Detection of Dopamine Deficiency for Parkinson's Disease Diagnosis with Machine Learning and Structural MRI

Carolina Vicente

Instituto Superior Técnico, Lisboa, Portugal

October 2021

Abstract

Parkinson's disease is a neurological disorder that affects 1% of the population over 60. Multiple diseases cause similar symptoms, but Parkinson's disease is characterized by dopaminergic neuronal loss. This leads to dopamine deficiency, which can be detected with a DaTscan. Subjects initially diagnosed with Parkinson's, but who have a negative exam are grouped as patients with Scans Without Evidence of Dopamine Deficiency. The present work aims to achieve the distinction of subjects with and without dopamine deficiency with a structural Magnetic Resonance Imaging scan. Images from 311 subjects from the PPMI database were processed with *FreeSurfer* into 689 features. Data was then divided into 2 categories, with 70% allocated to training sets, and 30% set aside for test sets. Cross-validation and a validation set, made up of 10% of the training data, were used to compare different modelling approaches. An existing Machine Learning pipeline was used as a baseline approach. Multiple algorithms were compared. For feature selection, the features were partitioned into sets according to brain region, and as an alternative, features from robust Principal Component Analysis. The baseline approach overfitted, with accuracies of 96.6% and 54.5% for training and validation sets, respectively. All other simpler approaches resulted in underfitting or overfitting, with the highest validation balanced accuracy being 80.42% and 62.67% for cross validation. These were tested in the independent test set where the highest balanced accuracy was 50.60%.

Keywords: Parkinson's Disease, SWEDD, Machine Learning, Magnetic Resonance Imaging

1. Introduction

Parkinson's Disease (PD) is a neurological disorder that currently affects 1% of the population older than 60 years [1], with more than 6 million individuals affected worldwide [2], and it is expected to affect a greater percentage of the population in the decades to come [3]. In regard to its pathogenesis, most cases of PD are idiopathic (unknown cause), although there are known genetic and environmental contributions [2], with the greatest risk factor being age [3].

Pathologically, PD is characterised by a loss of dopaminergic neurons and the presence of Lewy bodies, which are abnormal aggregations of protein that develop inside nerve cells in the midbrain [3]. This leads to non-motor symptoms, like sleep disorders, and motor symptoms, for instance tremors and rigidity, the latter two being the most well-known PD symptoms.

Diagnosing Parkinson's Disease

There is currently no definitive test for the diagnosis of PD in the living. Diagnosis requires post-mortem examination of the brain for neuronal loss and depigmentation of the substantia nigra, in addition to the presence of Lewy bodies in the brain stem [4, 5].

Patients usually seek clinical help when motor symptoms start. Parkinsonism is a general term for a group of neurological disorders that cause those motor symptoms such as tremors, slow movement, and stiffness, and PD is the disease that most commonly explains a Parkinsonism case. The diagnosis process for a patient presenting symptoms relies on the expertise and experience of clinicians to distinguish and identify the underlying disease, based mainly on observational signs and symptoms, brain exams and response to medication [5].

The accuracy of PD diagnosis still has room for improvement, with current studies reporting an accuracy of 83.9% when done by experts and 73.8% when given by non experts [6].

DaTscan and SWEDD

DaTscan SPECT (single-photon emission computerized tomography) is a highly accurate exam, with 98% sensitivity and 100% specificity, in detecting dopamine deficiency in subjects with Parkinsonism [2]. The use of DaTscan for the diagnosis of PD has been studied [7], but it does not add enough to the diagnostic assessment to make it worthwhile [2],

since this exam is expensive and not easily available.

However, there are subjects that are clinically diagnosed with PD, whom after post-mortem examination or via DaTscan are detected to not have dopamine deficiency. These subjects are grouped as SWEDD (Scans Without Evidence of Dopamine Deficiency) but it should not be considered a diagnosis, since these subjects may present a plethora of different diseases other than PD, for instance supranigral parkinsonism and vascular parkinsonism [8].

Hypothesis and objectives

The present work study aims to achieve the distinction of subjects with and without dopamine deficiency with a structural Magnetic Resonance Imaging scan.

The main objective is to study if the hypothesis is possible. Starting by trying an already existing approach as a baseline, then exploring different algorithms, data transformations and machine learning methods. Validate these methods to choose the best one, to then test it in an independent set. If the hypothesis is achieved, then get some insights on how the distinction is done, for instance, what brain regions are the most important.

In the following 2 Background section an overview of what has been done in literature regarding the hypothesis and how the MRI processing works. Followed by 3 Implementation section where the methods used are detailed. In 4 Results section all the results are shown with a discussion on the work. Finally in 5 Conclusion section there is an explanation if the hypothesis was fulfilled along with some ideas for future work.

2. Background

2.1. The promise of Machine Learning

Machine learning (ML) has been used to distinguish PD subjects from others, through the use of symptoms, speech, movement patterns, and neuroimaging data [9–11].

Some reviews comment on the possibility of exploring Deep Learning to accomplish this [12–14], but some of the most common critiques to this approach are the lack of data quantity and that improvement in the accuracy does not seem worthwhile given the resulting loss of interpretability.

