Intelligence-Based Medicine 11 (2025) 100232

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.sciencedirect.com/journal/intelligence-based-medicine

Image-based machine learning model as a tool for classification of [^{18}F] PR04.MZ PET images in patients with parkinsonian syndrome

Maria Jiménez [a,b,*], Cristian Soza-Ried [b,c,d], Vasko Kramer [b,c], Sebastian A. Ríos [e], Arlette Haeger [b], Carlos Juri [f,g], Horacio Amaral [b,c], Pedro Chana-Cuevas [h,i]

[a] Industrial Engineering Department, University of Chile, Chile
[b] Nuclear Medicine and PET/CT Center PositronMed, Santiago, Chile
[c] PositronPharma SA, Santiago, Chile
[d] Universidad de las Américas, Veterinary Medicine and Agronomy Faculty, Natural Science Insittute, Santiago, Chile
[e] Business Intelligence Research Center, Industrial Engineering Department, University of Chile, Chile
[f] Department of Neurology, Faculty of Medicine, Pontifical Catholic University of Chile, Santiago, Chile
[g] Department of Neurology, Hospital Sotero del Río, Santiago, Chile
[h] Movement Disorders Center, Santiago, Chile
[i] Faculty of Medical Sciences, University of Santiago de Chile, Santiago, Chile

ARTICLE INFO

Keywords:
Parkinson's disease
Machine learning
Positron emission tomography
[^{18}F]PR04.MZ PET tracer

ABSTRACT

Parkinsonian syndrome (PS) is characterized by bradykinesia, resting tremor, rigidity, and encapsulates the clinical manifestation observed in various neurodegenerative disorders. Positron emission tomography (PET) imaging plays an important role in diagnosing PS by detecting the progressive loss of dopaminergic neurons. This study aimed to develop and compare five machine-learning models for the automatic classification of 204 [^{18}F] PR04.MZ PET images, distinguishing between patients with PS and subjects without clinical evidence for dopaminergic deficit (SWEDD). Previously analyzed and classified by three expert blind readers into PS compatible (1) and SWEDDs (0), the dataset was processed in both two-dimensional and three-dimensional formats. Five widely used pattern recognition algorithms were trained and validated their performance. These algorithms were compared against the majority reading of expert diagnosis, considered the gold standard. Comparing the accuracy of 2D and 3D format images suggests that, without the depth dimension, a single image may overemphasize specific regions. Overall, three models outperformed with an accuracy greater than 98 %, demonstrating that machine-learning models trained with [^{18}F]PR04.MZ PET images can provide a highly accurate and precise tool to support clinicians in automatic PET image analysis. This approach may be a first step in reducing the time required for interpretation, as well as increase certainty in the diagnostic process.

1. Introduction

Parkinsonian syndrome (PS) is an umbrella term for a constellation of heterogeneous and complex neurodegenerative disorders involving dopaminergic neurons within the substantia nigra pars compacta (SNpc) [1–3]. Among PS, Parkinson's disease (PD) is the most common, and its prevalence is dramatically rising in Latin American countries [4–6], a trend attributable to industrialization and enhanced life expectancy [4].

The diagnosis of PS is mainly based on clinical manifestations [7], requiring a high degree of expertise. To assist in this process the International Parkinson and Movement Disorder Society (MDS) published [8] and validated [9,10] guidelines to assure reproducibility in the diagnosis across medical centers and clinicians. These guidelines have been updated to include information on comorbidities such as diabetes mellitus (type II) [11–13], cognitive deficits [14–16], low plasma urate levels in men [17–20] and lifestyle factors such as physical inactivity [21–28], known to increase risk of PD. This updated MDS criteria now emphasized the utility of neuroimaging [29–34] to address the challenges of earlier and differential PS diagnosis, enhancing diagnostic certainty and reinforcing clinician's decisions [35,36]. The reduction of presynaptic biomarkers, such as dopamine transporters (DAT), indicates the loss of nigrostriatal neurons and the subsequent dopaminergic deficit in the striatal regions of the brain. [^{18}F]PR04.MZ is a PET tracer with high affinity, selectivity, and specificity for DAT, suitable for detecting

* Corresponding author. Nuclear Medicine and PET/CT Center PositronMed, Santiago, Chile.
E-mail address: mjimenag52@gmail.com (M. Jiménez).

https://doi.org/10.1016/j.ibmed.2025.100232
Received 27 December 2024; Received in revised form 17 February 2025; Accepted 2 March 2025
Available online 4 March 2025

dopaminergic deficit in the striatum and SNpc even before the onset of motor symptoms [37–42].

However, the process of image processing, modeling, and statistical analysis, requires specialized software packages, often limited to certain clinical centers. Furthermore, analyzing scan images, particularly those of complex anatomical organs like the brain, requires significant skill and experience to accurately detect anomalies or changes. While new radiotracers such as [^{18}F]PR04.MZ are promising, their routine clinical application relies on reproducible methods of PET imaging quantification to minimize variability between readers and produce standardized clinical reports. Artificial intelligence (AI) offers a compelling approach to interpreting PET imaging data. The rapid advancements in machine and deep learning applied to diagnostic neuroimaging have demonstrated promising outcomes in various clinical scenarios [43–49]. Computer-aided diagnostic (CAD) methods using neuroimaging exhibit increasing levels of sensitivity (recall) and specificity for accurate diagnoses [35,36], while AI-based approaches facilitate statistical prediction through classification methods.

Image-based machine learning models are widely employed as CAD tools for classification tasks. These models apply different techniques such as support vector machines (SVM), which establish a hyperplane for binary classification and optimize the margin between the hyperplane and the classes [50–52]. Random forest [53], an ensemble method, employs decision trees on data subsets, averaging their predictions to enhance accuracy and mitigate overfitting; this approach has proven effective for classifying patients with PD [54]. Logistic regression models facilitate binary classification by transforming linear regression values into a range between 0 and 1 using a logistic function. And K-nearest neighbor (K-NN) identifies the K closest data points based on Euclidean distance to form a classified set. In contrast, artificial neural networks (ANN) are more intricate, simulating human brain neural networks and excelling in image processing tasks. Comprising layers that capture pattern complexity, ANN have demonstrated remarkable performance in this domain.

In this article, we aim to identify a suitable machine-learning model to facilitate the diagnostic using [18F]PR04.MZ PET/CT imaging. We specifically examine and compare the performance of SVM, Random Forest, Logistic Regression, K-NN, and Neural Networks. Our assessment focuses on critical parameters such as precision, recall, F1-score, accuracy and AUC to classify between patients with PS from subjects without evidence for dopaminergic deficit (SWEDD).

2. Methods

2.1. Patient population

In this study, 204 subjects (118 men, 86 women, ages 61.6 ± 12.9 years, range 20–90 years), were referred to the Center for Nuclear Medicine & PET/CT PositronMed, Santiago, Chile, for PET/CT imaging using [^{18}F]PR04.MZ to evaluate the integrity of nigrostriatal neuronal. The compiled database was split into two groups: 129 PET scans (63 %) destined to train the models and 75 (37 %) for testing them.

