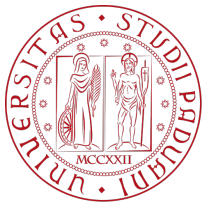

University of Padova, 11th June 2023
Homework

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

Andrea Bertola[1], Prateek Gupta[2], Nadine La Salvia[3], Mariasole Pasinato[4]

Abstract
Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by dopaminergic denervation and neuronal loss. Current diagnostic approaches primarily rely on visual inspection of DAT-SPECT images and clinical assessment of motor and non-motor symptoms, calling for more objective and sensitive evaluation methods. Radiomics analysis offers a promising avenue to address this need. In this cross-sectional study, we analyzed a dataset comprising 177 radiomics features extracted from DAT-SPECT images of the putamen region in 33 PD patients and 20 matched healthy controls (HC). Demographic data were available for both groups, while clinical assessment was limited to the PD group. Our objectives were to investigate significant differences in radiomic features between PD and HC, explore potential correlations between radiomics and clinical scales, and evaluate the classification performance of radiomics using machine learning methods. Utilizing a multiple-step feature selection procedure, we identified four key features of interest. Statistical comparisons revealed significant differences ($p < 0.0001$) between PD and HC for all four features using t-test and Mann-Whitney tests. In terms of correlation analysis with clinical scales, only one feature demonstrated a significant association ($p < 0.05$) specifically with non-motor symptoms scales, namely UPDRS (part I) and MMSE. For classification, Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and Logistic Regression (LR) were employed. SVM and LR yielded an accuracy of 94%, ROC-AUC of 0.95, and Matthews' Correlation Coefficient (MCC) of 0.88, while LDA achieved an accuracy of 100%, ROC-AUC of 1.0, and MCC of 1.0. Despite the limitations stemming from a small dataset, our findings underscore the potential of radiomic features in detecting PD-related pathophysiological changes on DAT-SPECT scans, suggesting their promise as biomarkers for PD detection.

Keywords
Radiomics — Parkinson's Disease — DAT SPECT

[1]Department of Information Engineering, *Bioingeneering for Neuroscience*, ID: 2055557, andrea.bertola@studenti.unipd.it
[2]Department of Information Engineering, *ICT for Life and Health*, ID: 2080676, prateek.gupta@studenti.unipd.it
[3]Department of Information Engineering, *Bioingeneering for Neuroscience*, ID: 2086866, nadine.lasalvia@studenti.unipd.it
[4]Department of Information Engineering, *Bioingeneering for Neuroscience*, ID: 2087527, mariasole.pasinato@studenti.unipd.it

Contents

1. Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disease characterised by various motor and non-motor symptoms [1, 2]. The pathological mechanism that correlates with symptoms is the neuronal loss in the substantia nigra with dopaminergic denervation of the striatum, accompanied by the formation of Lewy bodies in the degenerating neuron [3]. There has been a rapid rise in the prevalence of PD, leading to increased global burden in terms of death and disability, which has more than doubled over the last two decades [4, 5, 6]. This results in a need for a reliable biomarker for the early diagnosis, prognosis, and monitoring of the disease.

Currently, medical imaging modalities like magnetic resonance imaging (MRI), positron emission tomography (PET), and single photon emission computed tomography (SPECT) have been found effective and increasingly being used in the early diagnosis of PD [7]. Visual inspection of DAT-SPECT images has been the most common approach in the diagnosis of PD, but a more objective assessment and quantitative analysis may be more sensitive to early detection of the disease [8, 9].

Radiomics analysis [10] is increasingly being applied to imaging techniques to potentially provide more quantitative methods for the detection and tracking of disease progression. Specifically, there is a growing body of literature focusing on the application of radiomics analysis to DAT-SPECT imaging, leveraging its potential in this domain.

Rahmim et al. found significant correlations between textural radiomic features and clinical measures, such as disease rating scales and disease duration [9]. More recently, Rahmim et al. showed that radiomics analysis of longitudinal DAT-SPECT images, performed by correlating textural features with motor and cognitive functions, results in an improved prediction of PD outcome [11]. However, these studies have primarily focused on assessing textural features, overlooking other classes of radiomic features.

Shiiba et al. showed that a radiomics signature from DAT-SPECT-derived radiomic features can successfully discriminate PD patients from healthy individuals and classification performances significantly improve when radiomics features are employed in addition to more traditional semi-quantitative indicators such as SUR [12].

In this study, we want to assess both the discriminative power of radiomics and their association with PD symptoms severity. In particular, we first aim to investigate whether significant differences in radiomic features can be observed comparing a PD and a healthy control group. Considering the relevant results shown by Shiiba et al. [12] employing radiomics for classification, we expect them to perform well even in group comparison. We then propose to check for any correlation between radiomic features and a pool of clinical scales for the evaluation of the disease symptoms, similarly to the work of Rahmim et al. [9]. Finally, we'll investigate the performance of radiomics in the individual classification of PD patients and healthy controls. From the previously cited work of Shiiba et al. [12], we hypothesize that, even starting from a large set of radiomics features, a relatively small subset of properly selected ones will be enough to achieve significant performances in classification.

2. Materials and Methods

2.1 Dataset

The dataset used in this report and all the details presented in this section, unless otherwise explicitly specified, have been derived from the PET NODE REPOSITORY available at King's College London.

2.1.1 Sample characteristics

The dataset consists of a set of radiomics features derived from DAT SPECT imaging scans acquired from 33 Parkinson's disease (PD) patients and 20 healthy controls (HC), matched for age and gender.