There is a large quantity of articles aiming at the diagnosis of PD through the use of classical ML, and reviews that attempt to compare them [15–18]], albeit this is currently difficult to do since there is no reporting standard, and most often insufficient details describing the employed analysis pipeline. The most common algorithms reported to be used in several of the reviews are SVM, with some mentions of Random Forest, Naive Bayes and Logistic Regression. The accuracy in distinguishing between healthy controls and PD subjects varies between 80% and 100%, which are unusual results to obtain in a ML model, with a proper independent test, especially with data as complex as the brain. Moreover, distinguishing between healthy controls and different diseases tends to yield good results, while trying to differentiate diagnosis tends not to lead to such good results [15]. It should be noted that accuracy is used in every comparison in the reviews but it can be misleading because of unbalanced datasets.

PD vs SWEDD

A search query in PubMed and Scopus for articles that tried to accomplish the separation of groups by dopamine (PD, SWEDD, and control groups) with MRI and machine learning, returned very few studies, see Table 1. Of these results, only one uses sMRI T1w. The others use DTI images and classical ML algorithms with features extracted from the images (DTI tracts), reporting an accuracy of up to 97%.

Table 1 includes one of Diana Prata's lab unpublished data using sMRI T1w to distinguish PD from SWEDD with ML. It reports an accuracy of 97.4%, 73.3% and 65.3% in separating PD vs Control, PD vs SWEDD and Control vs SWEDD, respectively, by choosing adequate MRI slices. In particular, this suggests that Control and SWEDD groups might be not easily distinguishable.

2.2. Image preprocessing

There are multiple resources that can be used to process MRI scans, the purpose of which is to turn these 3D images into features that can be isolated and extracted. These can be for example surface areas, folding indexes, or volumes of regions of the brain.

FreeSurfer [24] provides a full processing stream for structural MRI data [25], including: skull stripping (Figure 1), gray-white matter segmentation, reconstruction of cortical surface models, labeling of regions on the cortical surface, as well as subcortical brain structures (Figure 2).

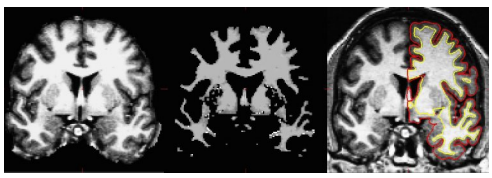

Figure 1: Three stages from the FreeSurfer cortical analysis pipeline: A - Skull stripped image. B - White matter segmentation. C - Surface between white and gray (yellow line) and between gray and pia (red line) overlaid on the original volume [24]

Table 1: Comparison table of studies that differentiate PD, Control and SWEDD

Reference	MRI type	Features	Model	Sample Size	PD	Control	SWEDD	Test	Accuracy
[19]	DTI	DTI tracts	SVM	142	37%	37%	27%	42%	81.25%
[20]	DTI	DTI tracts	SVM quadratic	48	54%	-	46%	LOOCV	72.5%
[21]	DTI	DTI tracts	SVM linear	80	33%	33%	33%	LOOCV	77.92%
[22]	DTI	DTI tracts	SVM	77	48%	-	52%	LOOCV	97%
[23]	sMRI T1w	Image slices	CNN	197	43%	43%	15%	30%	60%-80%

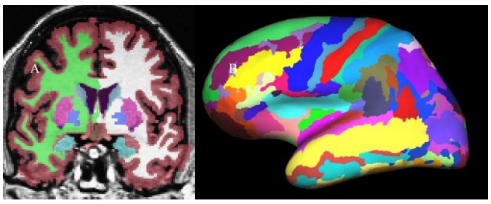

Figure 2: A - Volume-based labeling. Note that cortical gray matter and white mater are represented by single classes. Also note that there are separate labels for the structures in each hemisphere. B- Surface-based labeling. [24]

3. Implementation

3.1. Data Source

The MRI T1w came from the PPMI database [26]. PPMI is a landmark observational study that makes its data set and biorepository available to academia and industry. The PPMI study divides its enrolled subjects into different groups, called *Research Groups*.

3.2. Image Selection

For each subject it was selected only one image, to avoid subject-bias in the hopes of achieving the best classification facilitating the interpretation of results and gathering of insights. In the PPMI database there are 761 subjects with an MRI and DaTscan. From this group various choices in parameters and restrictions were chosen:

1. MRI had to have *Field Strength* = 3T, left 471 subjects;
2. MRI had to be *3D*, left 470 subjects;
3. MRI had to be *MPRAGE*, left 407 subjects;
4. MRI without considerable noise or ghosting, left 402 subjects;
5. MRI had to have *Repetition Time* = 2300, left 337 subjects;
6. MRI and DaTscan within 12 months, left 316 subjects

So, each subject had to have a DaTscan result and an MRI T1w, within one year so as to be as close to the ground truth diagnosis as possible. Furthermore, the MRI T1w had to have good quality and minimal noise and ghosting, selected via a manual visual inspection. Additionally it had to have the following parameters: *MPRAGE*, *Field Strength*=3T, *Repetition Time (TR)*=2300. These parameters were chosen to avoid possible bias that could exist due to relations between the different parameters and labels.