All procedures performed in this study were in accordance with the ethical standards of the institutional and national research committee and with the principles of the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. This is a retrospective analysis of previously acquired PET data approved by the regional ethics committee board (CEC SSM Oriente, permit 20,140,520) and written informed consent has been obtained from all participants.

2.2. PET imaging acquisition and analysis

All images were acquired on a state-of-the-art WB PET/CT scanner (Biograph mCT Flow, Siemens Healthineers, Erlangen, Germany). Patients were allowed to rest for 55 min after injection of 184.0 ± 30.0 MBq (range: 94.7–234.6 MBq) [^{18}F]PR04.MZ before being placed head-first supine in the PET/CT scanner. A low-dose CT was performed for attenuation correction before the emission scan was performed in LIST-mode from 60 to 90 min post-injection (p.i.). PET images were corrected for random, scatter, attenuation, and time-of-flight, reconstructed by an ordered subset expectation maximization (OSEM) algorithm (2 iterations and 21 subsets) followed by post-reconstruction smoothing (Gaussian, 4 mm FWHM), corrected for motion if necessary, and averaged, following the manufacturer's instructions (Siemens). Images consisted of 110 planes of 256 × 256 voxels of 1.59 × 1.59 × 1.5 mm^3.

For routine clinical analysis and reporting, representing the gold standard of our analysis, PET scans were reframed into 6 static frames of 300 s each, reviewed, corrected for motion and averaged for the period of 60–90 min post-injections. PMOD software (Version 3.4, Zurich, Switzerland: PMOD Technologies LLC. Available from: https://www.pmod.com/) was used to co-registered images to CT scans and to normalize to Montreal Neurological Institute (MNI) space. Volumes of interest (VOI) were outlined for the left and right anterior, posterior, and total putamen, caudate nucleus, SNpc, and cerebellum. Specific binding ratios (SBR) were calculated from SUVmean values as follows:

$$SBR = \frac{SUVregion - SUVcerebellum}{SUVcerebellum}$$

Ratios between anterior and posterior putamen were calculated as a measure of a rostro-caudal gradient (RCG):

$$RCG = \frac{SBRanterior\ putamen}{SBRposterior\ \ putamen}$$

VOIs were generated from a three-dimensional maximum probability atlas that included the putamen. The separation between the anterior and posterior putamen was arbitrarily set at 55 % and 45 % respectively. This separation proved to be the ideal ratio to reflect the RCG observed in parkinsonian syndromes [55].

2.3. General workflow

[^{18}F]PR04.MZ PET images were acquired and collected to create a PET image dataset. Later, three independent expert readers assessed and labeled two-hundred-four PET images. The images were preprocessed to generate 2D and 3D datasets for the subsequent training and analysis of five different ML models. Final statistical analysis was performed to evaluate accuracy, precision, recall, F1-score and AUC (Fig. 1).

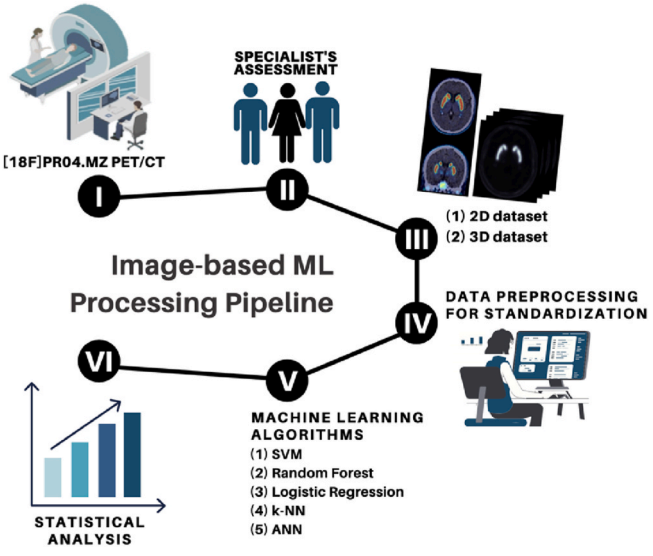

Fig. 1. Image-based ML processing pipeline: I) PET imaging with [^{18}F]PR04.MZ, II) Image classification, III) creation of 2D and 3D datasets, IV) preprocessing of datasets, V) training of ML model, VI) analysis.

2.4. Classification of PET images

All PET images (axial and coronal views) and quantitative signal-to-background-ratio (SBR) values (Fig. 2) were assessed. A pathological deficit was defined as a decrease of more than 40 % (eq. to 2 x SD) from the mean value of the healthy control group in any region.

Each scan was assigned as follows: normal (0) or abnormal (1), showing absence (0) or presence (1) of asymmetric binding, absence (0) or presence (1) of a rostro-caudal gradient, and being compatible (1) or not (0) with PS. The final label was assigned to a scan without evidence for dopaminergic deficit (0) or a scan showing dopamine deficiency compatible with PS, based on the agreement of at least two readers.

2.5. Imaging preprocessing and feature selection: 2D and 3D datasets

The data preprocessing pipeline was designed to produce two distinct dataset dimensions; therefore, each input was processed differently (Fig. 3). For the 2D dataset, the preprocessing stage focused on selecting the region of interest (ROI), targeting the striatal and midbrain regions within each image. This work required standardized images in a normalized space for which we used the three-dimensional maximum probability atlas [55]. Additionally, due to acquisition and anatomical variability, some images exhibited differences in the positioning of the ROI, requiring an automated localization process. This adjustment also affected the standardization of image resolution, leading to the generation of resized pseudo-color images with a resolution of 80 × 40 pixels (Fig. S1). This process can be followed in Fig. 3, left workflow.

The preprocessing of the 3D dataset was centered on the volume of interest (VOI), resulting in 10 consecutive 2D images covering the striatal and midbrain regions (see Fig. 3, right workflow). The unsupervised model applied to achieve this selection is described as follows.

1. From the acquired 3D image, 45 % and 30 % were removed from the upper and lower slices, respectively. This standardized inner volume was used as a first filter to select the VOI (Fig. 4).
2. Density-based spatial clustering of applications with noise (DBSCAN) is a technique of cluster building around an arbitrary initial point. The 'eps' parameter defines the radius within which neighboring points are considered, and if these points meet the required 'min samples'—the minimum number of points necessary—the area forms a cluster. Otherwise, it's classified as noise. The sklearn.cluster. DBSCAN function was used to select the slice with the highest density as the central slice of the output-optimized brain box. Therefore, the cluster shows the tracer uptake in the target region of interest (ROI), i.e. caudate nucleus, anterior and posterior putamen of both left and right brain hemispheres (Fig. 5).
3. Finally, a set of ten slices were selected to encompass the areas of highest density, resulting in a scan output with reduced resolution, as illustrated in Fig. 6.