In addition to the imaging assessment, all the patients (both PD and HC) underwent a clinical visit. Clinical scores, computed using the clinical scales described in *Appendix A*, were reported only for PD patients.

Subjects were chosen among males and females, between 30 and 85 years old (inclusive). Enrolled subjects for the PD group were required to have idiopathic PD according to the Movement Disorder Society Clinical Dignostic Criteria and to be classified between Stage 1 to 3 inclusive on the modified Hoehn and Yahr scale for PD severity. HCs were selected among non-demented subjects (i.e. score higher than 25 points in the Montreal Clinical Assessment).

2.1.2 Imaging assessment

A SPECT imaging scan was performed for all the participants using [^{123}I]-FP-CIT as the radioligand, targeting the dopamine transporter system (SPECT DAT scan).

The imaging data pipeline included the segmentation of brain anatomical regions using the CIC v2.0 anatomical atlas and co-registration to the subject's structural MRI. VOI analyses were performed on putamen (both left and right), with cerebellum grey matter as a reference region. Standardise Uptake Value Ratio (SUVR) was used as the main parameter of interest and was calculated as the ratio of the putamen VOI count density divided by the cerebellum count density. This measure approximates the binding potential when the radioligand is in equilibrium at the target site.

Radiomics features were extracted for each subject using the MIRP Python package.

2.2 Feature selection

The most critical issue with radiomics-based analyses is the dimension of data: that is, radiomics datasets often have fewer samples than features. In these cases, it is particularly important to extract meaningful features, in order to allow a faster computation and to obtain a clearer view of the underlying processes producing a given output [13]. Many feature selection methods were proposed over time, but still, there is no

clear solution on which works best to remove redundant and irrelevant features in radiomics context [14]. In this study, we decided to apply a multiple-step feature selection. Similarly to the works in [15, 16, 17], firstly, a filter method was exploited to greatly reduce the number of features on the basis of their scores in a specified statistical test; then a more complex algorithm was employed to fine-tune the selection of the most representative features.

We employed Python as the primary programming language for implementing the feature selection process. The radiomics dataset and the response variable (vector containing the HC/PD classification for each subject) were divided into a training set and a test set using a random split (70% training set and 30% test set), while ensuring that the HC/PD ratio of the whole dataset was maintained in both the subsets. We performed the feature selection procedure only on the training set, in order to keep the test set completely hidden from our classifier.

To assess the relationship between each feature in the training set and the response variable, Spearman's correlation coefficient was calculated and used for the filtering step [17]. Spearman's correlation is a non-parametric measure that captures monotonic relationships between variables. Furthermore, to address the issue of multiple comparisons inherent in evaluating a large number of features, an adjusted p-value for a 0.05 significance level was computed using the False Discovery Rate (FDR) correction [9]. FDR correction adjusts the significance threshold to control for the expected proportion of false positives [18]. We retained only those features that demonstrated a statistically significant association with the response variable, as determined by the adjusted p-value. Additionally, to ensure a strong correlation between the selected features and the response variable, we set a threshold of 0.6 for the correlation coefficient. Any features that possessed both a significant adjusted p-value and a correlation value exceeding 0.6 were retained for further analysis.

After standardizing the data, we applied logistic regression with L^1 penalty (that is, LASSO generalized to GLMs with binomial distribution for classification problems) on the subset of variables obtained from the previous step. We used $C = 1/\lambda = 10$ as regularization parameter. This step allowed us to identify a subset of features that are most predictive for the response variable, while simultaneously reducing the impact of irrelevant or redundant variables.

Finally, to address potential residual redundancy among the selected features, we conducted an additional analysis to examine Spearman's correlation between the retained variables. By calculating pairwise correlation coefficients, we aimed to identify and remove any redundant features that exhibited high correlation with each other. We checked for correlation values above 0.9 and significant p-values (significance level 0.05). This step ensured that the final set of features captured distinct and independent information.

By following these rigorous feature selection procedures, we aimed to identify a subset of features that are statistically significant, highly correlated with the response variable and not redundant. The complete pipeline for feature selection is summarized in Figure 1.

2.3 Univariate analysis

All statistical analyses and data processing for this section were performed using MATLAB. Data from all subjects were used to perform the following procedures, not only from the training set.

2.3.1 Group matching analysis

According to the data collection protocol, subjects with Parkinson's disease (PD) and healthy controls (HC) were matched for age and gender. To ensure the validity of this matching and assess the comparability of the two groups, we checked the results for age and gender and investigated additional demographic (education years) and anthropometric (height, weight, BMI) information.

For the numerical variables (i.e. excluding gender), we first examined the gaussianity assumption using the Shapiro-Wilk test. Gaussian variables were assessed for group matching using the independent samples t-test, with a significance level set at 0.05. Non-Gaussian variables were analyzed using the non-parametric Wilcoxon-Mann-Whitney test, also with a significance level of 0.05.

For the categorical variable gender, we employed the Chi-squared test to evaluate whether there were significant differences in the ratios of males and females between the two groups, again using a significance level of 0.05.

2.3.2 Comparison between PD and control groups

To investigate the differences between the healthy controls and PD patients, we used the subset of variables selected in the previous feature selection step.

First, we assessed the gaussianity assumption of the selected variables using the Shapiro-Wilk test at a significance level of 0.05 [19].

Next, we examined the correlation between the selected variables and the demographic and anthropometric information of the subjects (age, gender, education years, weight, BMI), using Spearman's correlation coefficient; we used FDR to adjust the p-values at 0.05 significance level and account for multiple comparisons.

