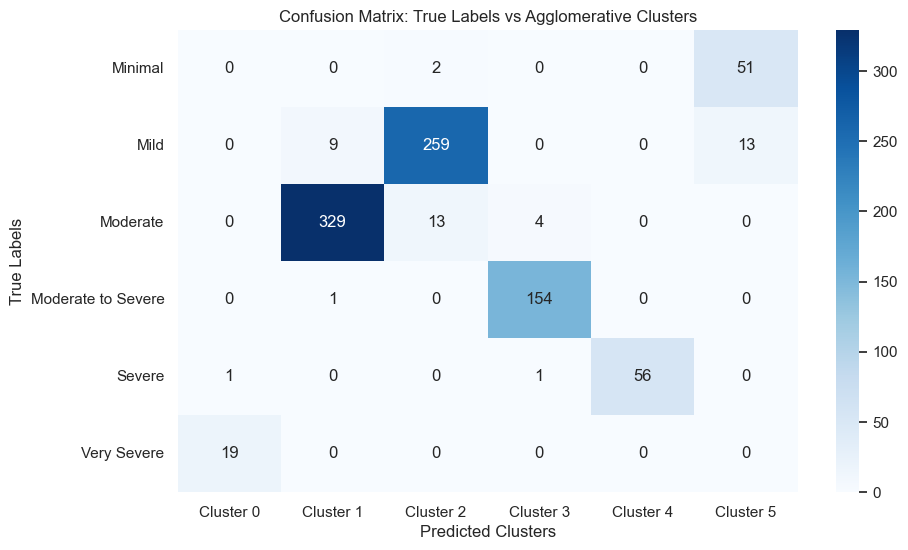
and -104.59. This feature extraction was instrumental in understanding the underlying factors influencing disease progression. After examining the top 20 feature contributions to LD1 and LD2, we found that 18 characteristics were shared by the two discriminants Which indicates that some features influence both linear discriminants. Major axis length and minor axis length were found to be unique factors in LD1, whereas solidity and LBP energy were the standout characteristics in LD2. These differences highlight the unique features of each discriminating value ​​to distinguish between different stages of Parkinson's disease progression: shape-related features are the main focus of LD1, whereas LD2 captures texture and stiffness components of DaTscan imaging data. The statisticalAnalyses revealed significant associations between key features extracted from LDA and Parkinson's disease severity state. In the Point Biserial Correlation results, DATSCAN\_PUTAMEN\_L\_ANT showed a strong negative correlation of -0.280 (0.05 > p = 6.49e-18), while gabor\_energy (0.151, 0.05 > p = 4.65e-06) and gabor\_entropy (0.139, 0.05 > p = 2.57e-05) exhibited positive correlations which indicate their clinical relevance in assessing disease progression. ANOVA results further supported these findings, with DATSCAN\_PUTAMEN\_L\_ANT yielding an F-statistic of 17.64 (0.05 > p = 1.05e-16), along with gabor\_energy (F = 7.44, 0.05 > p = 7.46e-07) and gabor\_entropy (F = 7.18, 0.05 > p = 1.32e-06), highlighting substantial variability across severity groups. Overall, these results underscore the significance of the top features extracted from the LDs to determine disease severity and progression. By clustering the LDA values using two methods which are K-Means



Clustering (KMC) in *Figure 6* and Agglomerative Clustering (Agg) in *Figure 5*. We compare both clusters with the true label in *Figure 7* and found Both methods revealed distinct clusters corresponding to different disease states.

The K-Means clustering obtained a silhouette score of 0.5115 means the clusters are well-defined and distinct but the agglomerative clustering obtained a slightly higher silhouette score of 0.5123.

Moreover, Agglomerative clustering achieved ARI of 0.8512 and NMI of 0.8653, whereas K-Means generated somewhat lower scores of ARI of 0.7947 and NMI of 0.8316. Both approaches demonstrated high scores for the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). These results indicate that the underlying structure of the data is successfully captured by the LDA components, allowing clustering at specific disease severity levels and demonstrating that the data are well clustered in the LDA-reduced space. We compared the K-Means and Agglomerative clustering labels with the actual disease severity labels and observed slight overlaps between the "Mild" and "Minimal" groups, 5th and 2nd clusters for Agglomerative Clustering on *Figure 15*, and0th and 5th clusters for K-Means on *Figure 16*. These overlaps highlight areas where further refinement in the clustering might be necessary. This phenomenon could be attributed to the early stages of Parkinson's disease, where symptoms often overlap between "Minimal" and "Mild" patients, making it challenging to distinguish these stages clearly in both clinical assessment and clustering outcomes. The probability density function (PDF) *Figure 10* and box plot of LD1 *Figure 8* show a strong positive connection between LD1 scores and the degree severity stage. LD1 values show a steadily rising trend with increasing severity, suggesting that higher scores correspond to more advanced stages of the disease and the value of the LD1 is distinct. These results highlight the possibility of LD1 serving as a useful marker of the severity