There were 5 subjects whose DaTscan result did not match what would be expected from belonging to the *Research Group*, and these were excluded. Once the model is chosen and trained, they may be used to check what the model would classify them as.

Included data

From the 311 chosen subjects, 104 (33%) have a negative DaTscan, and 207 (67%) a positive DaTscan. With positive label, 71 are Control and 33 SWEDD, while the ones with negative label, 160 are PD, 46 are GenCohort PD, and 1 GenReg PD.

3.3. Resulting Features

The pipeline used from Freesurfer processes sMRI wT1 into 689 features, 8 of them are copies or masks and can be removed. The remaining are separated between left and right hemisphere (*lh*, *rh*), and these features can be divided between 41 different brain regions. These features can be of 9 different types: *thickness*, *volume*, *curvind*, *wm*, *thicknessstd*, *foldind*, *gauscurv*, *meancurv*, *area*.

3.4. Training, Validation and Testing Sets

An independent test set is important for testing the validity of the results obtained and to test the final model as a possible tool to be used in the clinical diagnosis process. Therefore, a representative and significant test set is important, so 30% of the data was reserved for the final testing. Some considerations regarding the diagnosis, sex and age, had to be accounted for, since the amount of data is not large enough to rely on randomness as a guaranteed means to achieve representation of all classes.

Age was balanced between , since it can be a confounding variable in MRI images, because age has an affect on the brain, for instance, it is negatively correlated with grey matter measures [27]. This is accomplished by ensuring the average age is the same in both sexes. Furthermore, the different diagnosis in each DaTscan label is also balanced such that both sexes have similar percentages of each diagnosis, and subjects with multiple images were

left to the test set, since they could bring a higher value in the verification of the reliability of the results. Due to there is a relationship between sex and PD diagnosis, as there is a higher percentage of men diagnosed [28], the balancing in the test set was done such that it is equivalent across diagnoses, with 1 female to every 1.5 male, so that the confidence estimates can be equal for both sexes in the test results analysis.

After imposing these restrictions, subjects were chosen randomly from the remaining options, to be either in the training or testing sets.

To compare different approaches either cross-validation can be used or validation in an independent validation set. For a validation set, 10% of the training set was selected randomly, and for the cross-validation all the training set was used.

3.5. Baseline Approach

Diana Prata's Lab had previously developed an MRI ML pipeline to create and train a model to aid diagnosis of Alzheimer's disease (under peer review).

This approach consists of using a voting system between 7 different classifiers: a linear support vector machine (l-SVM); a decision tree classifier (DT); a random forest classifier (RF); an extremely randomized tree classifier (ET); a linear discriminant analysis classifier (LDA); a logistic regression classifier (LR); a logistic regression classifier with stochastic gradient descent learning (LR-SGD).

All hyperparameters are chosen by using an evolution algorithm and cross validation.

This pipeline served as the baseline approach in the present study.

3.6. Data Transformations

Different transformations can be made to the data, for efforts of feature selection or to remove subject specific data to make comparing data between subjects fairer.

Brain Regions

The 672 features were divided into 41 sets related to certain brain regions. This transformation allows to have less features than subjects and it can also be a method to identify what brain regions are more important for the classification problem.

This idea originally purposed herein aimed at using *a priori* knowledge that exists of the features and data being used, instead of relying on naive search for feature selection.

Robust Principal Component Analysis

Features obtained by explaining 90% of variance with Robust Principal Component Analysis (RPCA) was used instead of the original 672, which transforms the problem into one where there are less features than subjects.

One possible problem with this transformation is that since the number of subjects is less than original features, the new features explain not just variance of features but also of subjects. And so these features might not be well chosen for generalizability and future data.

Normalize

Data normalization is a transformation that is performed before RPCA, but using it by it self can bring improvement to some models, such as SVM, since it allows the models not to give more significance to certain features with higher values.

This transformation is done such that for each feature the following calculation is done.

$$x' = \frac{x - \bar{x}}{std}$$

where std is the standard deviation, and $\bar{x}$ is the mean.

Relative

The features which are of type area or volume, can be divided by the total surface area of the brain, or total volume of the brain, so as to have features that can be comparable between subjects. These features become relative instead of absolute.

3.7. Balancing

This dataset was unbalanced, and as most machine learning algorithms perform better when training with balanced data, multiple techniques for balancing the data were explored.

Balancing in model

Most models implemented by *sklearn*, and in particular all models used herein, have a parameter that can be used so the model gives weights to the classes, and for balancing, this parameter can be set to balanced, `class_weight='balanced'`.

Undersampling

Another option is to perform undersampling, which consists of removing samples from the class with the highest amount, so as to balance them.

Oversampling

Another option is to perform overrsampling, which consists of repeating samples from the class with the least amount, so as to balance them.

