2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

SPECT Image Features for Early Detection of Parkinson's Disease using Machine Learning Methods*

Emmi Antikainen[1], Patrick Cella[2], Antti Tolonen[1], and Mark van Gils[1], *Member, IEEE*

***Abstract*— Millions of people around the world suffer from Parkinson's disease, a neurodegenerative disorder with no remedy. Currently, the best response to interventions is achieved when the disease is diagnosed at an early stage. Supervised machine learning models are a common approach to assist early diagnosis from clinical data, but their performance is highly dependent on available example data and selected input features. In this study, we explore 23 single photon emission computed tomography (SPECT) image features for the early diagnosis of Parkinson's disease on 646 subjects. We achieve 94 % balanced classification accuracy in independent test data using the full feature space and show that matching accuracy can be achieved with only eight features, including original features introduced in this study. All the presented features can be generated using a routinely available clinical software and are therefore straightforward to extract and apply.**

***Clinical relevance*— This work evaluates SPECT image features available from routinely used clinical software to support early diagnosis of Parkinson's disease and establishes high accuracy results.**

I. INTRODUCTION

Parkinson's disease (PD) is a devastating neurodegenerative disease with over 6 million diagnosed cases worldwide [1]. In the United States alone, it induces over $50 billion annual costs [2]. The gradually progressive symptoms can include slowness of movement (bradykinesia), tremor, muscle stiffness, problems with balance and speech, and autonomic dysfunction along with other symptoms [3]. With no existing cure for PD, early diagnosis and monitoring of the disease severity is essential in order to provide correct, timely interventions and more effective treatment.

Neuroimaging serves as an ancillary diagnostic tool for differential diagnosis of PD [4]. Among different imaging techniques, single photon emission computed tomography (SPECT) with a dopamine-transporter (DAT) binding tracer such as DaTSCAN (Ioflupane I-123) stands out as the most sensitive method for detecting early PD [4]. In SPECT, the radioactive tracer is injected into the bloodstream where it binds to active dopaminergic neurons. Specifically, DAT-SPECT provides information on the presynaptic dopaminergic transporters in the striatum, and this may be used to successfully discriminate between cases of early stage PD and subjects without striatal dopaminergic deficiency (e.g. healthy controls and patients with essential tremor) [5], [6].

Machine learning (ML) based decision support can provide objectivity and increased accuracy for diagnostics. Especially when the diagnosis relies on the visual inspection of medical images, ML can help physicians save time while improving both consistency and interobserver agreement, which in turn may improve the patient outcome. ML solutions have reached equal or even higher accuracy as compared to medical experts in diagnosing diseases from images [7], [8]. Thus, early diagnosis of PD from SPECT scans may be boosted with ML, specifically through supervised classification.

Supervised ML methods are able to yield multivariate classification models from labelled databases to distinguish patients with PD from healthy controls. Moreover, many conventional ML methods are explainable and interpretable, enabling the physicians to review the justification of the suggested diagnosis. Yet, the accuracy of such methods is highly dependent on the selected input features and their capacity for the classification task: they need to describe useful and relevant information to facilitate reliable classification.

In this study, we evaluate the potential of 23 SPECT image features for early diagnosis of PD. Fourteen features are obtained by processing the DAT-SPECT image with the routinely used clinical DaTQUANT software (GE Healthcare, Chicago, IL, USA) and the other 9 features are derived from those. We explore several ML classifiers to obtain the best prediction model and examine the importance of the features to gain insights for future work. We employ two distinct databases (276 and 370 subjects, 646 in total), using one for training and the other for testing, to enable reliable evaluation of both classification performance and feature importance.

Preceding studies have demonstrated promising results for early detection of PD by applying ML on various diagnostic modalities [9], [10], [11], [12]. Relevantly, Prashanth et al. used a 548 subject database with four striatal binding ratio (SBR) features from SPECT images [9]. They used Support Vector Machines (SVM) and Logistic Regression (LR) for classification, achieving an impressive 96 % accuracy. However, they used 10-fold cross validation, which may give optimistic figures as compared to an independent test set. Other ML based studies have also reported high accuracies for this task but with relatively small datasets [13], [14], [15].