of the disease in individuals. For LD2 we observed that the correlation between severity stage *Figure 9* and LD2 is also positive, which means the higher the LD2 value the patient is in the more severe stage except for the very severe stage. For the very severe stage, the value of LD2 is less than -5 and the value of LD2 for other disease stages is greater than -5 *Figure 11*.

The use of the Kruskal-Wallis test for LD1 across disease severity groups produced a statistic of 826.57 with a p-value of 2.07e-176 << 0.05. Which suggests very significant differences among groups, the statistical analysis highlighted the relevance of our findings. These findings were further supported by an ANOVA, which showed significant variation in LD1 between severity groups (F = 3761.64, p < 0.001). Significant mean differences were found between all severity pairings by post-hoc analysis using Tukey's HSD test on LD1 *Table 2*, the biggest mean difference of 22.5, was found between the groups classified as minimal and very severe. The mean difference of 20 between Mild and Very severe The results of the Kruskal-Wallis test for LD2 showed significant differences as well, with a statistic of 669.77 and a p-value < 1.68e-142. Significant variability was seen in the LD2 ANOVA findings (F = 1083.48, p < 0.001). Tukey's HSD confirmed significant mean differences across all severity pairings, with a noteworthy 16.60 difference between the Very Severe and Severe groups among them *Table 3*.

|  |  |  |
| --- | --- | --- |
| **Mean Difference of disease with the last stage using LD1** | | |
|  |  | **Mean difference** |
| Minimal | Very Severe | 22.5 |
| Mild | Very Severe | 20 |
| Moderate | Very Severe | 16.6 |
| Moderate to Severe | Very Severe | 12 |
| Severe | Very Severe | 6.9 |

|  |  |  |
| --- | --- | --- |
| **Mean Difference of disease with the last stage using LD2** | | |
|  |  | **Mean difference** |
| Minimal | Very Severe | -11.7 |
| Mild | Very Severe | -13 |
| Moderate | Very Severe | -14.6 |
| Moderate to Severe | Very Severe | -16 |
| Severe | Very Severe | -16.5 |

The results of the Chi-square test for various clinical features extracted from some feature ranking techniques like Random forest, Recursive Feature Elimination (RFE), L1 Regularization (Lasso), and Mutual Information for Feature Selection prove their significance in assessing disease severity of Parkinson's disease. Most of the features ranked by these ranking techniques are clinical features which are categorical data so we use the Chi-square test *Figure 12*. By analyzing that **NHY** demonstrated a strong association with a Chi-square statistic of 81.66 (p = 0.0000000000347), indicating its critical role in various severity stages. Similarly, features such as **NP2DRES** (chi2 = 53.13, p = 0.0000000276) and **NP2RISE** (chi2 = 46.89, p = 0.00000488) showed significant relationships, suggesting their potential as key indicators in clinical assessments. Other significant features included **NP2EAT** (chi2 = 45.96, p = 0.000000613) and **NP2TURN** (chi2 = 59.77, p = 0.00000000148. We plot the heatmap categorical feature (MDS-UPDRS) across the disease severity *Figure 14* which gives a clear view of some feature values that are changing with the change of disease severity level. Overall, our findings not only elucidate the complex interplay between imaging features and clinical assessments in Parkinson's disease but also highlight the potential of advanced machine learning techniques, such as LSTM and LDA, in enhancing the predictive accuracy and understanding of disease progression.

## **5. Discussion**

We can see from *Table 4*, that all of the ML-based features selection techniques are prioritizing clinical data over image data. Even though RFFE tends to favor image-extracted data more compared to the other three techniques, the top five features in RFFE is also from MDS-UPDRS clinical motor data. Among the 59 clinical motor data, RFE, Fisher’s Score and Mutual Information score selected 20 significant moto features that are important where the selected features using Fisher’s score and MI score are almost equal.

