Investigating the use of radiomics for analysis of DAT SPECT imaging in Parkinson's Disease

Sofiia Tatosh
dept. of Information Engineering
University of Padova
Padova, Italy
sofiia.tatosh@studenti.unipd.it

Griselda Kolici
dept. of Information Engineering
University of Padova
Padova, Italy
griselda.kolici@studenti.unipd.it

Filippo Canderle
dept. of Information Engineering
University of Padova
Padova, Italy
filippo.canderle@studenti.unipd.it

***Abstract*—Radiomics features have emerged as a promising approach to extracting quantitative information from medical images, enabling the characterization and analysis of diseases like Parkinson's. In this study, we aimed to investigate the potential of radiomic features in distinguishing between patients with Parkinson's disease and healthy controls. Additionally, we sought to explore the correlation between these radiomic variables and the severity scores of Parkinson's disease. Furthermore, we employed machine learning techniques to develop predictive models using the radiomic features dataset. [1] [2]**

Initially, a comprehensive set of radiomic features was extracted from dopamine transporter (DAT) single-photon emission computed tomography (SPECT) scans obtained from patients diagnosed with Parkinson's disease and matched controls. Statistical analysis was performed to assess the discriminative power of these features in differentiating patients from controls. Next, we examined the relationship between the radiomic variables and severity scores of Parkinson's disease. By utilizing established severity assessment scales, such as the Unified Parkinson's Disease Rating Scale (UPDRS), we aimed to identify potential associations and quantify the influence of radiomic features on disease severity.

Furthermore, machine learning models were constructed to predict the presence of Parkinson's disease and classify individuals into patient or control groups based on the extracted radiomic features. Various algorithms, such as support vector machines, random forests, and others, were employed to develop robust and accurate predictive models.

Overall, our findings demonstrated the utility of radiomic features in Parkinson's disease research. The exploratory data analysis revealed significant differences between patients and controls, suggesting the potential of these features as biomarkers for disease detection. Moreover, The statistical analysis emphasized that there is a weak but significant relationship between particular radiomics features and severity scores in Parkinson's disease (PD).

Finally, the machine learning-based prediction models demonstrated promising results in accurately classifying individuals as patients or controls, showcasing the potential clinical applications of radiomic features in diagnosing and managing Parkinson's disease.

***Index Terms*—Parkinson's disease, Radiomics, Statistical analysis, Machine Learning**

I. Background

DAT SPECT (Single-Photon Emission Computed Tomography) is a nuclear imaging technique that measures the density and activity of dopamine transporters (DAT) in the brain. In Parkinson's disease, there is a loss of dopaminergic neurons, leading to decreased DAT levels. Through the DAT SPECT examination, it is possible to assess the amount of DAT present in an individual's brain, providing information on the function and integrity of dopaminergic neurons. This information can be used to diagnose Parkinson's disease and determine its severity [2].

Recent literature has seen research efforts combining radiomics features and DAT SPECT to establish a relationship between these features and the severity level of Parkinson's disease, ultimately enabling personalized therapy for patients through a precision medicine approach. The integration of radiomics features with DAT SPECT offers the potential to extract additional and detailed information regarding several characteristics of DAT tracer in the brains of Parkinson's patients. These features may provide valuable insights into the neurodegenerative process and aid in developing personalized treatment strategies [4].

This study aims to go into depth about the relationship between radiomics features and Parkinson's disease (PD) and investigate whether there are any associations between these features and the disease. This engineering and statistical approach to a medical problem aims to explore potential mathematical relationships among these types of medical features and assess the feasibility of implementing a machine learning model capable of autonomously discriminating patients from healthy controls.

In summary, we want to answer three questions:

1) Is there any difference between controls and PD in radiomics features (or combined score)?
2) Are the radiomic features or any derived scores associated with PD clinical symptoms severity?
3) Are the radiomics features capable of distinguishing patients from controls?

II. Materials and methods

A. Dataset

The dataset used in this research was obtained from the PET NODE REPOSITORY at King's College London. It includes radiomics features derived from DAT SPECT imaging scans of 33 idiopathic PD patients and 20 matched healthy controls. SPECT imaging scans were conducted using the

radioligand [123I]FP-CIT, targeting the dopamine transporter system. Standardized Uptake Value Ratios (SUVRs) were calculated to approximate the binding potential of the putamen region, with the cerebellum serving as the reference tissue.

The dataset utilized for the present study is a comprehensive collection of three distinct sub-datasets, namely radiomics features, demographic features, and clinical features, aimed at investigating Parkinson's disease (PD) severity [3].
Radiomics features were extracted for each subject using the MIRP Python package. The radiomics features sub-dataset comprises 177 columns and 53 rows, with 33 rows representing PD patients and 20 rows corresponding to healthy controls. These radiomics features encompass a wide range of quantitative imaging measurements acquired from each subject, capturing intricate details of PD-affected brain regions. The columns in this sub-dataset include features such as peak location, statistical descriptors (mean, variance, skewness, kurtosis, median, minimum, 10th percentile, 90th percentile, maximum, interquartile range, range, median absolute deviation), morphological characteristics (volume, area, sphericity, asphericity, compactness), and various imaging histogram descriptors. Additionally, texture-based features derived from intensity and co-occurrence matrices provide further insights into the spatial distribution and patterns within the images.
The demographic features sub-dataset includes 8 columns and 53 rows, matching the subjects in the radiomics sub-dataset, providing demographic information such as age, gender, education level, height, weight, and BMI. The clinical features sub-dataset consists of 33 rows, corresponding to the PD patients, and includes scores indicative of PD severity, encompassing motor and non-motor impairments and medication dosage assessed through the LEDD Total, UPDRS I-IV scores, UPDRS total, NMSQ score, MMSE score, and MoCA score. Integrating these multimodal datasets offers a comprehensive approach to investigating the relationships between radiomics features, demographic factors, and clinical scores, enabling a deeper understanding of Parkinson's disease and facilitating the development of accurate assessment models.

B. Research methods

Our study was structured into four main sections: exploratory data analysis, statistical analysis, the implementation of machine learning models and sensitivity analysis. Each section aimed to address specific research questions:

- Exploratory Data Analysis: This section aimed to investigate whether there were any differences in radiomics features (or combined scores) between healthy controls and individuals with Parkinson's disease (PD).
- Statistical Analysis: In this section, we examined the association between radiomics features or derived scores and the severity of PD clinical symptoms. We utilized established clinical assessment scales, such as the Unified Parkinson's Disease Rating Scale (UPDRS), to quantify the relationship and explore potential prognostic factors.
- Machine Learning Models: The section focused on assessing the capability of radiomic features to distinguish between patients with PD and healthy controls. Various machine learning algorithms, including logistic regression, support vector machine, and random forest, were implemented to develop predictive models that accurately classify individuals.
- Sensitivity Analysis: The final section was devoted to evaluating sensitivity of models' results to covariates, group matching and data quality (the implact of outliers, imbalanced dataset)

1) Exploratory Data Analysis: In this part, we aimed to investigate the characteristics of a dataset involving examining demographic and clinical variables, focusing on gender, age, education, and BMI. The dataset was divided into subsets based on gender, patients, and controls. For each subset, we calculated the mean and standard deviation of the variables of interest.