This study demonstrates that eight SPECT image features are sufficient for high accuracy classification of early PD. To the best of our knowledge, two of the features are novel,

*This work was supported by GE Healthcare UK

[1]Emmi Antikainen is with VTT Technical Research Centre of Finland Ltd., 33101 Tampere, Finland emmi.antikainen@vtt.fi

[2]Patrick Cella is with GE Healthcare, MA 01752, USA patrick.cella@ge.com

[1] Antti Tolonen was with VTT Technical Research Centre of Finland Ltd. and is now with Combinostics Ltd., 33100 Tampere, Finland, antti.tolonen@combinostics.com

[1] Mark van Gils was with VTT Technical Research Centre of Finland Ltd. and is now with Faculty of Medicine and Health Technology, Tampere University, 33720 Tampere, Finland, mark.vangils@tuni.fi

whereas all of them are either directly available or derived from DaTQUANT. These features can be further studied for early PD indicators or applied in clinical decision support.

II. MATERIALS AND METHODS

We extracted the 23 SPECT image features together with two complementing demographic features from two distinct databases. Using one database at a time for training and the other for testing, we developed several supervised classifiers to obtain the best prediction model for PD. Finally, we evaluated feature importance and examined the classifiers again with the most promising set of features.

A. Study cohort and data

The study data (see Table I) consists of two independent, retrospective databases collected in distinct studies [16], [17]. The included data covers a total of 646 subjects (299 positive for PD) from age 30 to 89 (65.6 on average, 246 females).

TABLE I
DATABASE DETAILS.

Database	Subjects	Females	Males	Age range	PD cases
DB1	276	125	151	38-89	124
DB2	370	121	249	30-84	175

Database DB1 was collected by Booij et al. and consists of three multicenter trials [16]. It contains some cases of dementia with Lewy bodies (DLB) which results in similar appearance of the scan but derives from a different pathology. Thus, we consider them positive PD cases. The patients were diagnosed by at least two clinicians in consensus in two trials, and by one clinician in the other trial. We excluded 28 cases with uncertain diagnosis, leaving 276 SPECT images.

Database DB2 was collected by Marek et al. [17]. From the 32-month multisite study, 195 healthy controls and 175 PD cases were randomly selected for this study. The diagnoses for the patients were provided by one experienced clinician at each study site.

B. Feature extraction

All included features are presented in Table II. The images were processed using the DaTQUANT software (GE Healthcare, Chicago IL). It automatically calculates tracer uptake as compared to the background region in the occipital cortex, i.e., striatal binding ratio, over several striatal regions. It also generates the putamen to caudatus uptake ratio on the right and left sides individually, along with the uptake asymmetry between the two sides for both caudate and putamen.

The directly available feature set was augmented by the most affected side (i.e. lower SBR of either right or left) of putamen and posterior putamen, and seven features, which as far as we know are original. They exploit a comparison to a built-in database of normal subjects, which DaTQUANT employs to correct for normal age-related decline. The comparison yields z-scores and percentage of deviation from the mean of age-matched normal uptake. Four of the novel features describe the number of regions deviating n standard deviations (SD) from normal uptake, where $n \in \{1.5, 1.6, 1.7, 2.0\}$. They count the caudate and the anterior and posterior putamen regions where $z < -n$. The remaining features describe the length of the striatum (right, left, and the most affected side), defined as the number of contiguous caudate segments (0 to 3) for which $z > -1.6$, a threshold selected for abnormal uptake. However, as PD tends to progress through the putamen to the caudate, the length was set to zero if the caudate $z < -1.6$, even if other parts of the ipsilateral striatum were above the threshold.

TABLE II
DEMOGRAPHIC AND SPECT IMAGE-BASED FEATURES.