In ANOVA, a higher F-statistic value indicates that the variance between groups is relatively large compared to the variance within groups. On the other hand, a lower p-value indicates that the null hypothesis can be rejected which means there is a statistically significant difference between the means of the features. After we conducted ANOVA across severity, we found image features: Gabor Energy, Gabor Entropy, Brightness, Mean, Standard Deviation are giving greater score in terms of F-statistics value and p-value. When ANOVA was conducted on image features across Visit: Skewness, Kurtosis, Standard Deviation, Brightness and Mean were giving higher F-statistics value and lower p-value.

Following that, ANOVA on categorical MDS-UPDRS clinical data across disease severity shows, NP2DRES (dressing), NP2HOBB (doing hobbies and other activities), NP2RISE (getting out of bed, car, or deep chair), NP3BRADY (global spontaneity of movement), NP3RISNG (arising from chair) are the top five features in terms of F-statistics and p-value scores among the 20 significant features. Alternatively, across visit we can see NP2TURN (turning in bed), NP1SLPD (daytime sleepiness), NP2DRES (dressing), NP2RISE (getting out of bed, car, or deep chair), NP2EAT (eating tasks) are giving higher F-statistics score and lower p-value score.

Kruskal-Wallis across severity on DaTscan image-extracted data also gives higher H-statistics score and lower p-value to: Gabor Energy, Gabor Entropy, Brightness, Mean, Standard Deviation. Afterwards, the Kruskal-Wallis test across visit on the same set of features shows Standard Deviation, Brightness, Mean, Skewness, Kurtosis have higher H-statistics and lower p-value score.

When we tested the MDS-UPDRS clinical data across severity with Kruskal-Wallis test, we found a bit different result than ANOVA giving higher H-statistics and lower p-value score in NP2DRES (dressing), NP2HOBB (doing hobbies and other activities), NP3BRADY (global spontaneity of movement), NP2RISE (getting out of bed, car, or deep chair), NP3POSTR (posture). Similar difference can be seen across visit on the same clinical data, NP2TURN (turning in bed), NP1SLPD (daytime sleepiness), NP2DRES (dressing), NP2EAT (eating tasks) and NP2WALK (walking and balance) scoring higher H-statistics value and lower p-value.

The Chi-squared test was performed on the categorical MDS-UPDRS clinical data across severity, and the result suggests that highest chi-squared scoring features, such as NP2DRES (dressing), NP3SPCH (speech), NP2RISE (getting out of bed, car, or deep chair), NP2EAT (eating tasks) and NP3TTAPR have a strong association with disease severity where NP2DRES (dressing), NP3SPCH (speech), NP2EAT (eating tasks), NP2RISE (getting out of bed, car, or deep chair) and NP2TURN (turning in bed) have the lowest p-value score.

The Chi-squared test across visit reveals significantly higher chi-squared values in NP2TURN (turning in bed), NP1SLPD (daytime sleepiness), NP2DRES (dressing), NP2RISE (getting out of bed, car, or deep chair) and NP2EAT (eating tasks) indicating a stronger association with patient visits and lower p-value in NP2TURN (turning in bed), NP2DRES (dressing), NP1SLPD (daytime sleepiness), NP2EAT (eating tasks) and NP2RISE (getting out of bed, car, or deep chair) suggests that these features are statistically significant.

From *Table 3*, LDA with LD1 scaling reveals the 20 most informative features contributing to disease severity. The features are sorted based on their scaling scores with Contrast being the most significant feature followed by Gabor Energy, Mean, Brightness and Area. In contrast, features like Convex Area, Equivalent Diameter, Filled Area, Energy, Homogeneity, Standard Deviation, Gabor Entropy, Skewness have negative scaling values, demonstrating that these features are less important for the progress of disease severity. Our experiment also exhibits that DaTscan image-extracted features are more important than MDS-UPDRS clinical data because out of 25 image-extracted features 20 of them were selected as important but not one clinical motor feature is present in that list.

At the same time, LDA with LD2 reveal a distinct set of notable features than LD1. Contrary to LD1, Area emerges as the most prime feature in LD2, followed by Homogeneity, Shannon Entropy, Contrast and Kurtosis. However, Equivalent Diameter, Filled Area, Energy, Standard Deviation, Gabor Entropy, Skewness have negative scaling values in LD2 as well, suggesting they are not the key features contributing to the progression of disease severity.