3.8. Exploration for best model

The main idea was to test simple models, in order to avoid overfitting, which happened in the baseline approach, explained below in detail. Comparing multiple options of combining different algorithms, balancing and transformation seems to be the best approach as, in this way, it is possible to choose the best ones to test in the reserved set.

Here, the multiple options were attained using combinations of alternatives. Each combination is composed choosing one of each alternative in the following bullet points:

1. Algorithm: Logistic Regression, Perceptron, Ridge Classifier, Random Forest, Support Vector Machine
2. Validation: 10% of training set, Cross-validation
3. Balancing: None, Over, Under, Balanced in model
4. Features: All features, the subset of features for each of the 41 brain regions
5. Transformation: None, Normalize, Relative, RPCA (only when all features are selected)

Algorithms were chosen for their interpretability, and the hyperparameters were chosen with cross-validation and grid search.

Besides the balancing types mentioned, types which create artificial data were not considered due to the medical nature of the data. Moreover, when using cross-validation and oversampling simultaneously it could be the case that the training subsets in the folds could have the same data points as in the testing fold, since *sklearn* is not prepared to handle this situation. Thus, this method was programmed from scratch.

The RPCA transformation is only applied when using all features, since this and selecting a subset of features from a brain region are different feature selection methods.

In sum, there were a total of 160 combinations if all features are selected, and 4920 when the subset of features corresponds to each of the 41 brain regions.

For each combination the following metrics were calculated: Accuracy and Balanced Accuracy.

4. Results

4.1. Label distribution

In the dataset, 33% have a negative DaTscan and 67% have a positive DaTscan. Both training and testing set have the same distribution.

4.2. Research Groups Distribution

Subjects that have a negative DaTscan belong to either the Control group or SWEDD group, while subjects that have a positive DaTscan belong to either the PD group or GenCohort PD group or GenReg PD group. Since there was a high percentage of subjects from the GenCohort PD group that had multiple images in the same day, and these were to be reserved to the test set, the distributions of training and testing sets are not equal. The final distributions in both sets can be seen in Table 2.

Table 2: Distribution of research groups, in the training and testing sets.

	Control	**SWEDD**	**PD**	**GenCohort PD**	**GenReg PD**
Train	23%	11%	53%	13%	1%
Test	23%	11%	47%	19%	0%

4.3. Age distribution

The mean and standard deviation (std) age across all the data is 61.3 ± 10.1. To test the significance of age in the DaTscan result, the Mann-Whitney test was used, which resulted in a $p = 0.163$, so there is not significant evidence to reject that the age distribution is equal for both labels.

Table 3: Age mean and std across label, in the train and test sets.

	DaTscan negative	**DaTscan positive**
Train	59.1 ± 11.7	61.8 ± 9.4
Test	61.8 ± 10.0	62.5 ± 9.4

Table 4: Age mean and std across research group, in the train and test sets.

	Control	**SWEDD**	**PD**	**GenCohort PD**	**GenReg PD**
Train	58.2± 12.0	61.1± 10.8	61.5± 9.0	62.6± 11.0	70.9
Test	62.0± 9.6	61.4± 11.3	60.6± 9.8	67.2± 6.5	-

4.4. Sex distribution

The distribution of the sexes across all the data is 36% female and 64% male. To test the significance of sex in the DaTscan result, the Person Chi-Square test was used, which resulted in a $p = 0.762$, so there is not significant evidence to reject that the sex distribution is equal for both labels.

Table 5: Sex distribution Female - Male, across label, in the train and test sets.

	DaTscan negative	**DaTscan positive**
Train	36% - 64%	34% - 66%
Test	41% - 59%	40% - 60%

4.5. Correlation analysis results

By analysing the number of zero entries in the features, four features were found to only have value zero: *Right-non-WM-hypointensities*, *Left-non-WM-hypointensities*, *Right-WM-hypointensities* and *Left-WM-hypointensities*, and they were removed. Moreover, the features *5th-Ventricle*

Table 6: Sex distribution Female - Male, across research group, in the train and test sets.

	Control	SWEDD	PD	GenCohort PD	GenReg PD
Train	35% 65%	39% 61%	35% 65%	29% 71%	-
Test	41% 59%	40% 60%	39% 61%	44% 56%	100% 0%

and *non-WM-hypointensities* only have 4 and 25 non-zero entries, which can indicate that they may not be useful for the models, but they were left in the dataset.

By checking the features with correlation higher/lower than 0.95/-0.95 the following five features were removed, which we assumed would be redundant, and with the correlation we confirm they don't have significant information.

4.6. Robust Principal Component Analysis results

The RPCA returns 29, 50 and 129 features that explain 80%, 90% and 100% variance, respectively.

4.7. Baseline approach results

Using the pipeline explained in Section 3.5 resulted in a model which its performance in shown in Table 7, with the results of labeling both the training dataset as well as the testing dataset. Across all metrics, we can see the model performs significantly worse in labelling the testing dataset when compared to the results from the training dataset. This was a clear sign that the model was overfitting the data, and a possible explanation would be the high complexity of the model and chosen hyperparameters.

Table 7: Results from baseline approach.