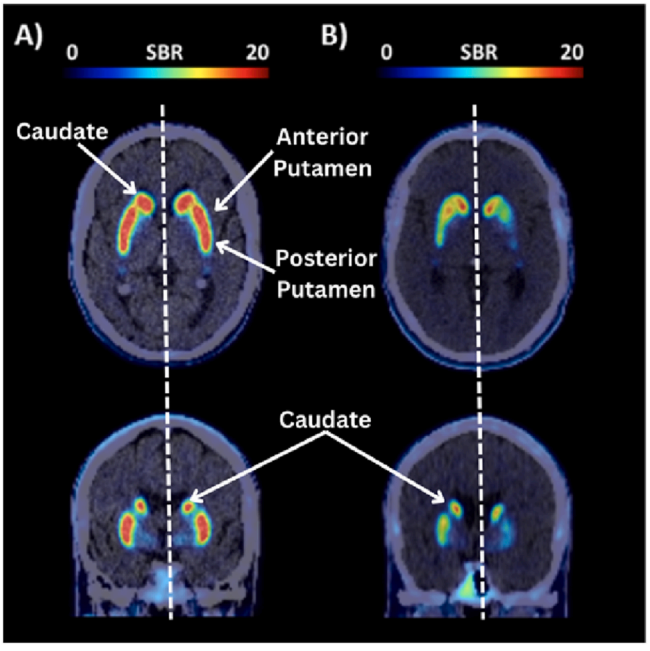

Fig. 2. Axial and coronal section images at the level of the normalized striatum in A) a SWEDD patient and B) a PS patient. Figure also includes anatomical information, i.e. caudate nucleus, anterior and posterior putamen, and dotted line separating left and right brain hemispheres.

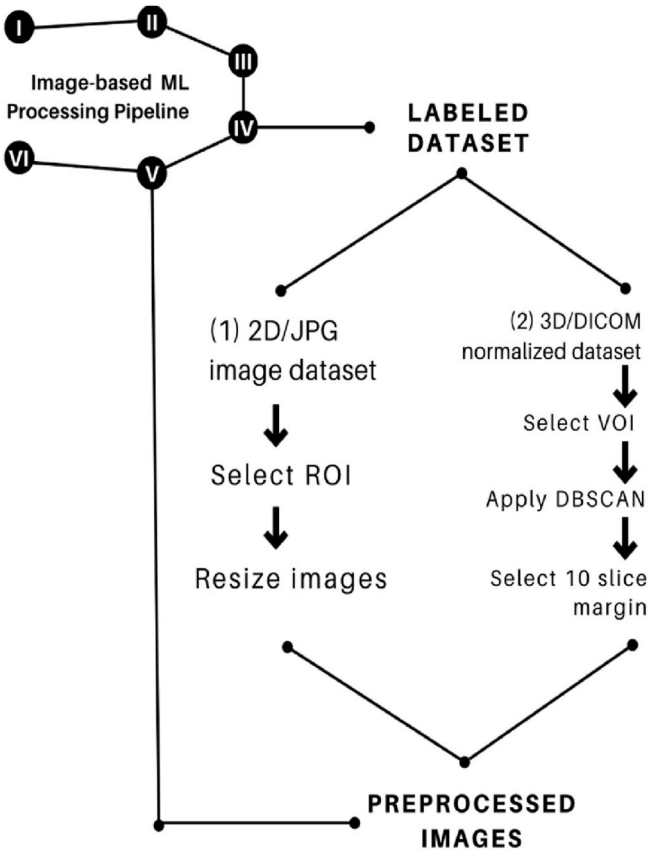

Fig. 3. Data preprocessing for standardization pipeline.

The success of this process allowed the reduction of the input images for the machine learning classification, before splitting them into the training/validation/test set for the supervised learning classification.

2.6. Machine learning model pipeline

After the dataset labeling and image preprocessing, 63 % of the images were exclusively used to train a ML model. To achieve our goal, we compared five different classifiers (SVM, random forest, logistic Regression, K-NN and neural network) and systematically applied a 10-fold cross-validation method (Fig. 7).

The performance of each classifier was then evaluated on the remaining 37 % of images, using labels assigned by expert readers (Step 6).

2.6.1. 2D-dataset

The 2D dataset was divided as described in Table 1. A total of 129 2D images were assigned to the training set, while the remaining 75 were used for testing.

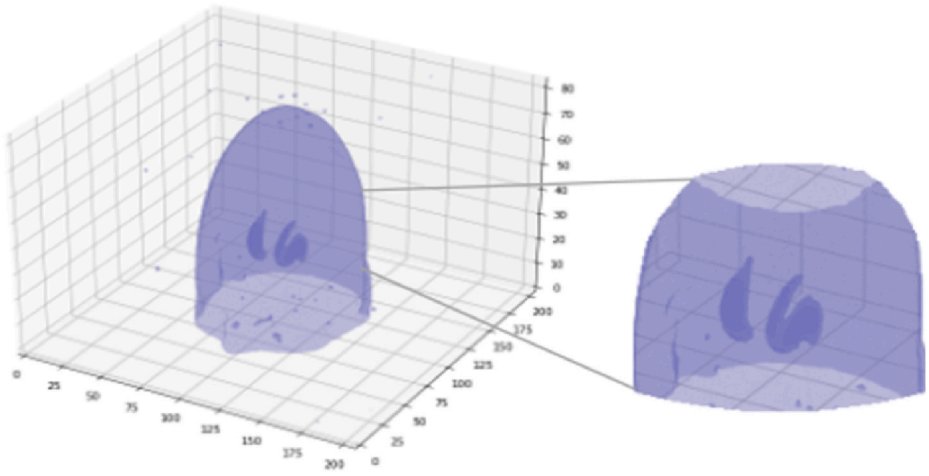

Fig. 4. 3D superior-inferior margin removal.

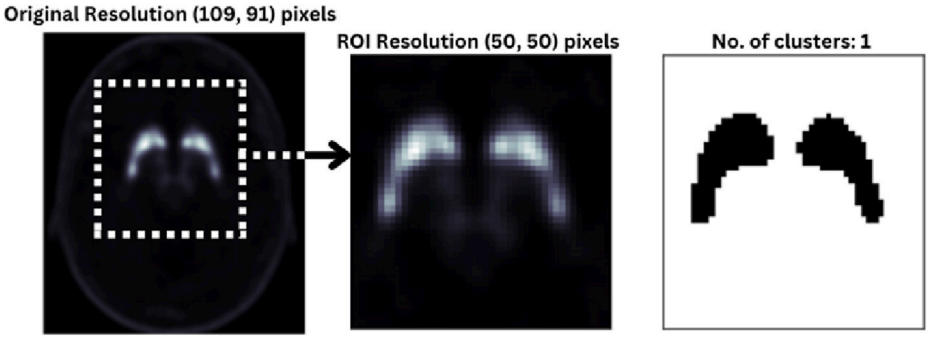

Fig. 5. Tracer uptake in PET images.