First, we compared the mean age and standard deviation between males and females in the entire dataset. We utilized visualizations, including histograms and box plots, to explore the age distribution by gender. The analysis revealed that the mean age for males and females was comparable.

Next, we focused on the comparison between patients and controls. We examined the mean and standard deviation of age, education (years), and BMI for these two groups separately. Violin plots were employed to visualize the distributions of these variables. The results demonstrated that the mean age, education, and BMI between patients and controls were comparable.

To investigate the possible differences in radiomics features for HC and PD, we moved forward with selecting a smaller subset of the features that appear to be the most discriminant concerning the assigned group. First, we calculated a pairwise Spearman correlation coefficient and p-value as a correlation statistical significance metric to select a set of features with a low correlation coefficient. Additionally, we performed a statistical analysis to identify variables with significant differences in standard deviation between patients and controls. The percentage difference in standard deviation for each variable was calculated, and a threshold of $\theta = 0.3$ was used to select the potential discriminant features. The variables with the highest differences were selected, and a t-test was conducted to compute the corresponding p-values. The top 10 features with the highest standard deviation differences were reported.

2) Statistical Analysis: In order to evaluate the possible association of the radiomics features with the PD symptoms severity, we exploited Spearman correlation, information gain, and ANOVA test approaches.

The Spearman correlation coefficient was calculated for each pair (f_i, s_j), where f_i is a radiomics feature and s_j is a clinical score presented in a clinical dataset. Each coefficient was evaluated using p-value, where H_0 hypothesis, correlation is not statistically significant, was rejected whenever $p_{(i,j)} \leq 0.1$ and accepted otherwise. Then, considering the acceptable correlation coefficient ($SCC_{(i,j)} > 0.2$), where

$SCC_{(i,j)}$ is Spearman Correlation Coefficient for (f_i, s_j)), meaning at least weak correlation is present, and low p-value, were the deciding factors for selecting a subset of features f that is associated to a clinical score s_j.

Another statistical measurement of radiomics association with clinical scores of PD patients was mutual information. Mutual information (MI) is a non-negative measure of the amount of information one random variable has about another variable [5]. In selecting the most informative features, this method is helpful because it quantifies the relevance of a feature subset to the continuous output variable (s_j). Based on the mutual information scores, we have selected the top 10 most informative radiomics features for every clinical score provided.

The third approach in our analysis was to investigate the differentiation potential of Parkinson's disease severity in the radiomics features. In order to perform this type of test, we first defined the UPDRS score cutoffs that define mild, moderate, and severe stages of PD symptoms. According to (Martínez-Martín et al., 2015), the MDS-UPDRS identifies mild, moderate, and severe stages of PD w.r.t. MDS-UPDRS Part I, II, III and IV scores as follows in Table I.

TABLE I
CUTOFF POINTS OF THE MDS-UPDRS

Severity	**Total Score**		
Levels according to	**Mild**	**Moderate**	**Severe**
MDS-UPDRS Part I	1–10	11-21	≥22
MDS-UPDRS Part II	1-12	13-29	≥30
MDS-UPDRS Part III	1-32	33-58	≥59
MDS-UPDRS Part IV	1-4	5-12	≥13

Having the subjects labeled according to each of the MDS-UPDRS parts, the cumulative score was assigned depending on what severity level is prevalent overall. The split we received was that 21 of 33 PD patients were labeled as having mild stages of Parkinson's disease, and 12 patients were given a moderate level. There were no patients with severe stages of Parkinson's according to this scale, so only the two groups were further analyzed. To know whether the two groups, mild and moderate, have any differences in the radiomics features, we performed a one-way ANOVA test that resulted in selecting the top 10 features that can differentiate the two groups. They were selected based on p-value with a standard threshold of $p_i \leq 0.05$.

3) Machine Learning applications: The methodology used in this study involved the application of several machine learning models to address the research question of whether radiomics features can distinguish patients from controls. The first step was feature selection using SVM-RFE (Support Vector Machine with Recursive Feature Elimination). SVM-RFE primarily aims to compute the ranking weights for all features and sort the features according to weight vectors as the classification basis [13]. Using SVM-RFE resulted in the selection of 10 features for each model. Alternative feature selection techniques, including Lasso regularization and Principal Component Analysis (PCA), were also explored. Lasso is a method of regression analysis that performs feature selection and regularization to improve prediction accuracy via penalized estimation functions [14]. Principal component analysis (PCA) reduces the dimensionality of such datasets, increases interpretability, and minimizes information loss. It does so by creating new uncorrelated variables that successively maximize variance [15].

The selected features were then used as inputs for various classification algorithms, including Random Forest, MKL SVM (Multiple Kernel Learning Support Vector Machine), and Logistic Regression. The dataset is divided into k (10) folds using StratifiedKFold. The task is performed iteratively for each fold of the data. Within each fold, the data is split into training and testing sets. Then the algorithms were trained on the training data and evaluated on the test data.

The performance of the models was assessed using various metrics. The test accuracy measured the overall accuracy of the models in correctly classifying instances from the test set. The sensitivity metric evaluated the models' ability to identify positive instances, i.e., patients, correctly. The area under the curve (AUC) value represented the models' discrimination ability.

4) Sensitivity Analysis: Sensitivity analyses play a crucial role in assessing the robustness of the findings or conclusions based on primary analyses of data in clinical trials [16]. This project aimed to evaluate the impact of outliers, class imbalance dealing techniques, and subgroup variations on the performance of the PCA and Random Forest model in analyzing radiomics data for Parkinson's disease patients and healthy controls. Firstly, the sensitivity analysis examined the effect of outliers on the model's performance. The analysis involved calculating z-scores and identifying outliers using a threshold of 3. Then, we explored the impact of different class imbalance dealing techniques on the model's performance. SMOTE technique was used firstly to address the class imbalance. The Synthetic Minority over Sampling Technique (SMOTE) is an enhanced sampling method in which new synthetic sampling is computed based on Euclidian distance for variables. As a result, the synthetic cases will have attributes with values similar to the existing cases and not merely replications as oversampling does, thus, increasing the representation of the minority class in the resulting dataset while reflecting the structure of the original cases. It has been shown that SMOTE is robust to the variation of unbalanced ratio with various classifiers [17]. In addition, a combination of undersampling and oversampling techniques was used. Oversampling strategy creates examples of the minority class to balance the dataset. It helps to improve the performance of models but presents the drawback of overfitting and introduces additional noise. The under-sampling strategy addresses the imbalance problem by eliminating the members of the majority class, showing an advantage in saving computation time while will discard potentially useful information. The hybrid method combines the over-sampling of the minority class with the