Feature	Availability from DaTQUANT
Age	-
Striatum, right	directly available
Striatum, left	directly available
Putamen, right	directly available
Putamen, left	directly available
Putamen, most affected side	derived
Caudate, right	directly available
Caudate, left	directly available
Anterior putamen, right	directly available
Anterior putamen, left	directly available
Posterior putamen, right	directly available
Posterior putamen, left	directly available
Posterior putamen, most affected side	derived
Putamen to caudatus ratio, right	directly available
Putamen to caudatus ratio, left	directly available
Caudatus asymmetry	directly available
Putamen asymmetry	directly available
No. of regions 1.5 SD from normal	derived
No. of regions 1.6 SD from normal	derived
No. of regions 1.7 SD from normal	derived
No. of regions 2.0 SD from normal	derived
Length of right striatum	derived
Length of left striatum	derived
Length of most affected striatum	derived
Sex	-

To complement the image features, we included two demographic features; age and sex. The categorical variable sex was transformed into two one-hot encoded features; female (F), and male (M). Moreover, the features were normalized by reducing the mean and adjusted to unit variance, to avoid biased input weighting due to differently ranged features.

C. Supervised classification models

We studied ML methods that have been proven to be successful in similar applications, based on their performances and robustness: SVM, LR, linear discriminant analysis (LDA), random forest (RF), gradient boosted (GB) regression trees, and AdaBoost (AB) [18], [19].

For SVM, we used radial basis function (RBF) kernel to enable non-linear classification boundaries. For LR, we used L2 regularization, which is also employed in SVMs.

D. Model validation

The hyperparameters for each classifier were optimized by applying stratified 10-fold cross-validation in the training data set. Subsequently, the classifiers were trained using the full training database and tested on the held out database.

TABLE III

CLASSIFIER (COLUMNS) PERFORMANCE (%, ROWS) WITH THE 95 % CONFIDENCE INTERVAL WHEN TRAINED ON DB1 AND TESTED ON DB2.

	SVM	LR	LDA	RF	GB	AB
Accuracy	93.8±2.5	93.5±2.5	83.5±3.8	**94.6**±2.3	94.3±2.4	94.1±2.4
Balanced accuracy	93.7±2.5	93.4±2.5	83.5±3.8	**94.4**±2.3	94.2±2.4	93.9±2.4
PPV	95.2±3.2	95.2±3.2	82.0±5.6	**97.0**±2.6	96.4±2.8	96.4±2.9
NPV	92.6±3.6	92.1±3.7	84.9±5.1	**92.7**±3.6	92.6±3.6	92.2±3.7
Sensitivity	**91.4**±4.1	90.9±4.3	83.4±5.5	**91.4**±4.1	**91.4**±4.1	90.9±4.3
Specificity	95.9±2.8	95.9±2.8	83.6±5.2	**97.4**±2.2	96.9±2.4	96.9±2.4
AUC	93.7±2.7	93.4±2.7	83.5±4.2	**94.4**±2.5	94.2±2.5	93.9±2.6

TABLE IV

CLASSIFIER (COLUMNS) PERFORMANCE (%, ROWS) WITH THE 95 % CONFIDENCE INTERVAL WHEN TRAINED ON DB2 AND TESTED ON DB1.

	SVM	LR	LDA	RF	GB	AB
Accuracy	**94.6**±2.7	89.9±3.6	92.8±3.1	93.8±2.8	91.3±3.3	90.9±3.4
Balanced accuracy	**94.6**±2.7	90.0±3.5	92.9±3.0	94.0±2.8	91.4±3.3	91.3±3.3
PPV	**92.9**±4.5	86.9±5.8	90.0±5.2	90.8±4.9	89.1±5.4	86.1±5.8
NPV	96.0±3.2	92.5±4.3	95.2±3.5	**96.6**±3.0	93.2±4.0	95.7±3.4
Sensitivity	96.0±3.8	91.1±5.0	94.4±4.1	**96.0**±3.5	91.9±4.8	95.2±3.8
Specificity	**94.1**±3.8	88.8±5.0	91.4±4.4	92.1±4.3	90.8±4.6	87.5±5.3
AUC	**94.6**±2.9	90.0±3.9	92.9±3.3	94.0±3.0	91.4±3.7	91.3±3.7

Then, the roles of the databases were switched to gain a more realistic estimate of model generalization to new data.

The selected performance metrics are accuracy, balanced accuracy, positive predictive value (PPV), negative predictive value (NPV), sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). Given N test subjects, we apply normal approximation to estimate 95 % binomial confidence interval

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/N}, \tag{1}$$

for all metrics $\hat{p}$ apart from AUC, whose confidence interval is estimated using the standard error defined in [20].