For the variables that followed a gaussian distribution, we performed an independent samples t-test to compare the means between the HCs and PD group [19]. We used a significance level of 0.05 for detecting statistically significant differences. For the variables that did not exhibit a gaussian distribution, we utilized the non-parametric Wilcoxon-Mann-Whitney test for group comparison. We set the significance level at 0.05 for assessing significant differences [12].

2.3.3 Correlation analysis with clinical scales

To assess the relationship between the selected radiomic features and clinical scales relevant to PD, we conducted a correlation analysis using the Pearson correlation coefficient. We focused on examining the correlation between each selected

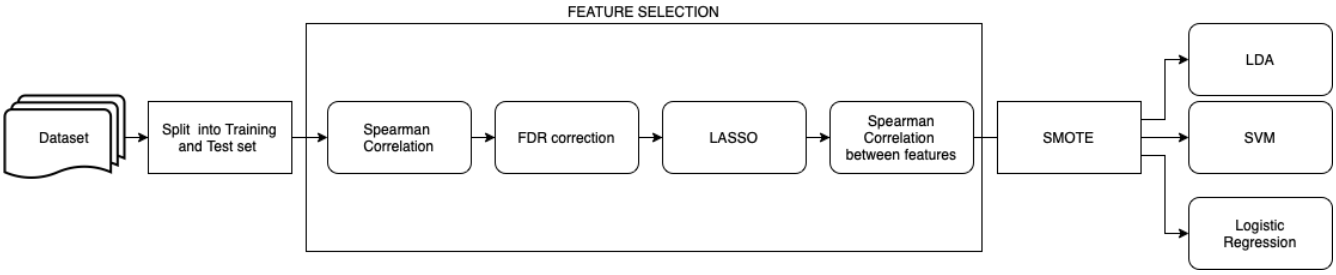

Figure 1. Pipeline for feature selection procedure and classification.

feature and the clinical scales, which are described in the *Appendix A* of this article. For each feature, we calculated Pearson correlation coefficient for each clinical scale, with a significance level of 0.05. To account for multiple comparisons, we employed FDR correction on the p-values obtained from the correlation analysis. The same procedure for correlation analysis between radiomics and clinical scores was employed in the study by Rahmim et al. [9].

2.4 Individual classification

Given the small number of subjects and selected features in our dataset, we decided to use simple and linear classifiers to ensure generalization and avoid overfitting [20]. The classifiers were trained using the same training set that was employed for the feature selection procedure.

To address the potential issue of imbalanced data, we applied the Synthetic Minority Over-sampling Technique (SMOTE) algorithm to oversample the training set and balance the two groups. This step is recommended to mitigate biases that could arise from an unbalanced dataset [21].

We selected three well-known linear classifiers for our analysis: Support Vector Machine with a linear kernel (SVM), Linear Discriminant Analysis (LDA) [22], and Logistic Regression (LR). These classifiers have been widely used in various classification tasks and have demonstrated effectiveness with linearly separable data. To optimize the performance of the classifiers, we performed 10-fold cross-validation. Within each fold, we employed a grid search to find the optimal hyperparameters for each classifier.

To evaluate the prediction performance of the classifiers on the test set, we used several well-established metrics. First, we calculated the accuracy, which measures the proportion of correctly classified instances over the total number of instances. Additionally, we assessed the confusion matrix, which provides a detailed breakdown of the classifier's performance in terms of true positives, false negatives, true negatives, and false positives. Furthermore, we plotted Receiver Operating Characteristic (ROC) curves, which illustrate the trade-off between the true positive rate and false positive rate for different classification thresholds. ROC curves are widely used to assess the performance of binary classifiers and provide a visual representation of their discriminative power. In addition to the aforementioned metrics, we employed Matthews' correlation coefficient (MCC), which evaluates the overall performance of a classifier by taking into account all four categories of the confusion matrix (true positives, false negatives, true negatives, and false positives). MCC is particularly useful when dealing with imbalanced datasets (like our test set, on which SMOTE algorithm was not applied), as it considers the proportion of both positive and negative elements in the dataset. It ranges in the interval [-1, +1], with -1 representing perfect misclassification and +1 indicating perfect classification; $MCC = 0$ is the expected value for the chance classifier [23, 24].

All the Machine Learning analysis was performed using Python.

3. Results

3.1 Selected features

The dataset contained 177 radiomic features. No data were missing from the dataset, so none of the features were excluded due to data quality issues. One feature was discarded in the first step of the feature selection process, for it had zero variance (constant feature) and therefore wouldn't be useful for the following analyses. The filtering step using Spearman's correlation left us with 132 features, showing that the vast majority of the radiomics were significantly ($p_{adjusted} < 0.05$) and highly correlated with the response vector. The penalized logistic regression reduced the subset of significant features to 9. We managed to further reduce the set of features by checking for the Spearman's correlation values between them as described in the *Feature selection* section. At the end of the whole procedure, four features remained:

1. *stat_skew*, skewness of the intensity histogram of the image;
2. *stat_qcod*, measure of the dispersion of the grey levels;
3. *ivh_v75*, volume fraction at 75% intensity, i.e. largest volume fraction that contains a grey level which has an intensity of at least 75% of the maximum grey level value present in the image;
4. *ivh_diff_v25_v75*, difference between volume fraction at 25% intensity and 75% intensity.

We noticed that, even slightly changing some of the parameters of the procedure, these four features were often selected, indicating an encouraging robustness of the process.