under-sampling of the majority class. Previous studies have reported successfully applying various re-sampling methods in radiomics analysis [18]. In addition, a sensitivity analysis based on gender-based subgroups using PCA (Principal Component Analysis) and Random Forest was performed. The data used in the analysis consists of two datasets: 'Radiomics' and 'Demographics Clinical.' These datasets were merged based on the standard identifier column, 'id_subject.' For the sensitivity analysis based on age subgroups using PCA and Random Forest, the analysis began by defining an age threshold set at 65. The subject IDs of individuals classified as old (age greater than or equal to the threshold) and young (age less than the threshold) are extracted from the dataset. The data is then divided into two subgroups: old and young. For the old subgroup, the corresponding data and labels are selected based on the subject IDs of old individuals. Similarly, the data and labels for the young subgroup are obtained using the subject IDs of young individuals. Next, the analysis is performed separately for each subgroup using PCA and Random Forest. Next, a sensitivity analysis through a subgroup analysis based on BMI (Body Mass Index) using PCA and Random Forest was done. Three subgroups are defined based on BMI ranges: "Underweight," "Healthy Range," and "Overweight/Obesity." For each subgroup, the corresponding indices of the merged DataFrame are identified based on the BMI range. The subgroup's data and labels were extracted accordingly. The average metrics include accuracy, sensitivity, specificity, and AUC.

III. RESULTS

1) Exploratory Data Analysis: The exploratory data analysis was divided into two parts. Firstly, we examined the dataset's gender-related characteristics to determine any demographic differences between patients and controls. Subsequently, we analyzed the radiomics features to assess whether there were differences between the two groups in terms of radiomic characteristics. The statistical analysis yielded several important findings.

First, we examined the gender-related characteristics of the dataset. The mean age for females was 62.24 years, with a standard deviation of 8.55 years, while for males, it was 61.17 years, with a standard deviation of 9.56 years (Figure 1). The age means for males and females in the entire dataset were comparable.

Next, we conducted a subset analysis to compare age, education (years), and BMI between patients and controls. The mean age for patients was 60.55 years, with a standard deviation of 9.52 years, while for controls, it was 63.10 years, with a standard deviation of 8.59 years. The age means for patients and controls in the entire dataset was comparable.

Regarding education (years), the mean for patients was 26.98, with a standard deviation of 4.99, while for controls, it was 27.12, with a standard deviation of 4.14. The education means for patients and controls in the entire dataset were comparable.

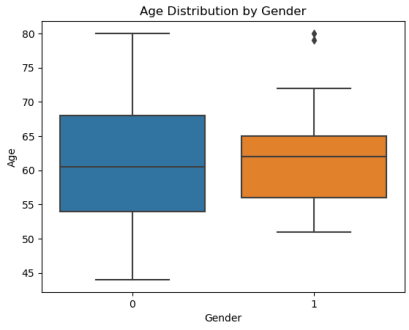

Fig. 1. Age distribution by gender

The mean BMI for patients was 15.79, with a standard deviation of 3.59, while for controls, it was 16.55, with a standard deviation of 3.44. The dataset's BMI means for patients and controls were comparable (Figure 2).

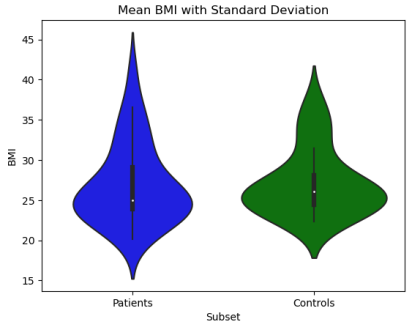

Fig. 2. BMI distribution for patients and controls

In the second part of the analysis, we also compared radiomic features between patients and controls. In order to do this, we first computed a pairwise Spearman correlation of radiomics features to see whether we could create a subset of features that are not correlated and hence independent of each other to reduce the dimensionality of the dataset. With our analysis in consideration, there were no features found with both weak to no correlation and p-value ≤ 0.05, which means no statistically significant low correlation pairs of features were not found. Thus, we computed the variance of every radiomic feature to find the ones with the highest variance and so highest amount of information retained from the entire dataset.

Among the features with the highest difference in standard deviation, we conducted independent t-tests and computed p-values, as we broadly explained in the *Methods* section. Table II presents the results for the top 10 features with the highest difference in standard deviation.

In summary, the analysis of demographic features did not reveal any quantitative differences between the two groups. However, examining the radiomics features demonstrated significant differences between patients and controls, as confirmed by statistical tests where the p-values were below the predetermined threshold($p \leq 0.05$). Additionally, based

TABLE II
P-VALUES FOR FEATURES WITH THE HIGHEST VARIANCE

Feature	P-value
ngt_complexity_3d_fbs_w0.0125	0.0189
dzm_ldhge_3d_fbs_w0.0125	0.0096
ngl_ldhge_d1_a0.0_3d_fbs_w0.0125	0.0117
cm_auto_corr_d1_3d_avg_fbs_w0.0125	0.0086
cm_contrast_d1_3d_avg_fbs_w0.0125	0.0002
ngl_hgce_d1_a0.0_3d_fbs_w0.0125	0.0091
rlm_srhge_3d_avg_fbs_w0.0125	0.0093
rlm_hgre_3d_avg_fbs_w0.0125	0.0092
rlm_lrhge_3d_avg_fbs_w0.0125	0.0088
szm_szhge_3d_fbs_w0.0125	0.0122

on the findings from the first part, it can be inferred that the observed differences between the two groups regarding radiomics features are not influenced by demographic features.

To visually represent the differences between groups based on the selected features, we computed a coefficient of variation and depicted the between-subject variability on the Figure 3.