	Accuracy	Balanced Accuracy	F1 score	Balanced F1 score	ROC AUC
Train	96.6%	94.9%	93.3%	96.5%	99.5%
Test	54.5%	43.8%	16.7%	52.2%	40.0%

4.8. Exploration for best model results

Results shown in this section follow the idea: first all the results are shown together along with some statistics, followed by details of the scores for the 5 best model combinations for all features or regions and for validation or cross validation. Finally the scores of testing of these best models are reported.

The balanced accuracy was used to compare all combinations, since accuracy is used in the literature but using the normal accuracy could be misleading since the data is unbalanced.

4.9. Overall results

There are more than 5000 possible combinations from following the approach explained in Section 3.8, and it is not possible to discuss them individually. So, to be able to evaluate overall how different approaches affect the results, the balanced accuracy average was calculated for different aggregations, which we can see in Tables 8, 9, 10 with the combinations that used all the features, and Tables 11, 12 and 13 for the combinations that used subsets of features relative to brain regions.

Table 8: Average balanced accuracies for all features and for each algorithm

Model	Train	Validation	CV
LR	79.34%	54.77%	47.73%
Perceptron	71.70%	58.91%	48.09%
RC	81.33%	58.70%	50.13%
RF	100.00%	53.93%	48.97%
SVM	75.00%	49.92%	50.00%

Table 9: Average balanced accuracies for all features and for each balancing type

Balancing	Train	Validation	CV
None	72.09%	53.21%	49.01%
Undersampling	91.91%	54.60%	49.01%
Oversampling	90.58%	59.25%	49.18%
In-model	71.32%	53.92%	48.73%

Table 10: Average balanced accuracies for all features and for each transformation

Transformation	Train	Validation	CV
None	80.61%	53.21%	50.14%
RPCA	79.78%	53.83%	48.15%
Normalize	88.65%	56.69%	47.55%
Relative	76.86%	57.25%	50.10%

Table 11: Average balanced accuracies when using brain regions features subsets and for each algorithm

Model	Train	Validation	CV
LR	56.66%	50.67%	49.35%
Perceptron	50.83%	50.33%	49.88%
RC	60.76%	51.00%	49.05%
RF	100.00%	53.21%	49.47%
SVM	75.61%	49.47%	50.02%

Table 12: Average balanced accuracies when using brain regions features subsets and for each balancing type

Balancing	Train	Validation	CV
None	61.96%	50.33%	49.70%
Undersampling	74.62%	50.80%	49.70%
Oversampling	74.27%	51.28%	49.97%
In-model	64.24%	51.33%	48.84%

4.10. Best model combinations results

For the tables in this section *B. Accuracy* represents balanced accuracy.

Table 13: Average balanced accuracies when using brain regions features subsets and for each transformation

Transformation	Train	Validation	CV
None	68.77%	50.94%	49.64%
Normalize	68.77%	50.94%	49.50%
Relative	68.77%	50.94%	49.52%

In Tables 14, 15, 16, 17 we see the results of the best combinations for all features, and for brain regions, both when using validation and cross-validation Balanced Accuracy .

The models were tested and outputted the results shown in Tables 18, 19, 16 and 17.

4.11. Discussion

Result

In regard to general results, a common tendency (seen across Tables 8 through 13) is that the balanced accuracy of the models is higher for the training set than for the validation set, with the latter being on average slightly above 50%. This shows that most models suffered from overfitting. Tables 8 and 11 in particular show that models using Random Forest as the algorithm were especially prone to this, seeing as the average balanced accuracy for the training set is 100%.

The oversampling strategy for balancing data tends to give better results (as seen in Tables 9 and 12), both from the lens of validation and cross-validation. In addition, there is no significant improvement from using any of the studied feature transformations (Tables 10 and 13).

Looking at the best models, the Ridge Classifier algorithm provides the best models for dealing with a large number of features (Tables 14 and 15), whereas Logistic Regression is most appropriate for the situations with less features (Tables 16 and 17). On the other hand, models using SVM as algorithm are not represented in the top models. Regarding data balancing, oversampling seems to lead to better models, whereas RPCA is not represented among the top models. Finally, the features from the brain regions *insula* and *temporalpole* seem to be of importance for this classification.

Although these results may seem promising, it should be noted that, since we have studied so many combinations, finding good results might simply be because we are bound to find some model that happens to work well with this particular dataset. This is why it is important to use a testing dataset, which is completely separate from the pipeline followed up until this point.

The testing results (Tables 18 through 21) show that, across all combinations, the top models selected before do not generalize well: for instance, the values of balanced accuracy do not go higher than 60%. Despite that, the majority of these models outclass the baseline considered for this work in terms of the balanced accuracy, as well as the other metrics calculated.

Considerations

The articles found and studied in this report show a lack of research attempts in using MRI to differentiate subjects with and without dopamine deficiency, and only a few using Diffusion MRI – with the most common MRI (sMRI Tw1) still surprisingly unexplored. Hence the hypothesis of the present study was to examine whether it would be possible to use sMRI to identify what patients would have a positive result from a DaTscan, with the use of Machine Learning.