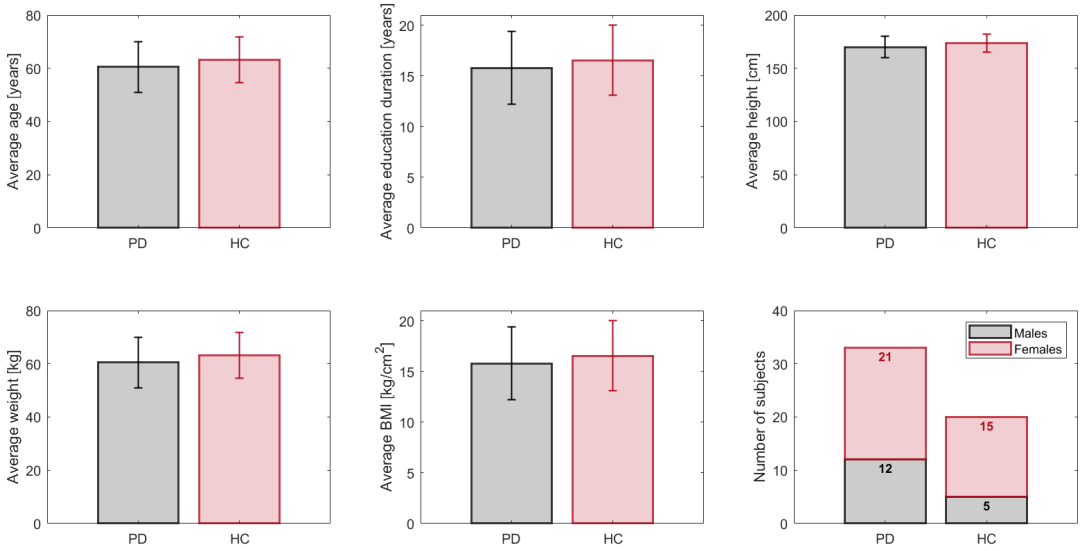

Figure 2. Group matching for demographic and anthropometric features. From top-left to bottom-right, age, education years, height, weight, BMI, and gender. None of the comparisons showed significant p-values when testing differences between the two groups.

3.2 Univariate analysis

3.2.1 Group matching analysis

Among the numerical features, age, education years, and height were found to follow a gaussian distribution based on the Shapiro-Wilk test. Therefore, we employed the independent samples t-test to compare the means between the PD and HC groups for these variables. None of the features showed significant differences: age ($t = -0.982$, $p = 0.331$), education years ($t = -0.761$, $p = 0.450$), and height ($t = -1.306$, $p = 0.197$).

Weight and BMI distributions from our data exhibited a significant departure from gaussianity. Consequently, we utilized the non-parametric Mann-Whitney test to assess the differences between the PD and HC groups for these variables. Similarly to the gaussian variables, no significant differences were observed between the groups: weight ($U = 832.5$, $p = 0.287$) and BMI ($U = 863$, $p = 0.614$).

Furthermore, the matching analysis confirmed that the PD and HC groups were well-matched in terms of gender. The Chi-squared test revealed no significant differences in gender ratios between the groups ($\chi^2 = 0.738$, $p = 0.390$).

Collectively, these results demonstrate that the PD and HC groups were well-matched for all the demographic and anthropometric features (Figure 2).

3.2.2 Comparison between PD and control groups

Testing the variables for gaussianity using the Shapiro-Wilk test, only *stat_qcod* was found to follow a gaussian distribution. The remaining variables did not meet the assumption of normality.

In the correlation analysis between the radiomic features and the demographic and anthropometric features, no significant correlations were observed. Therefore, no covariates were included in the subsequent group comparison analysis.

For the one variable that followed a Gaussian distribution (*stat_qcod*), the t-test revealed a significantly lower magnitude in HC than in PD ($t = -8.02$, $p \approx 10^{-10}$). For the remaining variables that did not exhibit a gaussian distribution, the Wilcoxon-Mann-Whitney test indicated significant differences between the groups for all of them. In particular, *stat_skew* was found to be significantly lower in PD ($U = 219$, $p \approx 10^{-9}$), while both *ivh_v75* and *ivh_diff_v25_v75* showed a significantly lower magnitude in HC (respectively, $U = 817$, $p \approx 10^{-7}$; $U = 844$, $p \approx 10^{-8}$). Results are shown in Figure 3.

3.2.3 Correlation analysis with clinical scales

After adjusting for multiple comparisons, significant correlations were observed for one specific feature, *ivh_v75*. The *ivh_v75* feature demonstrated a significant positive correlation with UPDRS I ($r = 0.4747$, $p_{adj} = 0.03$), indicating that higher values of *ivh_v75* were associated with increased severity of UPDRS I scores (Figure 4). Additionally, a significant negative correlation was found between *ivh_v75* and MMSE ($r = -0.4591$, $p_{adj} = 0.03$), suggesting that higher values of *ivh_v75* were associated with lower scores on the MMSE, indicating cognitive impairment (Figure 5).