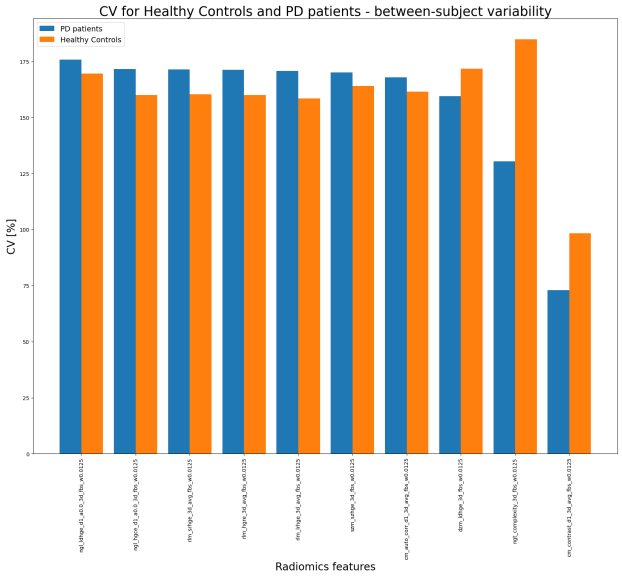

Fig. 3. Between-subject variability - Coefficient of Variation for HC vs. PD

As we can see, some of the CV [%] values are higher than 100%, which from the statistics itself, means that for some values standard deviation is higher than the mean. According to our further examination of the distributions of the features separately (both visually and statistically), we can conclude that high variability for these features is driven by the outliers present both in the subset of patients and the subset of healthy controls. Notably, the % variability between groups' standard deviation values should remain the same if removing outliers that affect certain radiomics features due to the nature of calculations. Due to a limited dataset, we considered proceeding without dropping outliers in the early stages since doing that did not bring up a different subset of features that are the most differentiable ones between groups.

2) Statistical Analysis: Spearman correlation was the first step in investigating the relationship between the clinical scores and radiomics features. The findings showed that most selected features were based on weak to moderate correlation supported by low p-values. Additionally, since the number of significant radiomics features differed for each clinical score, we received relevant results primarily for UPDRS II, III, and Total, along with NMSQ and MMSE. In Table III, we only provide the results for these clinical scores.

TABLE III
CORRELATION ANALYSIS BETWEEN FEATURES AND CLINICAL SCORES

Feature	SCC	p-value
	UPDRS II	
loc_peak_loc	-0.332	0.059
loc_peak_glob	-0.333	0.059
morph_comp_1	-0.346	0.049
morph_comp_2	-0.346	0.049
morph_sph_dispr	0.346	0.049
morph_sphericity	-0.346	0.049
morph_asphericity	0.346	0.049
morph_vol_dens_aee	-0.353	0.044
morph_vol_dens_conv_hull	-0.382	0.028
morph_area_dens_conv_hull	-0.399	0.021
	UPDRS III	
morph_vol_dens_aee	-0.409	0.018
morph_pca_elongation	-0.416	0.016
dzm_ldlge_3d_fbs_w0.0125	0.424	0.014
stat_qcod	-0.431	0.012
stat_cov	-0.45	0.009
morph_vol_dens_conv_hull	-0.46	0.007
morph_pca_flatness	-0.476	0.005
morph_area_dens_aabb	-0.49	0.004
morph_vol_dens_aabb	-0.497	0.003
morph_area_dens_conv_hull	-0.533	0.001
	UPDRS total	
morph_comp_1	-0.412	0.017
morph_comp_2	-0.412	0.017
morph_asphericity	0.412	0.017
morph_sph_dispr	0.412	0.017
morph_sphericity	-0.412	0.017
stat_cov	-0.415	0.016
morph_pca_flatness	-0.425	0.014
morph_vol_dens_aee	-0.429	0.013
morph_vol_dens_conv_hull	-0.465	0.006
morph_area_dens_conv_hull	-0.474	0.005
	NMSQ	
stat_skew	-0.31	0.079
ih_skew_fbs_w0.0125	-0.311	0.078
ivh_diff_v25_v75	0.326	0.064
ivh_v50	0.347	0.048
ivh_v25	0.35	0.046
ih_kurt_fbs_w0.0125	-0.368	0.035
stat_kurt	-0.368	0.035
	MMSE	
ivh_v50	-0.305	0.085
ih_qcod_fbs_w0.0125	-0.309	0.08
morph_pca_flatness	-0.336	0.056
stat_qcod	-0.361	0.039
ngt_contrast_3d_fbs_w0.0125	-0.365	0.037
ivh_v75	-0.421	0.015
ivh_v90	-0.503	0.003

The MDS-UPDRS scores were developed to evaluate various aspects of Parkinson's disease, like motor and non-motor experiences of daily living and motor complications. As we can observe from the table III, UPDRS scores have the majority of significant features related to the morphological ra-

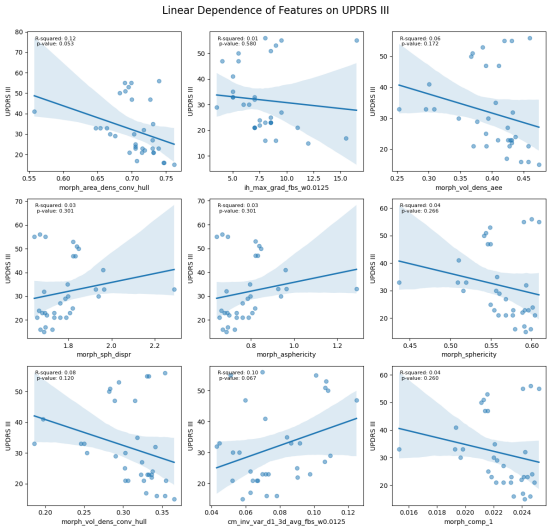

Fig. 4. Linear Dependence of Features selected based on Mutual Information score on UPDRS III.

diomics features. Evidently, numerous studies were conducted on morphological changes in the brain that depict the volume reduction in multiple brain regions like hippocampus [8]–[10], thalamus [8], [11], or posterior putamen nuclei [12]. It is worth noting that the negative correlation trend we observe in morphology radiomics to the severity of clinical scores is directly related to the supporting evidence of previous research.

Mutual information calculations did not depict significant results that would allow us to answer whether the radiomics features are associated with the clinical score severity. Moreover, the results obtained overlapped with the features we got in the Spearman correlation approach, so morphological radiomics features primarily drove the UPDRS scores, and the average mutual information (MI) score varied in the $0.2 - 0.3$ range. The most vital relationship was acquired between the morphological features and UPDRS III clinical score, which refers to assessing a patient's abilities at speech, facial expressions, hand movements, and gait [13], achieving the maximum of $MI_{(i,j)} = 0.43$.

A set of selected features and a linear dependence between each radiomics feature and UPDRS III is depicted in Figure 4. Most of the subplots do not reflect a significant linear dependence between the features and the target variable, meaning a linear relationship is not present between the clinical scores and radiomics features.

Based on the Spearman correlation results and mutual information, there is no relevant evidence of a direct, statistically significant, and strong relationship between radiomics features and clinical scores.

Considering a limited dataset of PD patients and, thus, lack of data points for performing a significant regression on the radiomics features, we have formed subgroups of PD subjects based on the disease severity level. As described in the Methods section, we obtained two groups, one labeled as mild and another as a moderate level of Parkinson's disease.