E. Feature importance analysis

Feature importance is directly available for some classifiers, e.g. in the weights of LR models, or in the mean decrease of impurity for tree-based models. However, for some models it is more complicated to derive. Therefore, we additionally applied univariate feature selection to assess feature importance on a general level. Correlation analysis using Pearson correlation coefficient was selected to further complement the analysis. Feature importance was evaluated within both databases individually, using the models trained on the corresponding full database.

Univariate selection is based on univariate statistical tests between the features and labels; the better the score, the more important the feature. Here we used the following scores: χ^2, ANOVA F-value (linear dependency), and mutual information (non-parametric).

III. RESULTS

A. Classification

Tables III and IV present the performance of each classifier on the two databases. For the tree-based methods, best results occurred at 100 decision trees. As seen in Table III, the RF classifier reached the most promising results in generalizing from the smaller DB1 to the samples of the larger DB2. In contrast, training on the larger DB2, the SVM classifier benefited from hyperparameter optimization and achieved more promising performance in terms of most metrics (see Table IV). Considering both training scenarios, the tree-based RF and GB classifiers and SVM show the best results.

B. Feature importance

Fig. 1 presents feature correlations within DB1 and DB2. The features can be divided into groups, which show low to moderate intergroup correlation, while consisting of highly correlated features within the group. The subset of features describing the anatomy of the brain were highly correlated with each other, excluding ratio-, asymmetry-, and length-based features. Similarly, the number of abnormal regions and length-based features showed high correlation, while correlating moderately with the anatomy-related features. The ratio- and asymmetry-based features on the other hand demonstrated higher correlation with the rest of the features in DB2 than they did in DB1. Age did not correlate significantly with any other features in either case.

The feature importance rankings in the two training scenarios are presented in Fig. 2. Importantly, univariate selection and tree-based importance ranked features similarly with respect to each other in both databases. Also LR weights gave low importance to age and sex, while still finding some of the same features important for classification, especially in DB1, where LR also reached closer to the RF performance.

Interestingly, the caudate features as well as ratio- and asymmetry-based features ranked mostly low in importance. Furthermore, when there were three alternatives to describe an anatomical feature (right, left, and the most affected side), the most affected side was ranked the most important (except by LR weights). The number of regions deviating more than