With the initial analysis, some patterns in the data seem to emerge which could indicate that a smaller set of features could be used. Furthermore, the baseline ML approach seems to be overffited, with balanced accuracy of the train and test set being 94.6% and 43.8%, respectively. This indicated that a possible path would be to use simpler models and method of finding hyperparameters.

The different models, when further tested in an independent set, suffered from underfitting or overfitting, but with the best models having higher scores than the baseline approach. The highest balanced accuracy achieved was 50.60%.

The option of using subsets of features that were relative to brain regions all ran into the same issues, but this would be a method of feature selection that could pinpoint what regions would be of most importance to this classification problem. The highest balanced accuracy achieved was 57.11% with features from the *temporalpole*.

Validation vs cross-validation

Comparing the validation and cross-validation results found across the top models tested, the balanced accuracy obtained from cross validation is a better predictor of the testing balanced accuracy, whereas the results using the usual validation are overly optimistic. This is in line with observations suggesting cross validation should be preferred in situations where small datasets are available, as is the case here.

5. Conclusion

The overall conclusion is that, although we were able to obtain models that outperform the baseline, the approaches used with the given dataset are not able to distinguish between subjects with and without dopamine deficiency with the features that were extracted from a sMRI. As thus, this work has no immediate clinical applicability.

Future Work

Multiples paths can be followed to further test if it possible to use MRI to identify what patients would

Table 14: Best combinations for all features, using validation.

ID	Model	Transformation	Balancing	B. Accuracy	Accuracy
1	Perceptron	relative	oversampling	77.08%	73.91%
2	Logistic Regression	relative	oversampling	75.00%	82.61%
3	Ridge Classifier	none	oversampling	71.25%	73.91%
4	Ridge Classifier	none	in-model	67.92%	69.57%
5	Ridge Classifier	none	none	67.92%	69.57%
6	Perceptron	normalize	none	67.92%	69.57%
7	Perceptron	normalize	oversampling	67.92%	69.57%

Table 15: Best combinations for all features, using cross-validation.

ID	Model	Transformation	Balancing	B. Accuracy	Accuracy
1	Ridge Classifier	none	oversampling	55.18%	52.92%
2	Ridge Classifier	relative	none	53.15%	65.47%
3	Ridge Classifier	relative	undersampling	53.15%	65.47%
4	Logistic Regression	relative	oversampling	53.06%	45.97%
5	Random Forest	normalize	oversampling	52.66%	59.35%

Table 16: Best combinations, using brain regions features subsets and validation.

Region	Model	Balancing	B. Accuracy	Accuracy
insula	Logistic Regression	oversampling	80.42%	78.26%
temporalpole	Random Forest	undersampling	77.92%	82.61%
rostralmiddlefrontal	Random Forest	undersampling	77.08%	73.91%
superiorfrontal	Ridge Classifier	in-model	77.08%	73.91%
cuneus	Random Forest	oversampling	74.58%	78.26%

Table 17: Best combinations, using brain regions features subsets and cross-validation.

Region	Model	Transformation	Balancing	B. Accuracy	Accuracy
posteriorcingulate	Ridge Classifier	*	in-model	62.67%	62.68%
paracentral	Logistic Regression	normalize	oversampling	59.51%	58.14%
cerebellum	Logistic Regression	normalize	oversampling	59.30%	51.15%
posteriorcingulate	Logistic Regression	*	in-model	59.17%	59.01%
insula	Random Forest	normalize	oversampling	58.43%	59.46%

Table 18: Testing models from Table 14, that use all features, obtained with validation

ID	Train B. Accuracy	B. Accuracy	Accuracy
1	61.34%	50.60%	62.77%
2	85.86%	48.49%	50.00%
3	100%	45.41%	48.94%
4	100%	43.00%	45.74%
5	100%	44.60%	47.87%
6	100%	45.36%	47.87%
7	98.97%	44.60%	47.87%

Table 19: Testing models from Table 15, that use all features, obtained with cross-validation

ID	Train B. Accuracy	B. Accuracy	Accuracy
1	100%	45.41%	48.94%
2	52.08%	49.19%	64.89%
3	100%	48.44%	48.94%
4	85.86%	48.49%	50.00%
5	100%	48.34%	62.77%

Table 20: Testing models from Table 20, that use brain regions, obtained with validation

Region	Train B. Accuracy	B. Accuracy	Accuracy
insula	56.90%	45.51%	51.06%
temporalpole	100%	57.11%	56.38%
rostralmiddlefrontal	100%	54.94%	58.51%
superiorfrontal	57.45%	39.01%	41.49%
cuneus	100%	47.33%	57.45%

Table 21: Testing models from Table 21, that use brain regions, obtained with cross-validation

Region	Train B. Accuracy	B. Accuracy	Accuracy
posteriorcingulate RC	64.36	45.26%	45.74%
paracentral	53.10%	54.08%	56.38%
cerebellum	58.28%	53.18%	53.19%
posteriorcingulate LR	63.68%	46.77%	45.74%
insula	100%	47.33%	57.45%

have a positive result from a DaTscan. One of more important ones would be to gather more data, which is an effort being pursued by Diana Prata's Lab, although medical data and MRIs is not data that is easily scalable. Using more homogeneous groups can be advantages, by separating the negative group, and then solving two binary problems. However, this would make the data available even more unbalanced.