TABLE IV
MILD AND MODERATE PD SUBGROUPS BASED ON UPDRS SCORES

Feature	Mild mean	Mild std	Moderate mean	Moderate std
LEDD Total	444.10	292.85	733.21	335.74
UPDRS I	7.05	3.73	11.67	5.01
UPDRS II	6.10	3.48	14.83	2.91
UPDRS III	25.19	8.47	43.25	10.78
UPDRS IV	1.33	2.57	4.50	3.64
UPDRS total	39.76	11.84	74.25	16.02
NMSQ	6.62	3.72	10.08	3.40
MMSE	29.38	0.90	29.25	1.01
MoCA	27.62	2.06	26.50	2.22

TABLE V
ANOVA TEST RESULTS ON MILD AND MODERATE PD SUBGROUPS

Feature	p-value
dzm_ldlge_3d_fbs_w0.0125	0.034
morph_sph_dispr	0.039
morph_asphericity	0.039
morph_sphericity	0.04
morph_comp_1	0.041
morph_comp_2	0.044
cm_joint_max_d1_3d_avg_fbs_w0.0125	0.045
cm_inv_diff_mom_norm_d1_3d_avg_fbs_w0.0125	0.047
stat_cov	0.047
cm_inv_diff_norm_d1_3d_avg_fbs_w0.0125	0.055

Table IV provides an overview of the features in the "Mild" and "Moderate" PD subgroups based on UPDRS scores. Using the two subgroups of PD patients, we performed a one-way ANOVA test to distinguish the most discriminant radiomics features and see whether there is a possibility to utilize the features to assign a particular patient to a specific subgroup. The results of the ANOVA test in Table V depict that these radiomics features can be used for determining which stage a Parkinson's disease patient is at (mild/moderate).

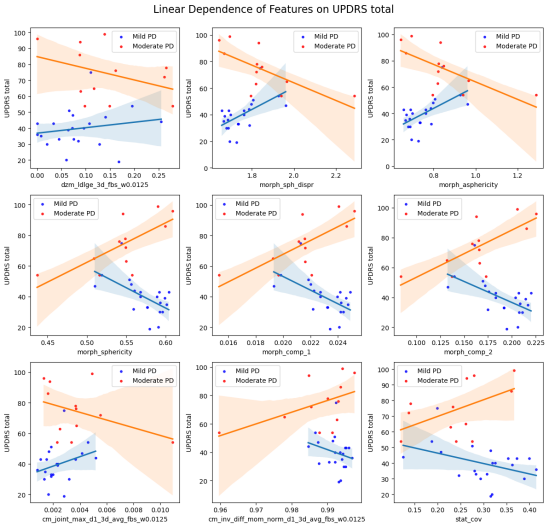

Fig. 5. Linear Dependence of Features selected based on ANOVA test on UPDRS Total.

Figure 5 depicts the differences between the two groups' trends w.r.t. the UPDRS Total score their assessment was granted. For instance, the morphological features of the second row of the plots show that for the mild severity of Parkinson's disease, the higher the value of the radiomics feature, the lower the clinical score that corresponds to it. The opposite happens for the moderate set since the trend is positive.

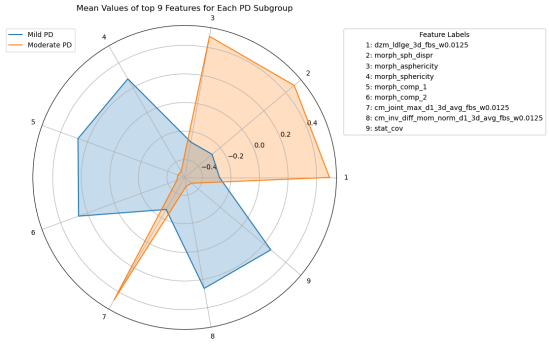

Fig. 6. Mean of top 9 scaled radiomics features for mild and moderate PD subgroups.

To explore the subgroup differences further, in Figure 6, we compare the mean values of the radiomics features for each group. This visualization is a good representation of the differential abilities of the radiomics features in terms of clinical severity groups.

In conclusion, we have found that a specific subset of radiomics features has the potential to identify the levels of Parkinson's disease severity. Based on all of the applied methods, the selected discriminative features primarily consist of morphological features, underlining a relationship between the brain's morphology with the Parkinson's disease severity levels.

3) Machine Learning applications: To address the project question of whether radiomics features can distinguish patients from controls, we employed several machine learning models.

Firstly, SVM-RFE (Support Vector Machine with Recursive Feature Elimination) for feature selection, followed by Random Forest, MKL SVM, and Logistic Regression as the classification algorithms. The SVM-RFE process selected 10 features in each model, resulting in impressive performance outcomes.

When evaluated on the test set, the models maintained a commendable level of performance. The test accuracy was recorded at 91%, indicating that the models could generalize well to unseen data. The sensitivity, which measures the ability to identify positive instances correctly, was 83%. Furthermore, the AUC value remained consistently high at 100%, signifying the models' excellent discrimination ability.

To further validate the robustness of our approach, we replicated the same feature selection method with two additional models, utilizing SVM-RFE with different classification algorithms. These models also selected 10 features and demonstrated exceptional performance on the training data, achieving an average accuracy, sensitivity, and AUC of 100%.

It is worth mentioning that the test accuracy for these models was slightly lower at 89%. However, both models maintained a consistent sensitivity of 89% and achieved an AUC of 96%, indicating their discriminatory solid power.

To explore alternative feature selection techniques, we introduced Lasso regularization combined with Random Forest, MKL SVM (Multiple Kernel Learning Support Vector Machine), and Logistic Regression. These feature selection algorithms selected 12 features and consistently achieved average accuracies, sensitivities, and AUCs of 97%, 100%, and 100%, respectively. The test accuracies, sensitivities, and AUCs for these models were comparable to the previous ones, with values of 91%, 83%, and 100%, respectively.

We further experimented with dimensionality reduction using Principal Component Analysis (PCA). When combined with MKL SVM or Random Forest, PCA produced models with average accuracies and AUCs of 100% and 100%, respectively. The test accuracies, sensitivities, and AUCs were also high, ranging from 83% to 100%. Tables VI and VII summarize the performance of the models above during the training and test phases, respectively.