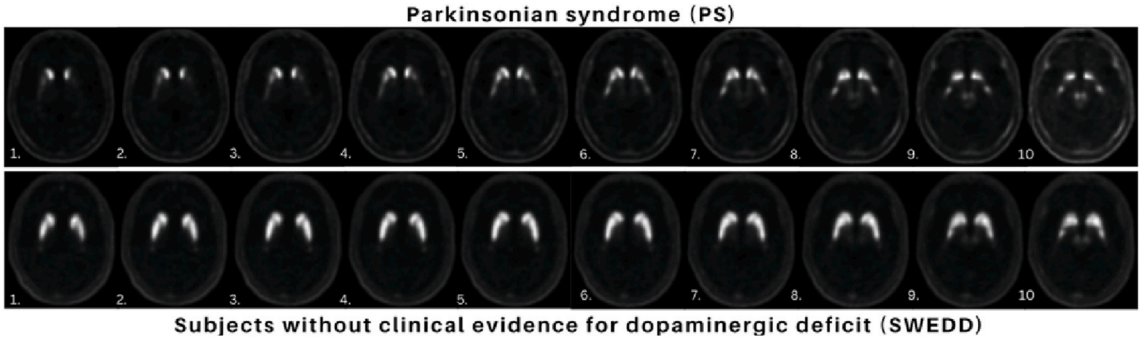

Fig. 6. Ten consecutive slice selection of PS (upper row) patients and SWEDD (lower row) subjects.

During the preprocessing stage (Fig. 3), the ROI was selected and, prior to resizing the scans to a lower resolution, the margins of each pseudo-color image were removed. This step allowed the input homogenization before running the algorithm.

To avoid overfitting, the training process incorporated a cross-validation technique, where the training dataset was split into 10 folds: 90 % used to train the model and 10 % to validate its performance after each iteration (Fig. 7).

For training of the preprocessed images (step 5 in Fig. 1) Python 3.8.2 scikit-learn functions (http://scikit-learn.org) were used to implement the ML models. GridSearchCV was used to find the best parameter grid across all possible combinations (Table 2).

TensorFlow keras (https://www.tensorflow.org/guide/keras/sequential_model) was used to implement the neural network. The neural network was composed of three convolutional layers, each featuring a 3 × 3 kernel layer and a ReLU activation function, followed by a 2 × 2 max pooling layer and one flatten with activation 'relu'. The classification layer included a dense layer with a 'sigmoid' activation function that allows full connectivity between neurons in preceding and succeeding layers. Finally, we compiled the model using a 'binary_crossentropy' loss, 'adam' optimizer and measured its performance with metric 'accuracy'.

Each model was executed and validated iteratively, comparing their performance in terms of precision, recall, F1- score and accuracy (Fig. 7).

2.6.2. 3D-dataset

The results of the preprocessed 3D dataset were submitted as input for the supervised classifier. This 3D array was flattened into a pixel vector. Similar to the approach used for the 2D dataset, all machine

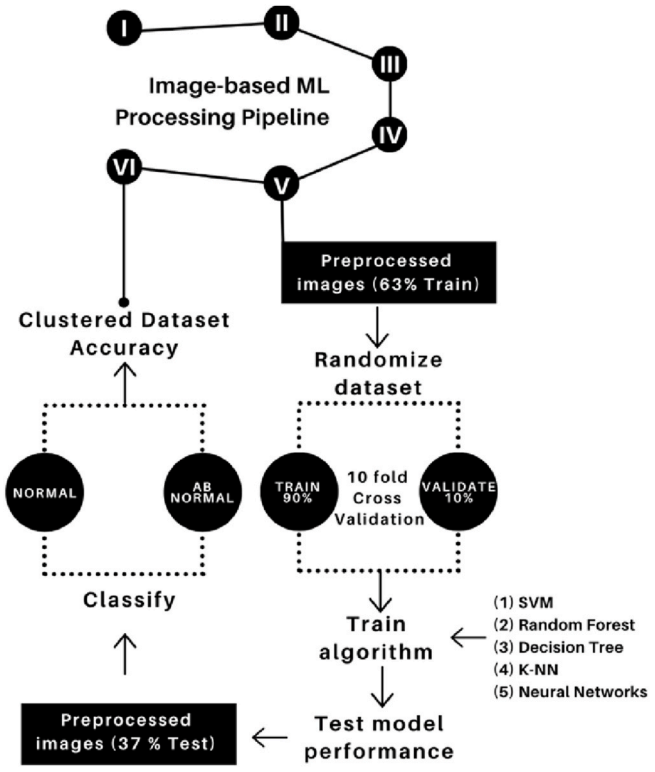

Fig. 7. Machine Learning model pipeline.

Table 1
Group composition and demographic details.

	SWEDD Training	PS Training	SWEDD Testing	PS Testing
Amount	54	75	39	36
Age	63.3 ± 13.9	59.0 ± 13.9	60.9 ± 11.5	65.3 ± 9.5

Table 2
Best parameter selection for each machine learning model using 2D dataset.

Model	GridSearchCV selection
SVM	SVC(C = 1, kernel = 'linear', probability = True, random_state = 21)
RF	RandomForestClassifier(criterion = 'entropy', random_state = 42)
LR	LogisticRegression(C = 5, penalty = 'l1', random_state = 21, solver = 'liblinear')
K-nn	KNeighborsClassifier(n_neighbors = 1)

learning models applied to the 3D-dataset employed the same hyperparameter optimization grid search technique (Table 3). To avoid overfitting and ensure the creation of a generalized model adaptable to unseen data, a 10-fold cross-validation technique was employed. The models were later compared using the previously mentioned indicators: precision, recall, F1- score and accuracy.

The Ax Adaptive Experimentation Platform (https://ax.dev) was used to select the best neural network combination of hyperparameters.

Table 3
Best parameter selection for each machine learning model using 3D dataset.

Model	GridSearchCV selection
SVM	SVC(C = 1, kernel = 'linear', probability = True, random_state = 0)
RF	RandomForestClassifier(n_estimators = 45, random_state = 0)
LR	LogisticRegression(C = 1, penalty = 'l1', random_state = 0, solver = 'liblinear')
K-nn	KNeighborsClassifier(n_neighbors = 2, weights = 'distance')

The metric defined in the network compiler was set to optimize 'binary accuracy'. AX library.get_next_trial(.) allows to iteratively create a neural network with a combination of parameters for 25 experimental trials. Finally, ax client.get_best parameters select the best set of parameters (Table 4).

To address overfitting, Ax platform was set to assign a 'learning_rate' within the range [0.0001, 0.001], and a 'dropout_rate' within [0.01, 0.05]. According to Table 4, 0.013 % is the best learning rate to get the minimum loss function, though it requires more epochs and computational resources, while 1.06 % dropout rate is the percentage of neurons dropped between the three hidden layers. Additionally, 'EarlyStopping' (https://keras.io/api/callbacks/early_stopping) with (patience = 20) was included in the models' callback. This ensures that training halts automatically when no further improvement is detected.

Accuracy and loss value metrics were used to select the best model. Graphing both indicators allowed the comparison of the model performance with the training and validation data after each iteration. Accuracy eases the visualization of how precisely it is to classify the pattern of an image; meanwhile, loss value indicates the error that adds each training epoch. The convergence behavior of both graphs indicates the best model for later classification of the test set. After setting the model to train for 500 epochs and configuring early stopping (patience = 20), the algorithm converged, as shown in Fig. 8, and was ready for final classification.