Another option is consider using other type of features extracted from Structural MRI, or use another modality such as DTI to try to replicate the results that exist in the literature. Furthermore, More Structural MRI can be used if for selecting them less restrictions on the parameters are made, but this can bring problems since the images may

be different enough that the models would detect them, and differentiate between them instead of the label that is wanted.

Finally more complex models, such as CNN or other deep learning methods may bring more insights, but since there is not a lot of data, this is not a clear path to take.

Acknowledgements

I want to thank my supervisors, David Matos and Diana Prata, and all the people with whom I worked with from Diana Prata's Lab, Vasco and Helena specially, for all the help and support.

I want to thank my parents for always supporting me and my brother for annoying me but still helping me.

Last but not least, I want to thank my friends that heard me through all my university years, Mariana Carrasco for always being there and for all the late night projects, João Mira for the support and our meetings, Henrique Santos for all the company and support and Ted for the distraction and always being available when I ask.

References

[1] M. Robert A Hauser. Parkinson disease, Apr 2021. URL https://emedicine.medscape.com/article/1831191-overview#a2.

[2] M. J. Armstrong and M. S. Okun. Diagnosis and Treatment of Parkinson Disease: A Review. *JAMA*, 323(6):548–560, 02 2020. ISSN 0098-7484. doi: 10.1001/jama.2019.22360. URL https://doi.org/10.1001/jama.2019.22360.

[3] C. Blauwendraat, M. A. Nalls, and A. B. Singleton. The genetic architecture of parkinson's disease. *The Lancet Neurology*, 19(2):170–178, Feb. 2020. doi: 10.1016/s1474-4422(19)30287-x. URL https://doi.org/10.1016/s1474-4422(19)30287-x.

[4] C. A. Haaxma, B. R. Bloem, G. F. Borm, W. J. G. Oyen, K. L. Leenders, S. Eshuis, J. Booij, D. E. Dluzen, and M. W. I. M. Horstink. Gender differences in parkinson's disease. 78(8):819–824, Aug. 2007. doi: 10.1136/jnnp.2006.103788. URL https://doi.org/10.1136/jnnp.2006.103788.

[5] J. Jankovic. Parkinson's disease: clinical features and diagnosis. *Journal of Neurology, Neurosurgery & Psychiatry*, 79(4):368–376, 2008. ISSN 0022-3050. doi: 10.1136/jnnp.2007.131045. URL https://jnnp.bmj.com/content/79/4/368.

[6] G. Rizzo, M. Copetti, S. Arcuti, D. Martino, A. Fontana, and G. Logroscino. Accuracy of clinical diagnosis of parkinson disease. *Neurology*, 86(6):566–576, 2016. ISSN 0028-3878. doi: 10.1212/WNL.0000000000002350. URL https://n.neurology.org/content/86/6/566.

[7] R. de la Fuente-Fernandez. Role of DaTSCAN and clinical diagnosis in parkinson disease. *Neurology*, 78(10):696–701, Feb. 2012. doi: 10.1212/wnl.0b013e318248e520. URL https://doi.org/10.1212/wnl.0b013e318248e520.

[8] R. Erro, S. A. Schneider, M. Stamelou, N. P. Quinn, and K. P. Bhatia. What do patients with scans without evidence of dopaminergic deficit (SWEDD) have? new evidence and continuing controversies. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(3):319–323, May 2015. doi: 10.1136/jnnp-2014-310256. URL https://doi.org/10.1136/jnnp-2014-310256.

[9] R. Deb, G. Bhat, S. An, H. Shill, and U. Y. Ogras. Trends in technology usage for parkinson's disease assessment: A systematic review. Feb. 2021. doi: 10.1101/2021.02.01.21250939. URL https://doi.org/10.1101/2021.02.01.21250939.

[10] J. M. Valverde, V. Imani, A. Abdollahzadeh, R. D. Feo, M. Prakash, R. Ciszek, and J. Tohka. Transfer learning in magnetic resonance brain imaging: A systematic review. *Journal of Imaging*, 7(4):66, Apr. 2021. doi: 10.3390/jimaging7040066. URL https://doi.org/10.3390/jimaging7040066.

[11] A. Segato, A. Marzullo, F. Calimeri, and E. D. Momi. Artificial intelligence for brain diseases: A systematic review. *APL Bioengineering*, 4 (4):041503, Dec. 2020. doi: 10.1063/5.0011697. URL https://doi.org/10.1063/5.0011697.

[12] L. Zhang, M. Wang, M. Liu, and D. Zhang. A survey on deep learning for neuroimaging-based brain disorder analysis. *Frontiers in Neuroscience*, 14, Oct. 2020. doi: 10.3389/fnins.2020.00779. URL https://doi.org/10.3389/fnins.2020.00779.