TABLE VI
PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Avg. Acc. (%)	Avg. Sens. (%)	Avg. AUC (%)
SVM-RFE + RF	97	100	100
SVM-RFE + MKL SVM	100	100	100
SVM-RFE + LR	100	100	100
Lasso + RF	97	100	100
Lasso + MKL SVM	100	100	100
Lasso + LR	100%	100	100
PCA + MKL SVM	97	100	100
PCA + RF	94	100	100
PCA + LR	94	92	100

TABLE VII
TEST PERFORMANCE COMPARISON OF DIFFERENT MODELS

Model	Test Acc. (%)	Test AUC (%)
SVM-RFE + RF	91	100
SVM-RFE + MKL SVM	89	96
SVM-RFE + LR	91	100
Lasso + RF	91	100
Lasso + MKL SVM	91	100
Lasso + LR	91	100
PCA + MKL SVM	83	100
PCA + RF	100	100
PCA + LR	94	100

Based on the tables provided, Model 8, which involved Principal Component Analysis (PCA) combined with Random Forest using K-fold cross-validation, exhibited remarkable performance among these models. This model achieved an average accuracy of 94% and an average sensitivity of 100%, indicating its ability to classify instances from the dataset accurately. The average Area Under the Curve (AUC) value of 100% further affirmed the model's excellent discrimination power. Additionally, during the testing phase, Model 8

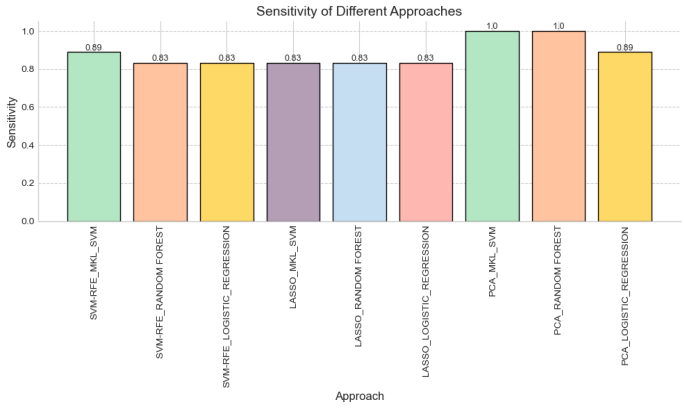

Fig. 7. Test sensitivity of different approaches

achieved perfect accuracy, sensitivity, and especially an AUC score of 100%, as depicted in Figure 7.

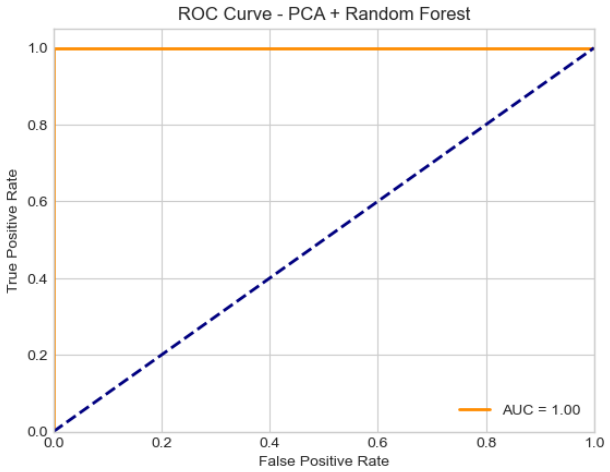

Fig. 8. ROC Curve of PCA and Random Forest model

However, it is essential to acknowledge the challenge posed by the limited dataset available for this study. With only 33 PD patients, the dataset size is relatively small. This limited sample size might affect the findings' generalizability and could introduce bias. Therefore, caution should be exercised when interpreting the results, and further validation with more extensive and diverse datasets is warranted.

In summary, the findings of this study suggest that the combination of PCA and Random Forest (Model 8) holds promise in accurately distinguishing PD patients from healthy controls based on radiomics features. While the limited dataset poses a challenge, this model's high accuracy, sensitivity, and AUC scores make it a strong candidate for further investigation and potential application in clinical settings.

4) Sensitivity analysis: The sensitivity analysis was conducted to evaluate the effect of outliers on the analysis of radiomics data for Parkinson's disease patients and healthy controls. The analysis included the calculation of z-scores and the identification of outliers using a threshold of 3. The results showed three outliers among the Parkinson's disease patients and nine outliers among the healthy controls.

When outliers were included in the analysis, the model demonstrated high average accuracy, sensitivity, and AUC values, with an average sensitivity of 100%. This suggests that the model could correctly identify all or almost all positive instances, even in the presence of outliers. The high accuracy and AUC values indicate the model's effectiveness in distinguishing between instances.

However, the average accuracy and sensitivity slightly decreased when outliers were excluded from the analysis. Although the average sensitivity remained at 100%, indicating that the model correctly identified all or almost all positive instances on average, the test sensitivity decreased to 80%. This suggests that the absence of outliers affected the model's performance in identifying positive instances on the test set.

TABLE VIII
PERFORMANCE COMPARISON

Outliers	Accuracy (%)	Sensitivity (%)	AUC (%)
Including	97	100	100
Removing	93	100	100

TABLE IX
TEST PERFORMANCE COMPARISON

Outliers	Accuracy (%)	Sensitivity (%)	AUC (%)
Including	100	100	100
Removing	92	80	97

The decrease in test sensitivity implies that the model had some difficulty correctly classifying positive instances when outliers were removed. It is important to note that the test accuracy remained relatively high, indicating that the model achieved a high percentage of correct classifications overall. However, the decrease in sensitivity suggests that the model might have been less effective in capturing the subtle patterns or characteristics of positive instances without the influence of outliers.

The AUC values remained consistently high in both cases, indicating that the model had a strong discrimination ability in distinguishing between classes, regardless of the presence or absence of outliers.

In summary, while the model achieved excellent performance, removing outliers resulted in a slight decrease in test sensitivity, suggesting a potential impact on the model's ability to identify positive instances correctly.

In addition, we showed the impact of different class imbalance dealing techniques on the performance of the PCA (Principal Component Analysis) + Random Forest model. Firstly, where no class imbalance dealing technique was applied, the model achieved high average accuracy, sensitivity, and AUC values. The average sensitivity of 100% indicates that the model correctly identified all or almost all of the positive instances on average. The high accuracy and AUC values further suggest the model's effectiveness in distinguishing between instances. The test accuracy, sensitivity, and AUC values of

100% indicate that the model performed exceptionally well on the test set.

In the second case, SMOTE (Synthetic Minority Oversampling Technique) was applied to address the class imbalance. After applying SMOTE, the number of samples for each class became balanced, with 24 samples for both the positive and negative classes. The model achieved an even higher average accuracy and maintained a high average sensitivity and AUC value. The average sensitivity of 97% suggests that the model correctly identified most of the positive instances on average, even with the class imbalance addressed through SMOTE. The test accuracy, sensitivity, and AUC values remained at a perfect score of 100%, indicating excellent performance on the test set.

In the third case, a combination of undersampling and oversampling techniques was used to address the class imbalance. The number of samples for each class was reduced to 11 through undersampling and then increased to 24 through oversampling. The model achieved a perfect average accuracy, sensitivity, and AUC, indicating its ability to classify instances accurately. However, the test accuracy decreased to 89%, suggesting that the model might not generalize as well to unseen data. The test sensitivity also decreased to 78%, indicating a lower ability to identify positive instances in the test set correctly. The test AUC remained high at 98%, suggesting good discrimination ability.