3.3 Individual classification

All three classifiers reached a training score of 0.98. Regarding the test set, the SVM and the LR models exhibited an accuracy rate of 94%, with only one misclassified subject out of the total 16 (resulting in a false negative). The ROC curve yielded an area under the curve (AUC) value of 0.95, indicating a high discriminatory power. Furthermore, MCC reached a value of 0.88, reflecting strong overall classification performance. LDA model achieved an accuracy of 100%,

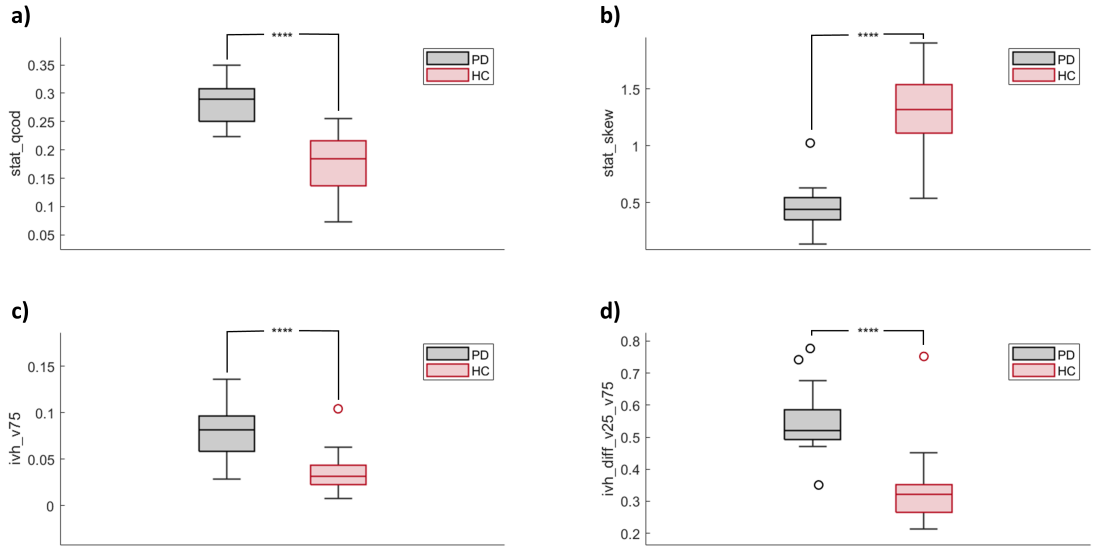

Figure 3. Group differences between PD and HC in selected radiomic features. a) *stat_qcod*, $p \approx 10^{-10}$; b) *stat_skew*, $p \approx 10^{-9}$; c) *ivh_v75*, $p \approx 10^{-7}$; d) *ivh_diff_v25_v75*, $p \approx 10^{-8}$. Four asterisks indicate significant differences with $p < 0.0001$.

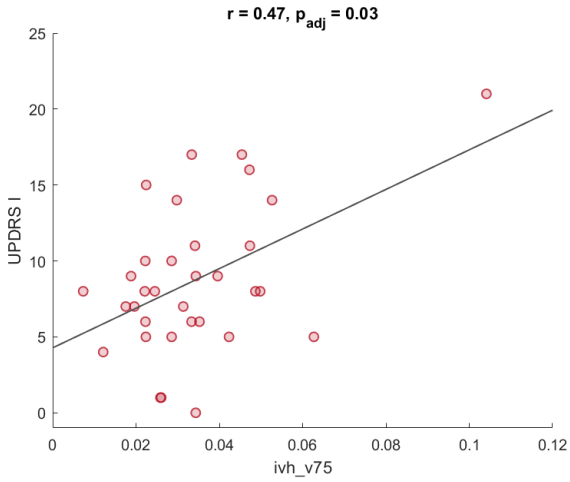

Figure 4. UPDRS I vs. *ivh_v75*. Pearson's correlation coefficient is 0.4747 and it was found to be significant ($p_{adj} < 0.05$).

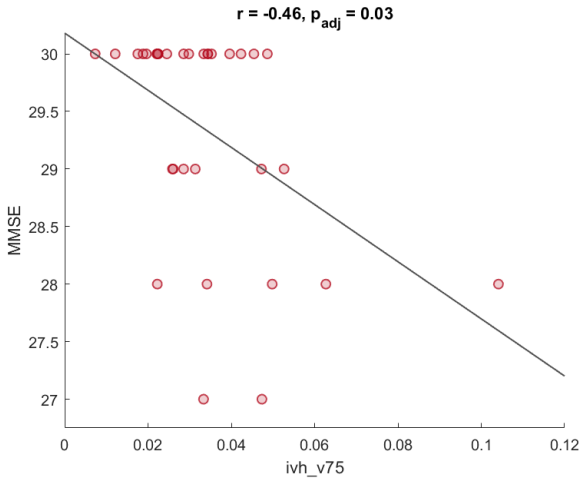

Figure 5. MMSE vs. *ivh_v75*. Pearson's correlation coefficient is -0.4591 and it was found to be significant ($p_{adj} < 0.05$).

correctly classifying all subjects in the test set. Table 1 provides a comprehensive summary of the classification results obtained using the three classifiers, while Figure 6 depicts the corresponding ROC curves.

4. Discussion

The present study aimed to investigate the potential of radiomic features extracted from DAT-SPECT scans to detect pathological changes associated with Parkinson's disease (PD) compared to healthy controls (HC). Our results demonstrated significant differences in the selected radiomic features between the two groups. Specifically, the analysis of the features *stat_qcod* and *stat_skew* revealed interesting differences between PD patients and healthy controls. PD patients exhibited higher dispersion of the histogram, as indicated by the increased values of *stat_qcod*. In contrast, healthy controls showed a less dispersed histogram and a higher skewness, suggesting that their grey levels tended to gather in a narrower section and lean towards one of the two extremes of the range. These findings are in line with previous studies that have highlighted the alterations in DAT-SPECT images associated with PD. In Rahmim et al. [9], increased UPDRS scores were associated with significantly lower values of homogeneity and higher values of dissimilarity of DAT-SPECT images, confirming the relationship between grey levels dispersion and PD. Increased dispersion in the histogram of PD patients may

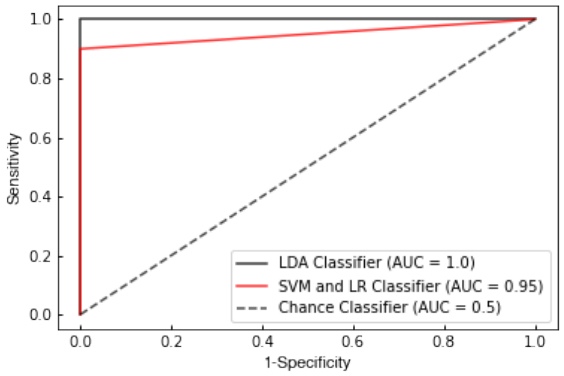

Figure 6. Receiver operating curves for SVM, LDA and LR models trained on selected radiomic features for the classification of PD and HC. SVM and LR models exhibit an AUC value of 0.95; LDA results in an AUC value of 1.0.