2.7. Statistical analysis

The inter-observer reliability between the clinicians and the image-based model identification was calculated using Cohen's exact Kappa (κ) [56]. We considered the level of agreement according to the following value ranges for κ: slight (0.00–0.20), fair (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), and almost perfect (0.81–1.00) [57]. Calculations were performed using 95 % confidence intervals (CI's), and $p < 0.05$ were considered significant. The statistical analysis was performed using R "IRR"–CRAN, package 3.5.1 version (http://www.R-project.org).

To analyze the performance of the five different models, accuracy (1), precision (2), recall (3) and F1-score (4) were evaluated using the values for true positive (TP), true negative (TN), false positive (FP), and false negative (FN) and according to the following equations.

	Predicted value		
		1	0
True value	0	$\overline{FP}$	$\overline{TN}$
	1	TP	FN

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$\frac{TP}{TP + FP} \quad (2)$$

$$\frac{TP}{TP + FN} \quad (3)$$

Table 4
Summary of the best set of parameters for neural network binary accuracy optimization.

Hyperparameter	Best combination
'learning rate'	0.00013850114679874224
'dropout rate'	0.010574775359670174
'num hidden layers'	5
'neurons per layer'	17
'batch size'	8
'activation'	'tanh'
'optimizer'	'rms'
'keras cv'	1.0000164359863488

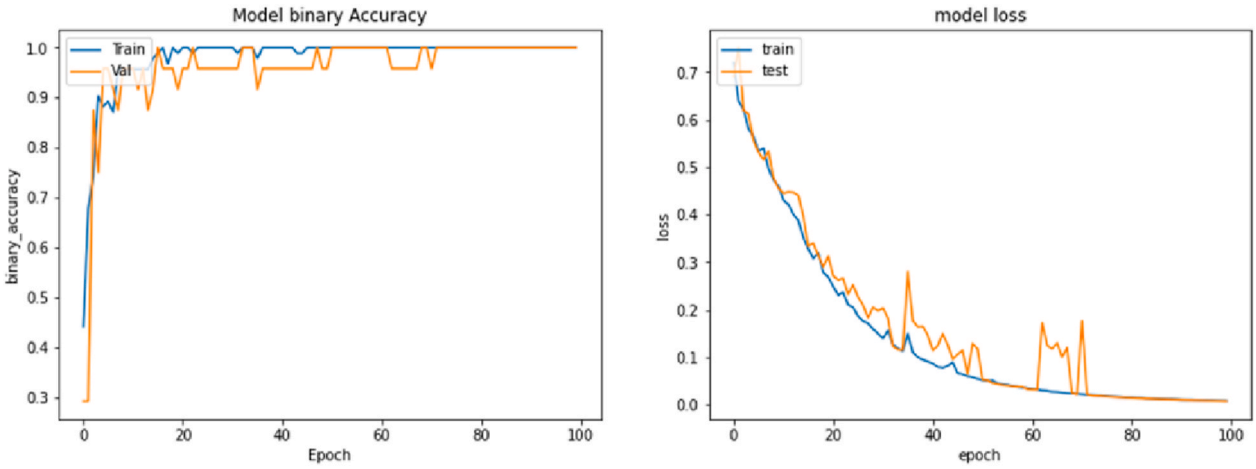

Fig. 8. Convergence for accuracy and loss values.

$$2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

Where accuracy measures the percentage of samples correctly classified by the model. Precision is the quality of the ML model in classification tasks rated according to the number of TP in a total of PS predictions. Recall reports the number of PS subjects the ML model can identify, and F-score measuring the performance of precision and recall between models.

In addition to the metrics mentioned above, Receiver Operating Characteristics (ROC) curves were also displayed. By showing the True Positive Rate (TPR) and False Positive Rate (FPR), ROC curves help to evaluate the performance of each classifier. In an illustration, the best ROC curve is the one closer to the top-left corner of the image, since the ideal scenario implies having high TPR and no FPR.

The Area Under the curve (AUC), in this case the ROC curve, rates on a scale from 0 to 1 the value of the area covered by the curve. This metric, along with other indicators, contribute to the comparison between models' performance.

3. Experimental results

Three independent expert readers blindly analyzed the [^{18}F]PR04. MZ PET/CT images. As an additional measure of control, the study included eighteen healthy volunteers, all of whom were correctly classified by each expert reader. The Kappa statistic showed an almost perfect value ($\kappa = 0.946$; $p < 0.001$, 95 % CI 0.918–0.966), indicating a 97 % agreement between the clinical and neural network diagnosis. Furthermore, only one case was misclassified by our algorithm (Table S1).

3.1. Performance with the 2D dataset

After training, the classifiers were tested with the remaining dataset portion (37 %). Table 5 shows the results obtained from the classifiers.

Table 5
Classification performance average five image-based ML classifiers with testing images from the 2D dataset.

Classifier	Precision	Recall	F1-score	Accuracy	AUC
SVM	0.86	0.84	0.82	82.67 %	0.9614
Random Forest	0.80	0.71	0.67	69.33 %	0.9375
LogReg	0.81	0.77	0.75	76.00 %	0.9621
K-NN	0.69	0.57	0.48	54.67 %	0.6636
NNs	**0.97**	**0.97**	**0.97**	**97.33 %**	**0.9993**

Neural networks scored higher in all indicators compared to the other ML model.

For the 2D dataset, the k-NN presented the lowest performance versus the other models (Table 5).

Fig. 9 presents the ROC curves and summarizes the AUCs. Both metrics contribute to the comparison task between the models. The results showed that NNs achieved an area under the ROC curve of 0.9993, coupled with superior precision, recall, F1-score and accuracy. This performance positions neural networks as the best model for the 2D input dataset.

3.2. Performance with 3D datasets

Table 6 shows the results achieved by each classifier for the 3D dataset according to the predicted probability of class membership (Table S1).

While k-NN underperformance with 2D images, results in the 3D dataset were significantly better. In general, results indicated that models consistently outperformed with 3D inputs compared to 2D datasets (Table 6, Fig. S2).

As in the 2D dataset, NNs performed well in identifying PS-compatible patients, closely matched by RF and LogReg. Notable, there was an important improvement in the models AUCs, specially for random forest and k-nearest neighbors (Fig. 10). Overall, models trained

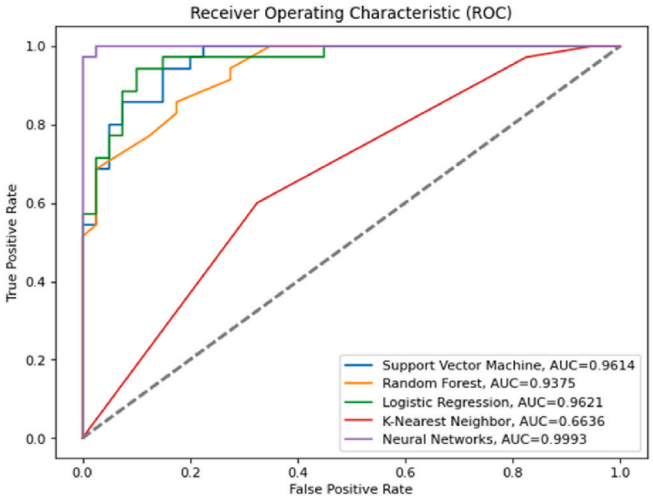

Fig. 9. ROC curve including all five models' performance. The AUC is in the legend for each model.