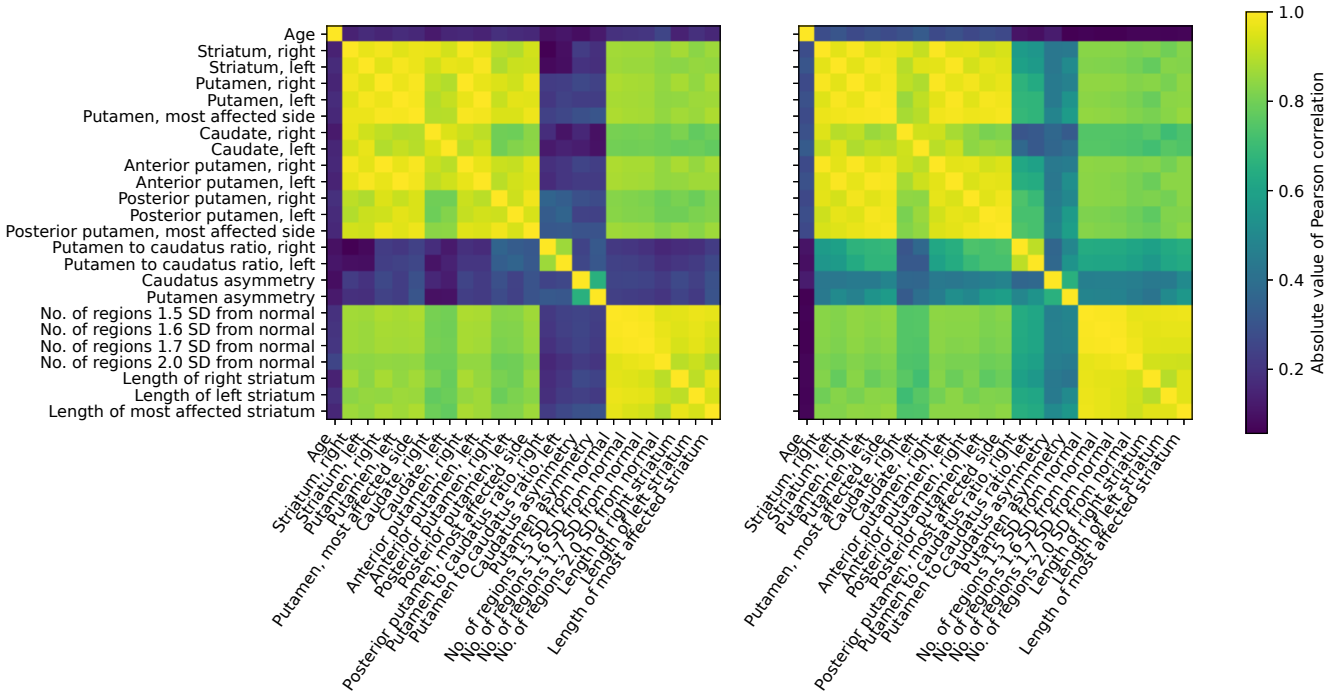

Fig. 1. Feature correlation (absolute value) within the DB1 (left) and the DB2 (right) databases.

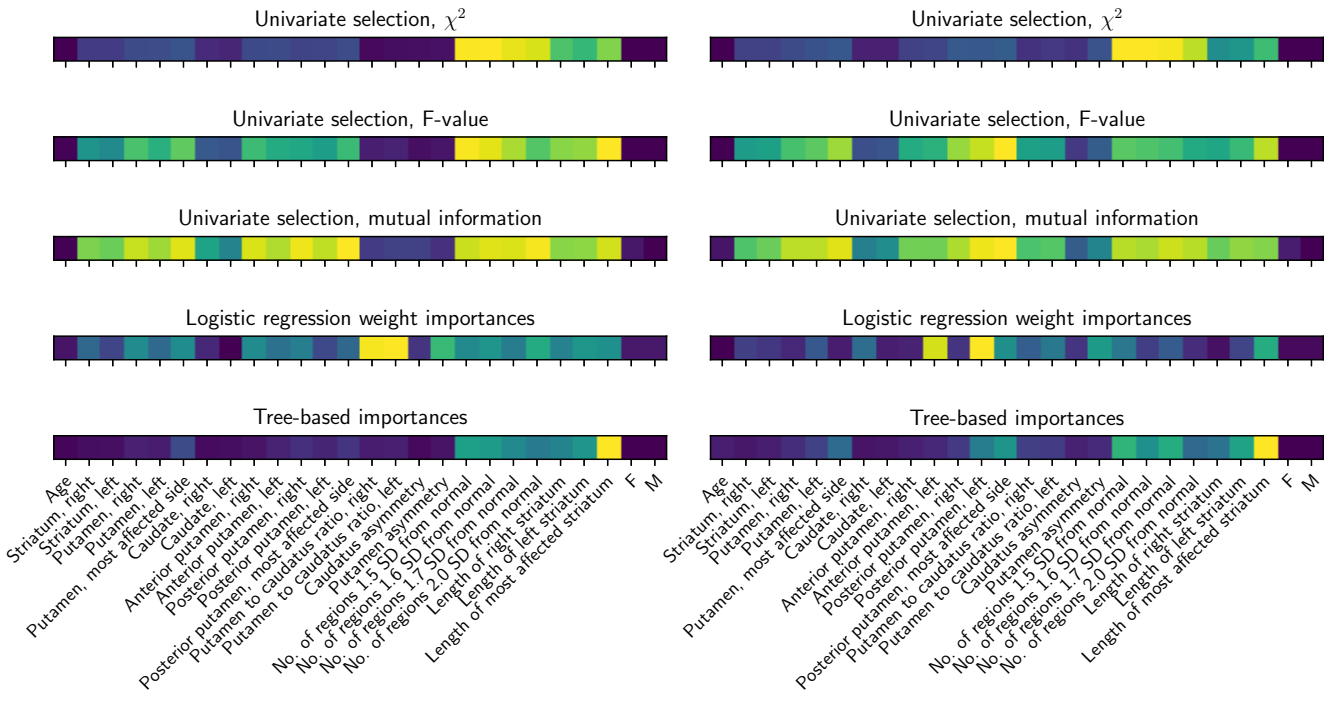

Fig. 2. Feature importance within the DB1 (left) and DB2 (right) datasets. For each method, i.e. each row, the brightest yellow indicates the most important feature while the darkest blue indicates the least important feature.