Table 1. Classification performance of various classification models using selected radiomic features for test set.

Classifier	ROC-AUC	Accuracy	MCC
SVM	0.95	0.94	0.88
LDA	1.0	1.0	1.0
Logistic Regression	0.95	0.94	0.88

reflect the loss of dopaminergic neurons in the basal ganglia, leading to reduced dopamine uptake and subsequent variations in radiotracer distribution. The higher skewness observed in healthy controls may indicate a more concentrated distribution of grey levels, suggesting preserved dopaminergic function in these individuals [3].

Furthermore, we examined the correlation between the selected radiomic features and the clinical scales to explore their potential association with disease severity and symptomatology. Among the features analyzed, only *ivh_v75* demonstrated significant correlations with clinical scales, specifically UPDRS I and MMSE. Consistently, higher values of *ivh_v75* were associated with higher scores in UPDRS I, indicating a worsening of non-motor symptoms, and lower scores in MMSE, indicating more severe cognitive impairment. Interestingly, we did not find any significant correlations between the radiomic features and scales related to motor symptoms. This may suggest that the feature captures specific aspects of non-motor symptoms and cognitive impairments in PD.

In our classification analysis, we employed three well-known classifiers (SVM, LDA, and LR) which achieved accuracy rates of 94%, 100%, and 94% respectively. These high accuracy values may be attributed to the inherent separability of the data, which is evident from the scatter plots (Figure 7) depicting the relationships between pairs of features. Even with the use of just two features, the data points appear to be easily and linearly separable. Considering that we trained our classifiers using four features, this greater feature space likely contributed to the high accuracy observed. In other words, our models are probably more complex than our dataset, leading to good classification performance. However, these results should not be generalized unless the same steps are validated on a bigger dataset.

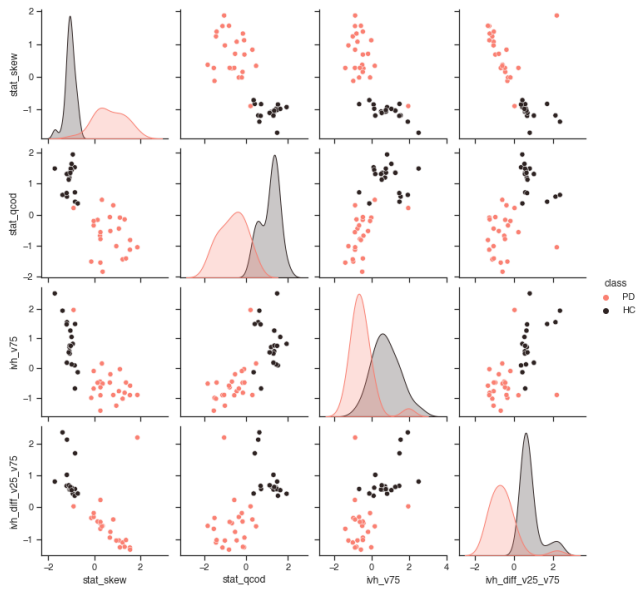

Figure 7. Scatter plots and distributions of selected radiomic features for the training set after applying SMOTE. On the diagonal, the distribution of each of the radiomics is shown, distinguishing PD (red) from HC (black). The out-of-diagonal charts are pairwise scatter plots of the features.

4.1 Limitations and future work

The utilization of radiomic features in general is not without its inherent limitations. Previous studies have highlighted the issue of feature instability: radiomic features can exhibit fluctuations even between imaging scans acquired within a short time frame [25]. Furthermore, the susceptibility of radiomic features to variation across different scanners and imaging protocols introduces substantial variability between centers and hospitals [26]. These challenges pose potential limitations to the widespread adoption of radiomics as a tool for objective and unbiased diagnosis, especially when compared to semi-quantitative indices or clinical scales. To address these concerns, future research on radiomic features should emphasize large-scale validation on diverse and adequately sized datasets collected from multiple sources. This will help assess and account for the aforementioned sources of variability, promoting robust and reliable application of radiomics in clinical practice.

Expanding the dataset size would also have the potential to enhance the generalizability of the generated models. As previously mentioned, a small dataset may not adequately

capture the full spectrum of population variability, resulting in a well-performing classifier that may exhibit limited effectiveness when confronted with new data. To gain further insight into the classification power of the classifiers developed in our study, we should conduct validation experiments on a larger and more diverse dataset. Such validation efforts would provide a more robust assessment of the classifiers' performance and their ability to accurately classify new instances.