Our attention-based LSTM model in *Table 5* demonstrates splendid performance in classifying the disease severity, with an overall testing accuracy of 0.96. Notably, the model achieves perfect precision, recall and f1-score in the Severe and the Very Severe class. The Moderate class also exhibits high precision and recall which are 0.97 and 1.00 respectively, suggesting our model can distinguish this class from others with a high degree of accuracy. The same decision also applies for the Mild and the Moderate to Severe class where both classes are equal with a f1-score of 0.95. Alternatively, the minimal class shows a relatively lower recall with f1-score of just 0.50. The macro average and weighted average metrics further support our model’s overall performance with f1-scores of 0.90 and 0.95, respectively. These results suggest that our attention-based LSTM model is a promising approach for disease severity classification tasks, particularly in Severe and Very Severe class.

In *Table 7*, the classification results before applying Linear Discriminant Analysis (LDA) to our multimodal dataset shows varying degrees of performance among the ML-based algorithms. Logistic Regression and Support Vector Machine (SVM) comes out as the top performing algorithms, with a F1-score of 0.82 and 0.81 respectively. It suggests that Logistic Regression and SVM are effective in capturing the underlying patterns in our multimodal data. In contrast, Decision Tree and Naïve Bayes algorithms exhibit relatively poor performance with F1-score of 0.42 and 0.48 respectively. Though KNN performs better than DT and NB, the F1-score of 0.55 does not recommend it as a good algorithm either. The ensemble methods, such as Random Forest, Gradient Boosting and XGB demonstrate mediocre performance with F1-score ranging from 0.60 to 0.62.

Applying LDA with LD1 and LD2 features significantly improves the performance of all classification algorithms which can be seen in *Table 6*. Notably, Random Forest, KNN, and XGB achieve remarkable testing accuracies of 0.91, 0.92, and 0.91 respectively, with corresponding F1-scores of 0.91, 0.92, and 0.90 respectively. Furthermore, the other algorithms, such as Logistic Regression, DT, Gradient Boosting, SVM, and Naïve Bayes, also exhibit similarly improved performance with equal testing accuracy and F1-score 0.90 for all algorithms. The improvement across all ML-based algorithms suggests that the LDA transformation has adequately minimized the dimensionality of our data, making it more distinct and simpler to classify. Following that, the results also imply that the LD1 and LD2 features extracted through LDA are highly informational and appropriate for our classification task which allows the algorithms to better distinguish between classes.