Table 6
Classification performance average of four ML classifiers and neural networks with testing images from the 3D dataset.

Classifier	Precision	Recall	F1-score	Accuracy	AUC
SVM	0.97	0.98	0.97	97.33 %	0.9871
Random Forest	**0.99**	**0.99**	**0.99**	**98.67 %**	**0.9925**
LogReg	**0.99**	**0.99**	**0.99**	**98.67 %**	**0.9684**
K-NN	0.91	0.91	0.91	90.67 %	0.9867
NNs	**0.99**	**0.99**	**0.99**	**98.67 %**	**0.9727**

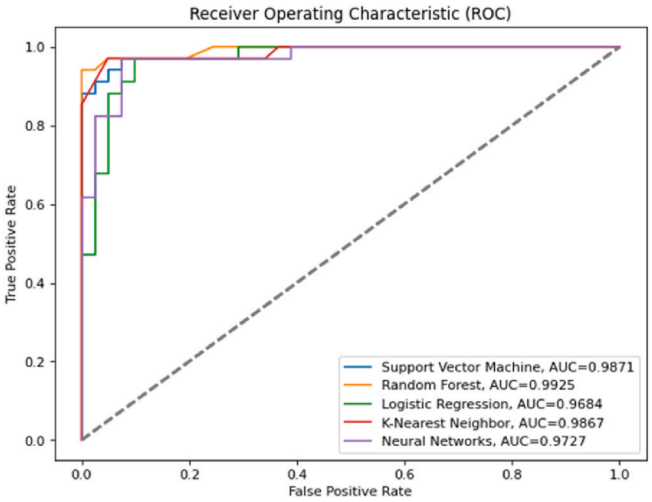

Fig. 10. ROC curve including all five models' performance. The AUC is in the legend for each model.

using pixel vectors from three-dimensional inputs consistently outperformed across all metrics (Fig. 11).

Fig. 11 suggests that 3D imaging offers additional information by capturing depth dimensions. Therefore, relying on just one image may place excessive emphasis on a specific region [58,59].

4. Discussion

PET imaging of DAT is a useful diagnostic method for monitoring dopaminergic neurodegeneration within the striatum and ventral area of the SNpc in movement disorders [1–3]. Previous studies presented PET imaging with [^{18}F]PR04.MZ as an excellent technique with high sensitivity for evaluation of patients with movement disorders [37–42].

In this article, we first selected and developed five different models that aim to classify [18F]PR04.MZ PET 2D and 3D scans. We measured and compared their performances across both 2D and 3D approaches. Although all models discriminate between PS and SWEDDs with high accuracy and precision, similar to the expert clinician's performance, models presented the highest performance in both datasets, with more than 97 % agreeing with the clinical diagnosis. Thus, our work presents an automated tool using machine learning to discriminate SWEDDs and PS patients, providing crucial support to less experienced neurologists in evaluating PET scans images during clinical routines.

Given its potential as a novel radiotracer, [^{18}F]PR04.MZ PET scans are expected to be extensively distributed in clinical centers. Our results demonstrate that machine learning models can effectively discriminate these PET scans with an accuracy exceeding 98 %. These models will thus serve as tools to aid in the analysis of larger datasets of neuroimages using this tracer.

Our results indicate that all models perform better with 3D images as input, suggesting that multidimensional data may accurately reflect the complex structure of the brain. A study comparing multiple approaches with different dimensions for brain MRI auto-segmentation images revealed that 3D approaches achieve higher performance for limited training data compared to 2D and even 2.5D [60]. However, that approach requires more computational memory.

To explore the use of publicly available tools for non-programmers, we compared our 2D dataset models' performance with the trial version of the Google platform algorithm. The 129 labeled 2D images were uploaded and used to train a model on the platform algorithm, while the remaining 75 images were reserved for testing (Table S3). The Google Cloud Vision API demonstrated superior sensitivity and specificity, achieving an accuracy rate of 93.33 % and a recall value of 0.94 (Table S3), surpassing the averages reported by individual clinicians. This suggests that exploring cost-effective alternatives, such as cloud-based business tools, may be worthwhile for employing ML in data classification tasks.

In machine learning, the level of trust in a model often correlates with its interpretability, due to the traceability of the prediction or decision steps. Models with low interpretability but high accuracy, known

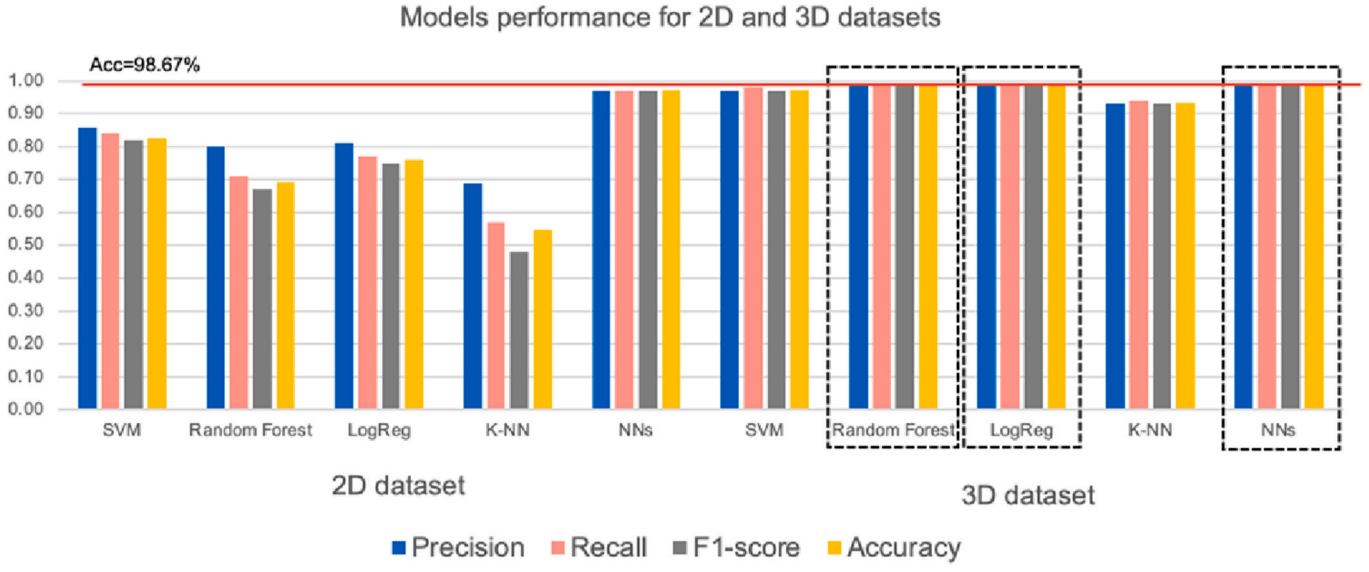

Fig. 11. Performance of the different classifiers based on accuracy, precision, recall and F1-score metrics used in this research: Support Vector Machine, Random Forest, Logistic Regression, k-Nearest Neighbor and Neural Network.

as black-box models, are less explainable. Among models with greater explainability, logistic regression stands out, while NNs, although more accurate, are less explainable (Fig. 12).