As regards the correlation analysis with clinical scales, it is important to note that it was performed only on PD patients; the absence of clinical data for the healthy control group is a limitation of our study. Including clinical assessments of healthy controls could provide further insights and potentially reveal stronger or novel correlations between the radiomic features and clinical measures. In addition, considering the radiomic features of the caudate region could improve the results in this type of analysis. In Rahmim et al. [9], caudate has shown to exhibit the highest correlation between textural features and severity of motor and cognitive symptoms, compared to putamen, which showed a lower correlation. By identifying robust associations between radiomic features and clinically meaningful parameters, we can establish a clearer understanding of the underlying biological and pathological processes reflected in these features. This, in turn, can facilitate the translation of radiomics into clinical practice by providing clinicians with meaningful and actionable insights for diagnosis, treatment planning, and monitoring of PD. Enhancing the interpretability of radiomic features, in fact, is crucial to gain the trust and acceptance of clinicians, who may approach radiomics with skepticism due to the sometimes elusive nature of their meaning.

Finally, it is important to remember that the optimal choice of classifiers in the context of radiomics remains an open question. In our study, we compared the performance of three distinct classifiers; however, future investigations could expand this comparison to include more sophisticated models such as Random Forest or XGBoost [17], which are often employed in scenarios involving larger and more intricate datasets. Similarly, in terms of feature selection, exploring alternative combinations of feature selection methods may yield valuable insights. Continued research in this field is warranted to identify the most suitable classifiers and feature selection approaches that can effectively leverage the potential of radiomics.

5. Conclusions

Our study demonstrated the potential of radiomic features in detecting pathological changes associated with Parkinson's disease (PD) in DAT-SPECT scans. The selected radiomic features exhibited significant differences between PD patients and healthy controls, indicating their relevance as potential biomarkers for PD detection. Additionally, our study has yielded promising findings regarding the correlation between these radiomic features and clinical scales, as well as their potential implications in understanding the underlying pathophysiological mechanisms. Looking ahead, further advancements in the field of radiomics hold promise for their integration into clinical practice, offering enhanced capabilities for diagnosis, treatment, and monitoring of Parkinson's disease.

References

[1] Ronald F Pfeiffer. Non-motor symptoms in parkinson's disease. *Parkinsonism & related disorders*, 22:S119–S122, 2016.

[2] Bernd Müller, Jörg Assmus, Karen Herlofson, Jan Petter Larsen, and Ole-Bjørn Tysnes. Importance of motor vs. non-motor symptoms for health-related quality of life in early parkinson's disease. *Parkinsonism & related disorders*, 19(11):1027–1032, 2013.

[3] Dennis W Dickson. Neuropathology of parkinson disease. *Parkinsonism & related disorders*, 46:S30–S33, 2018.

[4] Valery L Feigin, Emma Nichols, Tahiya Alam, Marlena S Bannick, Ettore Beghi, Natacha Blake, William J Culpepper, E Ray Dorsey, Alexis Elbaz, Richard G Ellenbogen, et al. Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the global burden of disease study 2016. *The Lancet Neurology*, 18(5):459–480, 2019.

[5] Günther Deuschl, Ettore Beghi, Franz Fazekas, Timea Varga, Kalliopi A Christoforidi, Eveline Sipido, Claudio L Bassetti, Theo Vos, and Valery L Feigin. The burden of neurological diseases in europe: an analysis for the global burden of disease study 2017. *The Lancet Public Health*, 5(10):e551–e567, 2020.

[6] Bastiaan R Bloem, Michael S Okun, and Christine Klein. Parkinson's disease. *The Lancet*, 397(10291):2284–2303, 2021.

[7] Marios Politis. Neuroimaging in parkinson disease: from research setting to clinical practice. *Nature Reviews Neurology*, 10(12):708–722, 2014.

[8] Ana M Catafau and Eduardo Tolosa. Impact of dopamine transporter spect using 123i-ioflupane on diagnosis and management of patients with clinically uncertain parkinsonian syndromes. *Movement disorders: official journal of the Movement Disorder Society*, 19(10):1175–1182, 2004.

[9] Arman Rahmim, Yousef Salimpour, Saurabh Jain, Stephan AL Blinder, Ivan S Klyuzhin, Gwenn S Smith, Zoltan Mari, and Vesna Sossi. Application of texture analysis to dat spect imaging: relationship to clinical assessments. *NeuroImage: Clinical*, 12:e1–e9, 2016.

[10] Janita E Van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into imaging*, 11(1):1–16, 2020.

[11] Arman Rahmim, Peng Huang, Nikolay Shenkov, Sima Fotouhi, Esmaeil Davoodi-Bojd, Lijun Lu, Zoltan Mari,

Hamid Soltanian-Zadeh, and Vesna Sossi. Improved prediction of outcome in parkinson's disease using radiomics analysis of longitudinal dat spect images. *NeuroImage: Clinical*, 16:539–544, 2017.

[12] Takuro Shiiba, Kazuki Takano, Akihiro Takaki, and Shugo Suwazono. Dopamine transporter single-photon emission computed tomography-derived radiomics signature for detecting parkinson's disease. *EJNMMI research*, 12(1):1–12, 2022.

[13] Augusto Destrero, Sofia Mosci, Christine De Mol, Alessandro Verri, and Francesca Odone. Feature selection for high-dimensional data. *Computational management science*, 6:25–40, 2009.

[14] Aydin Demircioğlu. Benchmarking feature selection methods in radiomics. *Investigative Radiology*, 57(7):433–443, 2022.

[15] Valentina Brancato, Nadia Brancati, Giusy Esposito, Massimo La Rosa, Carlo Cavaliere, Ciro Allarà, Valeria Romeo, Giuseppe De Pietro, Marco Salvatore, Marco Aiello, et al. A two-step feature selection radiomic approach to predict molecular outcomes in breast cancer. *Sensors*, 23(3):1552, 2023.