## **6.Conclusion**

**Reference**

1. Amoroso, Nicola, et al. "Complex networks reveal early MRI markers of Parkinson’s disease." Medical image analysis 48 (2018): 12-24.
2. Khachnaoui, Hajer, et al. "Enhanced Parkinson’s Disease Diagnosis Through Convolutional Neural Network Models Applied to SPECT DaTSCAN Images." IEEE Access (2023).
3. Junaid, Muhammad, et al. "Explainable machine learning models based on multimodal time-series data for the early detection of Parkinson’s disease." Computer Methods and Programs in Biomedicine 234 (2023): 107495.
4. Yagis, Ekin, Alba G. Seco De Herrera, and Luca Citi. "Generalization performance of deep learning models in neurodegenerative disease classification." 2019 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, 2019.
5. Prashanth, R., et al. "High-accuracy detection of early Parkinson's disease through multimodal features and machine learning." International journal of medical informatics 90 (2016): 13-21.
6. Harvey, Joshua, et al. "Machine learning-based prediction of cognitive outcomes in de novo Parkinson’s disease." npj Parkinson's Disease 8.1 (2022): 150.
7. Zhang, Xiaobo, et al. "Multi-level Graph Neural Network with Sparsity Pooling for Recognizing Parkinson’s Disease." IEEE Transactions on Neural Systems and Rehabilitation Engineering (2023).
8. Zhang, Xi, et al. "Multi-view graph convolutional network and its applications on neuroimage analysis for Parkinson’s disease." AMIA Annual Symposium Proceedings. Vol. 2018. American Medical Informatics Association, 2018.
9. Huang, Zhongwei, et al. "Parkinson’s disease classification and clinical score regression via united embedding and sparse learning from longitudinal data." IEEE Transactions on Neural Networks and Learning Systems 33.8 (2021): 3357-3371.
10. Hajianfar, Ghasem, et al. "Prediction of Parkinson’s disease pathogenic variants using hybrid Machine learning systems and radiomic features." Physica Medica 113 (2023): 102647.
11. Wingate, James, et al. "Unified deep learning approach for prediction of Parkinson's disease." IET Image Processing 14.10 (2020): 1980-1989.
12. Bhat, Shreya, et al. "Parkinson's disease: Cause factors, measurable indicators, and early diagnosis." Computers in biology and medicine 102 (2018): 234-241.
13. Abdulhay, Enas, et al. "Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease." Future Generation Computer Systems 83 (2018): 366-373.
14. Lundervold, Alexander Selvikvåg, and Arvid Lundervold. "An overview of deep learning in medical imaging focusing on MRI." Zeitschrift für Medizinische Physik 29.2 (2019): 102-127.
15. Majhi, Babita, et al. "An improved method for diagnosis of Parkinson’s disease using deep learning models enhanced with metaheuristic algorithm." BMC medical imaging 24.1 (2024): 156.
16. Poewe, Werner, et al. "Parkinson disease." Nature reviews Disease primers 3.1 (2017): 1-21.
17. Chaudhuri, K. Ray, and Anthony HV Schapira. "Non-motor symptoms of Parkinson's disease: dopaminergic pathophysiology and treatment." The Lancet Neurology 8.5 (2009): 464-474.
18. Hirschauer, Thomas J., Hojjat Adeli, and John A. Buford. "Computer-aided diagnosis of Parkinson’s disease using enhanced probabilistic neural network." Journal of medical systems 39 (2015): 1-12.
19. Suwijn, Sven R., et al. "The diagnostic accuracy of dopamine transporter SPECT imaging to detect nigrostriatal cell loss in patients with Parkinson’s disease or clinically uncertain parkinsonism: a systematic review." EJNMMI research 5 (2015): 1-8.
20. Adeli, Ehsan, et al. "Kernel-based joint feature selection and max-margin classification for early diagnosis of Parkinson’s disease." Scientific reports 7.1 (2017): 41069.
21. Salvatore, Christian, et al. "Machine learning on brain MRI data for differential diagnosis of Parkinson's disease and Progressive Supranuclear Palsy." Journal of neuroscience methods 222 (2014): 230-237.
22. Khachnaoui, Hajer, Rostom Mabrouk, and Nawres Khlifa. "Machine learning and deep learning for clinical data and PET/SPECT imaging in Parkinson's disease: a review." IET Image Processing 14.16 (2020): 4013-4026.
23. Taylor, Jonathan Christopher, and John Wesley Fenner. "Comparison of machine learning and semi-quantification algorithms for (I123) FP-CIT classification: the beginning of the end for semi-quantification?." EJNMMI physics 4 (2017): 1-20.
24. Raiaan, Mohaimenul Azam Khan, et al. "Mammo-Light: A lightweight convolutional neural network for diagnosing breast cancer from mammography images." Biomedical Signal Processing and Control 94 (2024): 106279.
25. Al Mahmud, Abdullah, et al. "SkinNet-14: a deep learning framework for accurate skin cancer classification using low-resolution dermoscopy images with optimized training time." Neural Computing and Applications (2024): 1-25.
26. Thakur, Ajay Kumar, et al. "A Deep Neural Network for Detecting Lung Cancer from CT Images." International conference on smart computing and cyber security: strategic foresight, security challenges and innovation. Singapore: Springer Nature Singapore, 2023.
27. Shamrat, FM Javed Mehedi, et al. "AlzheimerNet: An effective deep learning based proposition for alzheimer’s disease stages classification from functional brain changes in magnetic resonance images." IEEE Access 11 (2023): 16376-16395.
28. Kim, Hee E., et al. "Transfer learning for medical image classification: a literature review." BMC medical imaging 22.1 (2022): 69.
29. Khachnaoui, Hajer, Nawres Khlifa, and Rostom Mabrouk. "Machine learning for early Parkinson’s disease identification within SWEDD group using clinical and DaTSCAN SPECT imaging features." Journal of Imaging 8.4 (2022): 97.
30. Marek, Kenneth, et al. "The Parkinson progression marker initiative (PPMI)." Progress in neurobiology 95.4 (2011): 629-635.
31. Mining, What Is Data. "Data mining: Concepts and techniques." Morgan Kaufinann 10.559-569 (2006): 4.
32. Saeys, Yvan, Inaki Inza, and Pedro Larranaga. "A review of feature selection techniques in bioinformatics." bioinformatics 23.19 (2007): 2507-2517.
33. Ali Khan, Sajid, et al. "Kruskal‐Wallis‐Based Computationally Efficient Feature Selection for Face Recognition." The Scientific World Journal 2014.1 (2014): 672630.
34. Lin, Hao, and Hui Ding. "Predicting ion channels and their types by the dipeptide mode of pseudo amino acid composition." Journal of theoretical biology 269.1 (2011): 64-69.