TABLE X
PERFORMANCE COMPARISON

Method	Avg Acc. (%)	Avg Sens. (%)
SMOTE	98	97
Over & Undersampling	100	100

TABLE XI
TEST PERFORMANCE COMPARISON

Method	Test Acc. (%)	Test Sens. (%)
SMOTE	100	100
Over & Undersampling	89	78

The results indicate that applying class imbalance dealing techniques such as SMOTE can further enhance the model's performance, resulting in higher average accuracy and sensitivity. However, the combination of undersampling and oversampling techniques may decrease test accuracy and sensitivity, potentially indicating reduced generalization capability.

In the gender-based subgroup analysis using PCA and Random Forest, the average accuracy for both male and female subgroups was 94%, suggesting that the model could classify instances accurately for both genders on average.

TABLE XII
SENSITIVITY ANALYSIS BASED ON GENDER-BASED SUBGROUPS

Gender Subgroup	**Avg Accuracy (%)**	**Avg Specificity (%)**
Male	94	nan
Female	94	28

However, there was a notable difference in sensitivity between the male and female subgroups. The average sensitivity for males was 97%, indicating that the model correctly identified a high proportion of positive instances in the male subgroup. On the other hand, the average sensitivity for females was much lower at 44%, suggesting that the model had difficulty identifying positive instances in the female subgroup. It is important to investigate the reasons behind this significant difference and explore potential strategies to improve sensitivity in the female subgroup.

It is worth noting that the average specificity value is marked as "nan" (not a number) for the male subgroup, indicating that it was not calculated due to insufficient information provided in the results.

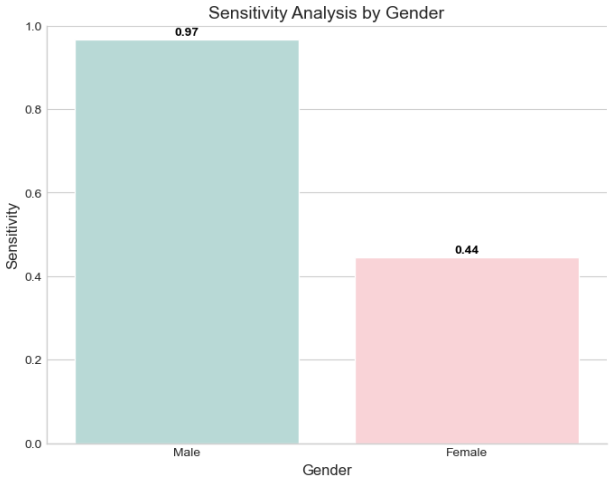

Fig. 9. Sensitivity based on Gender-based Subgroups

The sensitivity values were relatively high for the analysis based on age subgroups, indicating the model's ability to identify positive instances in both subgroups correctly. The AUC value of 100% for the young subgroup suggests excellent discrimination ability. However, the AUC value for the old subgroup was not available.

TABLE XIII
SENSITIVITY ANALYSIS BASED ON AGE SUBGROUPS

Age Subgroup	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
Old subgroup	90	90	90	nan
Young subgroup	91	95	85	nan

In the subgroup analysis based on BMI, no samples were available in the underweight subgroup, so the analysis was skipped. For the healthy range subgroup, the model achieved a high accuracy of 95% and a sensitivity of 100%, indicating its effectiveness in correctly classifying instances in this subgroup. The AUC value of 100% further confirms the model's excellent discrimination ability. In the overweight/obesity subgroup, the model achieved a slightly lower accuracy of 93% and a sensitivity of 85%. However, the AUC value remained high at 100%, suggesting strong discrimination capability.

These results highlight the importance of considering different subgroups when evaluating the model's performance.

TABLE XIV
SENSITIVITY ANALYSIS - SUBGROUP ANALYSIS BASED ON BMI

BMI Subgroup	Samples	Accuracy (%)	Sensitivity (%)
Healthy Range	24	95	100
Overweight/Obesity	29	93	85

The model performs better for males than females, indicating a potential gender-related difference in the model's predictive capability. Additionally, the model exhibits variations in performance across different age and BMI subgroups, with higher sensitivities observed in specific subgroups.

IV. DISCUSSION

Our study has demonstrated, through statistical analysis and correlation assessment, the potential to distinguish between PD subjects and healthy controls and evaluate the possible existence of a relationship between specific values of radiomics features and the severity level of the disease. This research contributes to the emerging theory that links radiomics features and DAT SPECT for an improved and precise diagnosis of Parkinson's disease.

One challenge encountered during the analysis and implementation of this study was working with a limited dataset size (a total of 53 individuals, including 33 PD patients and 20 healthy controls). Undoubtedly, having a more extensive dataset would have enabled a more precise statistical analysis and correlation assessment and the potential implementation of deep learning models instead of traditional machine learning models, which could have led to improved performance.

This limitation has compelled us to consider all subjects within our dataset without excluding potential outliers, which may alter our statistical analysis and correlations and introduce potential bias in our ML models. Given a significantly larger dataset, this step could have been included in our pipeline, undoubtedly leading to a more accurate result.

The results demonstrate the potential of radiomics features combined with machine learning models in distinguishing Parkinson's disease (PD) patients from healthy controls. Model 8, which utilized Principal Component Analysis (PCA) and Random Forest, exhibited remarkable performance with high accuracy, sensitivity, and AUC scores. The results suggest that the combination of feature selection and dimensionality reduction techniques can effectively identify relevant features and reduce the complexity of the dataset while maintaining strong predictive accuracy. However, as aforementioned, the limited sample size raises concerns about the generalizability of our findings and the potential for bias. To reduce these limitations, further validation with larger datasets is necessary. Furthermore, exploring the reproducibility and generalizability of our findings would be beneficial by applying the developed model to independent datasets or conducting external validation studies.

Additionally, the sensitivity analysis showed impressive performance across different gender, age, and BMI subgroups, with higher sensitivities observed in specific subgroups. E.g., understanding the reasons behind the lower sensitivity in the female subgroup requires further investigation. Factors such as differences in the dataset's characteristics, the distribution of Parkinson's disease among males and females, or potential confounding variables associated with gender may play a role. It would be beneficial to analyze and consider additional factors that could contribute to the observed disparity in sensitivity between genders. In addition, in the age-based subgroup analysis, the model demonstrated an AUC value of 100% for the young subgroup. However, further information regarding the AUC value for the old subgroup was not available, and additional analysis is needed to obtain a complete understanding of the model's performance in that subgroup.