[13] A. A.-A. Valliani, D. Ranti, and E. K. Oermann. Deep learning and neurology: A systematic review. *Neurology and Therapy*, 8(2):351–365, Aug. 2019. doi: 10.1007/s40120-019-00153-8. URL https://doi.org/10.1007/s40120-019-00153-8.

[14] A. D. Yao, D. L. Cheng, I. Pan, and F. Kitamura. Deep learning in neuroradiology: A systematic review of current algorithms

and approaches for the new wave of imaging technology. *Radiology: Artificial Intelligence*, 2(2):e190026, Mar. 2020. doi: 10.1148/ryai.2020190026. URL https://doi.org/10.1148/ryai.2020190026.

[15] J. M. Mateos-Pérez, M. Dadar, M. Lacalle-Aurioles, Y. Iturria-Medina, Y. Zeighami, and A. C. Evans. Structural neuroimaging as clinical predictor: A review of machine learning applications. *NeuroImage: Clinical*, 20:506–522, 2018. doi: 10.1016/j.nicl.2018.08.019. URL https://doi.org/10.1016/j.nicl.2018.08.019.

[16] J. Xu and M. Zhang. Use of magnetic resonance imaging and artificial intelligence in studies of diagnosis of parkinson's disease. *ACS Chemical Neuroscience*, 10(6):2658–2667, May 2019. doi: 10.1021/acschemneuro.9b00207. URL https://doi.org/10.1021/acschemneuro.9b00207.

[17] K. R. Bhatele and S. S. Bhadauria. Brain structural disorders detection and classification approaches: a review. *Artificial Intelligence Review*, 53(5):3349–3401, Oct. 2019. doi: 10.1007/s10462-019-09766-9. URL https://doi.org/10.1007/s10462-019-09766-9.

[18] E. U. Haq, J. Huang, L. Kang, H. U. Haq, and T. Zhan. Image-based state-of-the-art techniques for the identification and classification of brain diseases: a review. *Medical & Biological Engineering & Computing*, 58 (11):2603–2620, Sept. 2020. doi: 10.1007/s11517-020-02256-z. URL https://doi.org/10.1007/s11517-020-02256-z.

[19] L. Jin, Q. Zeng, J. He, Y. Feng, S. Zhou, and Y. Wu. A ReliefF-SVM-based method for marking dopamine-based disease characteristics: A study on SWEDD and parkinson's disease. *Behavioural Brain Research*, 356:400–407, Jan. 2019. doi: 10.1016/j.bbr.2018.09.003. URL https://doi.org/10.1016/j.bbr.2018.09.003.

[20] E. Matsusue, Y. Fujihara, K. Tanaka, Y. Aozasa, M. Shimoda, H. Nakayasu, K. Nakamura, and T. Ogawa. The utility of the combined use of 123i-FP-CIT SPECT and neuromelanin MRI in differentiating parkinson's disease from other parkinsonian syndromes. *Acta Radiologica*, 60(2):230–238, May 2018. doi: 10.1177/0284185118778871. URL https://doi.org/10.1177/0284185118778871.

[21] M. Kim and H. Park. Structural connectivity profile of scans without evidence of dopaminergic deficit (SWEDD) patients compared to normal controls and parkinson's disease patients. *SpringerPlus*, 5(1), Aug. 2016. doi: 10.1186/s40064-016-3110-8. URL https://doi.org/10.1186/s40064-016-3110-8.

[22] M. Kim and H. Park. Using tractography to distinguish SWEDD from parkinson's disease patients based on connectivity. *Parkinson's Disease*, 2016:1–10, 2016. doi: 10.1155/2016/8704910. URL https://doi.org/10.1155/2016/8704910.

[23] H. R. Pereira. *Classification of patients with parkinsonian syndromes using medical imaging and artificial intelligence algorithms*. Universidade Nova de Lisboa, 2018. URL https://run.unl.pt/handle/10362/61556.

[24] Freesurfer, Apr 2021. URL https://surfer.nmr.mgh.harvard.edu/fswiki.

[25] Freesurfer pipeline, Apr 2021. URL https://surfer.nmr.mgh.harvard.edu/fswiki/FreeSurferAnalysisPipelineOverview.

[26] Ppmi. URL https://www.ppmi-info.org/.

[27] V. Tavares, D. Prata, and H. A. Ferreira. Comparing SPM12 and CAT12 segmentation pipelines: a brain tissue volume-based age and alzheimer's disease study. *Journal of Neuroscience Methods*, 334:108565, Mar. 2020. doi: 10.1016/j.jneumeth.2019.108565. URL https://doi.org/10.1016/j.jneumeth.2019.108565.

[28] C. A. Haaxma, B. R. Bloem, G. F. Borm, W. J. G. Oyen, K. L. Leenders, S. Eshuis, J. Booij, D. E. Dluzen, and M. W. I. M. Horstink. Gender differences in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 78(8):819–824, Aug. 2007. doi: 10.1136/jnnp.2006.103788. URL https://doi.org/10.1136/jnnp.2006.103788.