Our CAD approach performed well on NN models in both two and three-dimensional datasets. However, interpretability is not included into these models. Some studies address interpretability in NN using techniques such as Local Interpretable Model Agnostic Explanations (LIME) [61]. Despite this, models like logistic regression and random forest achieved high scores in all the evaluated indicators, becoming better alternatives than NN.

4.1. Limitations and future work lines

Our analysis represents an initial effort to evaluate the performance of ML models in discriminating between PS compatible and SWEDDs using 2D or 3D [18F]PR04.MZ PET imaging. The dataset utilized was small due to the restricted number of patients available for the study. As the dataset expands, future research lines could incorporate augmentation techniques [62,63] to further enhance data availability and model robustness.

It is also important to mention that the accuracy and other performance metrics reported in this study, are based on the standard of truth established by three specialists and the image specifications employed. Therefore, all models trained using this information as a reference will perform in accordance with the provided data. Since the CAD system developed in this study aimed to replicate this assessment, it may not be exempt from misdiagnosis. Currently, the definitive diagnosis of Parkinson's disease (PD) can only be confirmed post-mortem through autopsy. Importantly, all subjects in this study who exhibited PD-compatible symptoms are still alive, implying that their diagnoses are based on PET image and clinical assessments.

As [^{18}F]PR04.MZ PET scans become widely used as a DAT tracer, future work lines with larger datasets have the potential to deliver more robust and generalizable results. Additional data could be incorporated into this approach to train models capable of distinguishing between various movement disorders syndromes often referred to as PS. Studies that apply such multiclass classification techniques are opening novel lines of research and applications involving image and computational methodologies [64]. Furthermore, within the context of [^{18}F]PR04.MZ PET/CT images, a dataset that includes a follow-up of symptoms progression and patient deterioration could facilitate the development of models for stage classification.

5. Conclusions

Advances in computational technologies, such as artificial intelligence in medical decision-making, have prompted ethical debates and critical questions about the role of AI in healthcare. One key consideration is the extent to which artificial intelligence will influence medical decisions. Additionally, it is pertinent to evaluate whether the increased accuracy achieved by machine learning tools contributes to giving greater trust in their application.

Our study demonstrated 97 % agreement between the diagnoses made by consensus expert assessment and those made using ML techniques. Although our data set is small, this represents an initial approach to training a machine learning tool with the potential utility as a valuable diagnostic aid and its integration into routine clinical workflow.

While NN are powerful classification tools, our research highlights that other pattern recognition algorithms, i.e. random forest and logistic regression, can be interpretable alternatives for image-based medical classification tasks when using 3D images. Thus, the adoption of these models in standardized clinical reports could substantially reduce the time required for patient analysis and improve diagnostic certainty.

Interested parties may inquire about access to the dataset directly from any of the authors, in accordance with the data sharing policies and procedures established for this study.

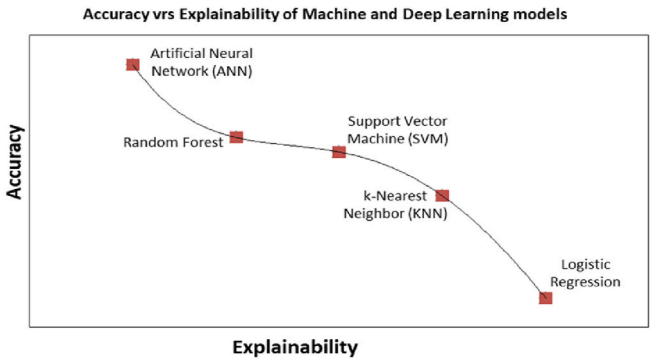

Fig. 12. Schematic representation of accuracy and explainability of the models used in this research: SVM, RF, LogReg, k-NN and neural network.

CRediT authorship contribution statement

Maria Jiménez: Writing – review & editing, Writing – original draft, Methodology, Investigation. **Cristian Soza-Ried:** Writing – review & editing, Writing – original draft, Investigation. **Vasko Kramer:** Writing – review & editing, Investigation. **Sebastian A. Ríos:** Methodology. **Arlette Haeger:** Investigation, Formal analysis. **Carlos Juri:** Investigation, Formal analysis. **Horacio Amaral:** Investigation, Formal analysis. **Pedro Chana-Cuevas:** Investigation, Formal analysis.

Ethical statement

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the principles of the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. This is a retrospective analysis of previously acquired PET data approved by the regional ethics committee board (CEC SSM Oriente, permit 20,140,520) and written informed consent has been obtained from all participants.

Funding

This research has received financial support from ANID, Chile (grant Fondecyt regular 1220908).

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Maria Jimenez, Cristian Soza-Ried, Vasko Kramer, Pedro Chana-Cuevas, Arlette Haeger & Carlos Juri reports financial support was provided by ANID, Chile (grant Fondecyt regular 1220908). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank Phillipe Salles and Paula Saffie Awad (all CETRAM) for their clinical inputs and reviewing the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmed.2025.100232.

References

[1] Berg D, Borghammer P, Fereshtehnejad SM, Heinzel S, Horsager J, Schaeffer E, Postuma RB. Prodromal Parkinson disease subtypes—key to understanding heterogeneity. Nat Rev Neurol 2021;17(6):349–61.

[2] Oxtoby NP, Leyland LA, Aksman LM, Thomas GE, Bunting EL, Wijeratne PA, Young AL, Zarkali A, Tan MM, Bremner FD, Keane PA. Sequence of clinical and neurodegeneration events in Parkinson's disease progression. Brain 2021;144(3):975–88.

[3] Surmeier DJ. Determinants of dopaminergic neuron loss in Parkinson's disease. FEBS J 2018;285(19):3657–68.

[4] Leiva AM, Martínez-Sanguinetti MA, Troncoso-Pantoja C, Nazar G, Petermann-Rocha F, Celis-Morales C. Chile lidera el ranking latinoamericano de prevalencia de enfermedad de Parkinson. Rev Med Chile 2019;147(4):535–6.

[5] León JB. Epidemiología de la enfermedad de Parkinson en España y su contextualización mundial. Rev Neurol 2018;66(4):125–34.

[6] Dorsey ER, Elbaz A, Nichols E, Abd-Allah F, Abdelalim A, Adsuar JC, Ansha MG, Brayne C, Choi JY, Collado-Mateo D, Dahodwala N. GBD 2016 Parkinson's Disease Collaborators. Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol 2018;17(11):939–53.

[7] Marsili L, Rizzo G, Colosimo C. Diagnostic criteria for Parkinson's disease: from James Parkinson to the concept of prodromal disease. Front Neurol 2018;9:156.