Moreover, our results demonstrate model performance variations across different BMI subgroups. Overall, higher sensitivities were observed in certain subgroups, indicating that age and BMI might influence the model's predictive capability. These findings suggest that accounting for these factors and potentially incorporating age and BMI-specific features could enhance the accuracy and sensitivity of the model, particularly in subgroups where sensitivity is relatively lower.

Further investigation and potential adjustment of the model or dataset could be warranted to improve performance in subgroups with comparatively lower sensitivity.

In this study, we aimed to outline a simple pipeline with a robust scientific methodology. We opted to utilize authentic medical features and dimensionality reduction techniques such as Principal Component Analysis (PCA), which yielded superior performance at the association between medical features and the medical diagnosis of a disease. Our study highlights the potential of radiomics features combined with machine learning models for distinguishing Parkinson's disease (PD) patients from healthy controls. The statistical analysis and correlation assessment revealed a relationship between specific radiomics values and the severity of the disease, supporting the emerging theory of utilizing radiomics features and DAT SPECT for improved diagnosis of Parkinson's disease. Although the limited dataset size posed challenges, further validation with larger datasets and exploration of subgroup-specific factors could enhance the accuracy and sensitivity of the models.

REFERENCES

[1] Rahmim, A., Salimpour, Y., Jain, S., Blinder, S. A., Klyuzhin, I. S., Smith, G. S., Mari, Z., Sossi, V. (2016). Application of texture analysis to DAT SPECT imaging: Relationship to clinical assessments. NeuroImage: Clinical, 12, e1-e9. https://doi.org/10.1016/j.nicl.2016.02.012

[2] Shiiba, T., Takano, K., Takaki, A., Suwazono, S. (2022). Dopamine transporter single-photon emission computed tomography-derived radiomics signature for detecting Parkinson's disease. EJNMMI Research, 12(1), 1-12. https://doi.org/10.1186/s13550-022-00910-1

[3] Zwanenburg, A., Leger, S., Vallières, M., Löck, S. (2016). Image biomarker standardisation initiative. ArXiv. https://doi.org/10.1148/radiol.2020191145

[4] Hatt, M., Cheze Le Rest, C., Antonorsi, N., Tixier, F., Tankyevych, O., Jaouen, V., Lucia, F., Bourbonne, V., Schick, U., Badic, B., Visvikis, D. (2021). Radiomics in PET/CT: Current Status and Future AI-Based Evolutions. Seminars in Nuclear Medicine, 51(2), 126-133. https://doi.org/10.1053/j.semnuclmed.2020.09.002

[5] Vergara, Jorge Estevez, Pablo. (2014). A Review of Feature Selection Methods Based on Mutual Information. Neural Computing and Applications. 24. 10.1007/s00521-013-1368-0.

[6] Pablo Martínez-Martín, Carmen Rodríguez-Blázquez, Mario Alvarez, Tomoko Arakaki, Víctor Campos Arillo, Pedro Chaná, William Fernández, Nélida Garretto, Juan Carlos Martínez-Castrillo, Mayela Rodríguez-Violante, Marcos Serrano-Dueñas, Diego Ballesteros, Jose Manuel Rojo-Abuin, Kallol Ray Chaudhuri, Marcelo Merello, Parkinson's disease severity levels and MDS-Unified Parkinson's Disease Rating Scale, Parkinsonism Related Disorders, Volume 21, Issue 1, 2015, Pages 50-54, ISSN 1353-8020, https://doi.org/10.1016/j.parkreldis.2014.10.026.

[7] Goetz, C.G., Tilley, B.C., Shaftman, S.R., Stebbins, G.T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M.B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A.E., Lees, A., Leurgans, S., LeWitt, P.A., Nyenhuis, D., Olanow, C.W., Rascol, O., Schrag, A., Teresi, J.A., van Hilten, J.J. and LaPelle, N. (2008), Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. Mov. Disord., 23: 2129-2170. https://doi.org/10.1002/mds.22340

[8] Devignes Q, Viard R, Betrouni N, Carey G, Kuchcinski G, Defebvre L, Leentjens AFG, Lopes R, Dujardin K. Posterior Cortical Cognitive Deficits Are Associated With Structural Brain Alterations in Mild Cognitive Impairment in Parkinson's Disease. Front Aging Neurosci . 2021;13:668559

[9] Wilson H, Niccolini F, Pellicano C, Politis M. Cortical thinning across Parkinson's disease stages and clinical correlates. J Neurol Sci . 2019;398:31–38

[10] van Mierlo TJ, Chung C, Foncke EM, Berendse HW, van den Heuvel OA. Depressive symptoms in Parkinson's disease are related to decreased hippocampus and amygdala volume. Mov Disord . 2015;30:245–252.

[11] Garg A, Appel-Cresswell S, Popuri K, McKeown MJ, Beg MF. Morphological alterations in the caudate, putamen, pallidum, and thalamus in Parkinson's disease. Front Neurosci . 2015;9:101.

[12] Nemmi F, Sabatini U, Rascol O, Péran P. Parkinson's disease and local atrophy in subcortical nuclei: insight from shape analysis. Neurobiol Aging . 2015;36:424–433.

[13] Huang ML, Hung YH, Lee WM, Li RK, Jiang BR. SVM-RFE based feature selection and Taguchi parameters optimization for multiclass SVM classifier. ScientificWorldJournal. 2014;2014:795624. doi: 10.1155/2014/795624. Epub 2014 Sep 10. PMID: 25295306; PMCID: PMC4175386.

[14] Wu L, Wang C, Tan X, Cheng Z, Zhao K, Yan L, Liang Y, Liu Z, Liang C. Radiomics approach for preoperative identification of stages I-II and III-IV of esophageal cancer. Chin J Cancer Res. 2018 Aug;30(4):396-405. doi: 10.21147/j.issn.1000-9604.2018.04.02. PMID: 30210219; PMCID: PMC6129566.

[15] Jolliffe Ian T. and Cadima Jorge 2016Principal component analysis: a review and recent developmentsPhil. Trans. R. Soc. A.3742015020220150202 http://doi.org/10.1098/rsta.2015.0202

[16] Thabane, L., Mbuagbaw, L., Zhang, S. et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC Med Res Methodol 13, 92 (2013). https://doi.org/10.1186/1471-2288-13-92

[17] Zhang Y, Oikonomou A, Wong A, Haider MA, Khalvati F. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. Sci Rep. 2017 Apr 18;7:46349. doi: 10.1038/srep46349. PMID: 28418006; PMCID: PMC5394465.

[18] Lv J, Chen X, Liu X, Du D, Lv W, Lu L and Wu H (2022) Imbalanced Data Correction Based PET/CT Radiomics Model for Predicting Lymph Node Metastasis in Clinical Stage T1 Lung Adenocarcinoma. Front. Oncol. 12:788968. doi: 10.3389/fonc.2022.788968