TABLE V

CLASSIFIER (COLUMNS) PERFORMANCE (%, ROWS) WITH THE 95 % CONFIDENCE INTERVAL USING THE REDUCED FEATURE SET ON THE MOST PROMISING CLASSIFIERS.

	SVM	RF	GB
Trained on DB1, tested on DB2			
Accuracy	93.2±2.6	**94.3**±2.4	92.4±2.7
Balanced accuracy	93.2±2.6	**94.1**±2.4	92.3±2.7
PPV	93.6±3.7	**97.0**±2.6	94.5±3.5
NPV	**92.9**±3.6	92.2±3.7	90.7±4.0
Sensitivity	**92.0**±4.0	90.9±4.3	89.1±4.6
Specificity	94.4±3.2	**97.4**±2.2	95.4±2.9
AUC	93.2±2.7	**94.1**±2.5	92.3±2.9
Trained on DB2, tested on DB1			
Accuracy	92.4±3.1	**93.5**±2.9	92.4±3.1
Balanced accuracy	92.6±3.1	**93.7**±2.9	92.6±3.1
PPV	88.7±5.4	**90.2**±5.1	89.3±5.3
NPV	95.8±3.3	**96.5**±3.0	95.2±3.5
Sensitivity	95.2±3.8	**96.0**±3.5	94.4±4.1
Specificity	90.1±4.7	**91.4**±4.4	90.8±4.6
AUC	92.6±3.4	**93.7**±3.1	92.6±3.4

1.5 SD below normal seemed the most important among similar features, except when ranked by mutual information.

Based on these results, we coupled four of the most important features (most affected side of both the putamen and the posterior putamen, length of the most affected striatum, and number of regions 1.5 SD below normal) with four features of limited importance (caudate right and left, caudatus asymmetry, and putamen asymmetry); the latter showed moderate or low correlation to the most important features and may hence efficiently contribute to discriminating between the two classes. The results achieved on the most promising models with the reduced feature set are presented in Table V. The results suffer only a minor decline, if any, after including only one third of the original feature set.

IV. DISCUSSION

This study explored 23 SPECT image features, coupled with age and sex, for early diagnosis of Parkinson's disease. We established 94 % balanced accuracy for classification using random forest. Importantly, we analyzed feature importance and showed similar results with only eight SPECT image features. The partly novel features derive from the DaTQUANT software and may thus be effortlessly replicated. As compared to previous works, improved model generalization to new data is expected due to the use of a fully independent test set [9].

Further improvements in the prediction accuracy may be reached via modern deep learning (DL) methods that do not require prior feature engineering, such as convolutional neural networks (CNN), possibly coupled with other approaches, e.g. generative adversarial networks (GAN). Yet, adaptation of such approaches to clinical use may depend on the future availability of explainable DL solutions and abundant data.

ACKNOWLEDGMENT

The authors thank Manoj Hegde (at GE Healthcare) for the work to process and quantify the exams with DaTQUANT.

REFERENCES

[1] E. R. Dorsey *et al.*, "Global, regional, and national burden of parkinson's disease, 1990-2016: a systematic analysis for the global burden of disease study 2016," *The Lancet Neurology*, vol. 17, no. 11, pp. 939–953, 2018.

[2] W. Yang, J. L. Hamilton, C. Kopil, J. C. Beck, C. M. Tanner, R. L. Albin, E. R. Dorsey, N. Dahodwala, I. Cintina, P. Hogan, and T. Thompson, "Current and projected future economic burden of parkinson's disease in the u.s." *npj Parkinson's disease*, vol. 6, no. 15, 2020.

[3] M. Politis, K. Wu, S. Molloy, P. G. Bain, K. R. Chaudhuri, and P. Piccini, "Parkinson's disease symptoms: The patient's perspective," *Movement Disorders*, vol. 25, no. 11, pp. 1646–1651, 2010.