[16] Runtao Yang, Chengjin Zhang, Lina Zhang, and Rui Gao. A two-step feature selection method to predict cancerlectins by multiview features and synthetic minority oversampling technique. *BioMed research international*, 2018, 2018.

[17] Nguyen Quoc Khanh Le, Truong Nguyen Khanh Hung, Duyen Thi Do, Luu Ho Thanh Lam, Luong Huu Dang, and Tuan-Tu Huynh. Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from mri. *Computers in Biology and Medicine*, 132:104320, 2021.

[18] Koen JF Verhoeven, Katy L Simonsen, and Lauren M McIntyre. Implementing false discovery rate control: increasing your power. *Oikos*, 108(3):643–647, 2005.

[19] Leonardo F Machado, Paula CL Elias, Ayrton C Moreira, Antônio C Dos Santos, and Luiz O Murta Junior. Mri radiomics for the prediction of recurrence in patients with clinically non-functioning pituitary macroadenomas. *Computers in Biology and Medicine*, 124:103966, 2020.

[20] D Salas-Gonzalez, JM Górriz, J Ramírez, IA Illán, M López, F Segovia, R Chaves, P Padilla, CG Puntonet, and Alzheimer's Disease Neuroimage Initiative. Feature selection using factor analysis for alzheimer's diagnosis using pet images. *Medical physics*, 37(11):6084–6095, 2010.

[21] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[22] Yimeng Fan, Chaoyue Chen, Fumin Zhao, Zerong Tian, Jian Wang, Xuelei Ma, and Jianguo Xu. Radiomics-based machine learning technology enables better differentiation between glioblastoma and anaplastic oligodendroglioma. *Frontiers in oncology*, 9:1164, 2019.

[23] Mohammad Hossin and Md Nasir Sulaiman. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1, 2015.

[24] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13, 2020.

[25] Stephen SF Yip and Hugo JWL Aerts. Applications and limitations of radiomics. *Physics in Medicine & Biology*, 61(13):R150, 2016.

[26] Shruti Atul Mali, Abdalla Ibrahim, Henry C Woodruff, Vincent Andrearczyk, Henning Müller, Sergey Primakov, Zohaib Salahuddin, Avishek Chatterjee, and Philippe Lambin. Making radiomics more reproducible across scanner and imaging protocol variations: a review of harmonization methods. *Journal of personalized medicine*, 11(9):842, 2021.

[27] Christopher G Goetz, Barbara C Tilley, Stephanie R Shaftman, Glenn T Stebbins, Stanley Fahn, Pablo Martinez-Martin, Werner Poewe, Cristina Sampaio, Matthew B Stern, Richard Dodel, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, 23(15):2129–2170, 2008.

[28] Kallol Ray Chaudhuri, Pablo Martinez-Martin, Anthony HV Schapira, Fabrizio Stocchi, Kapil Sethi, Per Odin, Richard G Brown, William Koller, Paolo Barone, Graeme MacPhee, et al. International multicenter pilot study of the first comprehensive self-completed nonmotor symptoms questionnaire for parkinson's disease: the nmsquest study. *Movement disorders: official journal of the Movement Disorder Society*, 21(7):916–923, 2006.

[29] Verna C Pangman, Jeff Sloan, and Lorna Guse. An examination of psychometric properties of the mini-mental state examination and the standardized mini-mental state examination: implications for clinical practice. *Applied nursing research*, 13(4):209–213, 2000.

[30] Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

A. Appendix: Clinical background

For clinical assessment of PD severity, quantitative and qualitative investigation tools are employed to examine the stage of the disease, its effects on everyday functional activities, and the degree of patients' cognitive impairment. Some of these scales are briefly described in this section.

MDS-UPDRS The Unified Parkinson's Disease Rating Scale consists of four parts: I) Non-motor experiences of daily living, II) Motor experiences of daily living, III) Motor examination, and IV) Motor complications. Questions from Part I and Part II have been designed to be completed by the patient or the caregiver in a questionnaire format. Part III provides for the objective investigation of parkinsonism by the clinician, while Part IV examines motor fluctuations and dyskinesias. Each question features five responses that are linked to commonly accepted clinical terms: Normal = 0, Slight = 1, Mild = 2, Moderate = 3, and Severe = 4 [27].

NMSQ The Non-Motor Symptoms Questionnaire is a screening tool designed to assess the presence of nonmotor symptoms in PD patients. The NMSQ is a self-completed questionnaire comprising 30 items investigating whether the patient has experienced the specified symptoms in the previous month. The questionnaire features 'yes' or 'no' as responses to each item and was therefore not designed as a quantitative scale; however, a total score can be computed by summing all the positive responses ('yes'). The questionnaire examines several relevant domains, such as memory and attention deficits, physiological malfunctions, depression, and sleep disorders [28].

MMSE The Mini-Mental State Examination is a clinical and research instrument designed to measure cognitive impairment. The screening must be administered by a health professional. It consists of 11 tasks, addressing orientation, memory, attention, language, and calculations. The patient is given a score for each task; the total possible score is 30 points. Generally, 23 is set as the threshold: scores of 23 and lower are indicative of cognitive impairment, while scores of 24 and higher indicate intact cognitive abilities [29].

MoCA The Montreal Clinical Assessment is another screening tool to evaluate mild cognitive impairment, often coupled with MMSE. The tested domains are short-term memory, visual-spatial abilities, executive functions, verbal abstraction, attention, concentration, working memory, language (naming, repetition, and fluency), and time and place orientation. The scores range from 0 to 30 points; the value of the optimal cut-off is still debated [30].