[4] E. Tolosa, G. Wenning, and W. Poewe, "The diagnosis of parkinson's disease," *The Lancet Neurology*, vol. 5, no. 1, pp. 75 – 86, 2006.

[5] S. Hesse, C. Oehlwein, H. Barthel, J. Schwarz, D. Polster, A. Wagner, and O. Sabri, "Possible impact of dopamine spect on decision-making for drug treatment in parkinsonian syndrome," *Journal of Neural Transmission*, vol. 113, pp. 1177–1190, 2006.

[6] A. M. Catafau and E. Tolosa, "Impact of dopamine transporter spect using 123i-ioflupane on diagnosis and management of patients with clinically uncertain parkinsonian syndromes," *Movement Disorders*, vol. 19, no. 10, pp. 1175–1182, 2004.

[7] D. Li, K. Kulasegaram, and B. D. Hodges, "Why we needn't fear the machines: Opportunities for medicine in a machine learning world," *Academic Medicine*, vol. 94, pp. 623–625, 2019.

[8] X. Liu *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The Lancet Digital Health*, vol. 1, no. 6, pp. e271 – e297, 2019.

[9] R. Prashanth, S. Dutta Roy, P. K. Mandal, and S. Ghosh, "Automatic classification and prediction models for early parkinson's disease diagnosis from spect imaging," *Expert Systems with Applications*, vol. 41, no. 7, pp. 3333 – 3342, 2014.

[10] R. Prashanth, S. Dutta Roy, P. K. Mandal, and S. Ghosh, "High-accuracy detection of early parkinson's disease through multimodal features and machine learning," *International Journal of Medical Informatics*, vol. 90, pp. 13–21, 2016.

[11] C. Kotsavasiloglou, N. Kostikis, D. Hristu-Varsakelis, and M. Arnaoutoglou, "Machine learning-based classification of simple drawing movements in parkinson's disease," *Biomedical Signal Processing and Control*, vol. 31, pp. 174 – 180, 2017.

[12] J. S. Almeida, P. P. Rebouças Filho, T. Carneiro, W. Wei, R. Damaševičius, R. Maskeliūnas, and V. H. C. de Albuquerque, "Detecting parkinson's disease with sustained phonation and speech signals using machine learning techniques," *Pattern Recognition Letters*, vol. 125, pp. 55 – 62, 2019.

[13] I. A. Illán, J. M. Górriz, J. Ramírez, F. Segovia, J. M. Jiménez-Hoyuela, and S. J. Ortega Lozano, "Automatic assistance to parkinson's disease diagnosis in datscan spect imaging," *Medical Physics*, vol. 39, no. 10, pp. 5971–5980, 2012.

[14] A. Rojas, J. Górriz, J. Ramírez, I. Illán, F. Martínez-Murcia, A. Ortiz, M. Gómez Río, and M. Moreno-Caballero, "Application of empirical mode decomposition (emd) on datscan spect images to explore parkinson disease," *Expert Systems with Applications*, vol. 40, no. 7, pp. 2756 – 2766, 2013.

[15] D. J. Towey, P. G. Bain, and K. S. Nijran, "Automatic classification of 123i-fp-cit (datscan) spect images," *Nuclear Medicine Communications*, vol. 32, pp. 699–707, 2011.

[16] J. Booij, J. Dubroff, D. Pryma, J. Yu, R. Agarwal, Lakhani, P., and P. H. Kuo, "Diagnostic performance of the visual reading of 123i-ioflupane spect images with or without quantification in patients with movement disorders or dementia," *Journal of nuclear medicine*, vol. 58, no. 11, pp. 1821–1826, 2017.

[17] K. Marek *et al.*, "The parkinson's progression markers initiative (ppmi) – establishing a pd biomarker cohort," *Annals of Clinical and Translational Neurology*, vol. 5, no. 12, pp. 1460–1477, 2018.

[18] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*. New York, USA: Springer, 2013.

[19] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge, MA, USA: The MIT Press, 2012.

[20] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (roc) curve." *Radiology*, vol. 143, no. 1, pp. 29–36, 1982.