[8] Postuma RB, Berg D, Stern M, Poewe W, Olanow CW, Oertel W, Obeso J, Marek K, Litvan I, Lang AE, Halliday G. MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 2015;30(12):1591–601.

[9] Postuma RB, Poewe W, Litvan I, Lewis S, Lang AE, Halliday G, Goetz CG, Chan P, Slow E, Seppi K, Schaffer E. Validation of the MDS clinical diagnostic criteria for Parkinson's disease. Mov Disord 2018;33(10):1601–8.

[10] Heinzel S, Berg D, Gasser T, Chen H, Yao C, Postuma RB. MDS Task Force on the Definition of Parkinson's Disease. Update of the MDS research criteria for prodromal Parkinson's disease. Mov Disord 2019;34(10):1464–70.

[11] Yue X, Li H, Yan H, Zhang P, Chang L, Li T. Risk of Parkinson disease in diabetes mellitus: an updated meta-analysis of population-based cohort studies. Medicine 2016;95(18).

[12] Yang YW, Hsieh TF, Li CI, Liu CS, Lin WY, Chiang JH, Li TC, Lin CC. Increased risk of Parkinson disease with diabetes mellitus in a population-based study. Medicine 2017;96(3).

[13] De Pablo-Fernandez E, Goldacre R, Pakpoor J, Noyce AJ, Warner TT. Association between diabetes and subsequent Parkinson disease: a record-linkage cohort study. Neurology 2018;91(2):e139–42.

[14] Darweesh SK, Wolters FJ, Postuma RB, Stricker BH, Hofman A, Koudstaal PJ, Ikram MK, Ikram MA. Association between poor cognitive functioning and risk of incident parkinsonism: the Rotterdam study. JAMA Neurol 2017;74(12):1431–8.

[15] Schrag A, Anastasiou Z, Ambler G, Noyce A, Walters K. Predicting diagnosis of Parkinson's disease: a risk algorithm based on primary care presentations. Mov Disord 2019;34(4):480–6.

[16] Weintraub D, Chahine LM, Hawkins KA, Siderowf A, Eberly S, Oakes D, Seibyl J, Stern MB, Marek K, Jennings D, Investigators PARS. Cognition and the course of prodromal Parkinson's disease. Mov Disord 2017;32(11):1640–5.

[17] Weisskopf MG, O'reilly E, Chen H, Schwarzschild MA, Ascherio A. Plasma urate and risk of Parkinson's disease. Am J Epidemiol 2007;166(5):561–7.

[18] Gao X, O'Reilly ÉJ, Schwarzschild MA, Ascherio A. Prospective study of plasma urate and risk of Parkinson disease in men and women. Neurology 2016;86(6):520–6.

[19] Chen H, Mosley TH, Alonso A, Huang X. Plasma urate and Parkinson's disease in the atherosclerosis risk in communities (ARIC) study. Am J Epidemiol 2009;169(9):1064–9.

[20] Kobylecki CJ, Nordestgaard BG, Afzal S. Plasma urate and risk of Parkinson's disease: a Mendelian randomization study. Ann Neurol 2018;84(2):178–90.

[21] Fang X, Han D, Cheng Q, Zhang P, Zhao C, Min J, Wang F. Association of levels of physical activity with risk of Parkinson disease: a systematic review and meta-analysis. JAMA Netw Open 2018;1(5):e182421.

[22] Yang F, Trolle Lagerros Y, Bellocco R, Adami HO, Fang F, Pedersen NL, Wirdefeldt K. Physical activity and risk of Parkinson's disease in the Swedish national march cohort. Brain 2015;138(2):269–75.

[23] Logroscino G, Sesso HD, Paffenbarger RS, Lee IM. Physical activity and risk of Parkinson's disease: a prospective cohort study. J Neurol Neurosurg Psychiatr 2006;77(12):1318–22.

[24] Sasco AJ, Paffenbarger RS, Gendre I, Wing AL. The role of physical exercise in the occurrence of Parkinson's disease. Arch Neurol 1992;49(4):360–5.

[25] Xu Q, Park Y, Huang X, Hollenbeck A, Blair A, Schatzkin A, Chen H. Physical activities and future risk of Parkinson disease. Neurology 2010;75(4):341–8.

[26] Chen H, Zhang SM, Schwarzschild MA, Hernan MA, Ascherio A. Physical activity and the risk of Parkinson disease. Neurology 2005;64(4):664–9.

[27] Sääksjärvi K, Knekt P, Männistö S, Lyytinen J, Jääskeläinen T, Kanerva N, Heliövaara M. Reduced risk of Parkinson's disease associated with lower body mass index and heavy leisure-time physical activity. Eur J Epidemiol 2014;29:285–92.

[28] Thacker EL, Chen H, Patel AV, McCullough ML, Calle EE, Thun MJ, Schwarzschild MA, Ascherio A. Recreational physical activity and risk of Parkinson's disease. Mov Disord 2008;23(1):69–74.

[29] Barber TR, Klein JC, Mackay CE, Hu MT. Neuroimaging in pre-motor Parkinson's disease. Neuroimage: Clinica 2017;15:215–27.

[30] King AE, Mintz J, Royall DR. Meta-analysis of 123I-MIBG cardiac scintigraphy for the diagnosis of Lewy body–related disorders. Mov Disord 2011;26(7):1218–24.

[31] Wen MC, Heng HS, Hsu JL, Xu Z, Liew GM, Au WL, Chan LL, Tan LC, Tan EK. Structural connectome alterations in prodromal and de novo Parkinson's disease patients. Park Relat Disord 2017;45:21–7.

[32] Dayan E, Browner N. Alterations in striato-thalamo-pallidal intrinsic functional connectivity as a prodrome of Parkinson's disease. Neuroimage Clin 2017;16:313–8.

[33] Knudsen K, Fedorova TD, Hansen AK, Sommerauer M, Otto M, Svendsen KB, Nahimi A, Stokholm MG, Pavese N, Beier CP, Brooks DJ. In-vivo staging of pathology in REM sleep behaviour disorder: a multimodality imaging case-control study. Lancet Neurol 2018;17(7):618–28.

[34] Saeed U, Compagnone J, Aviv RI, Strafella AP, Black SE, Lang AE, Masellis M. Imaging biomarkers in Parkinson's disease and Parkinsonian syndromes: current and emerging concepts. Transl Neurodegener 2017;6(1):1–25.

[35] Ravina B, Eidelberg D, Ahlskog JE, Albin RL, Brooks DJ, Carbon M, Dhawan V, Feigin A, Fahn S, Guttman M, Gwinn-Hardy K. The role of radiotracer imaging in Parkinson disease. Neurology 2005;64(2):208–15.

[36] Saeed U, Lang AE, Masellis M. Neuroimaging advances in Parkinson's disease and atypical Parkinsonian syndromes. Front Neurol 2020;11:572976.

[37] Riss PJ, Debus F, Hummerich R, Schmidt U, Schloss P, Lueddens H, Roesch F. Ex vivo and in vivo evaluation of [18F] PR04. MZ in rodents: a selective dopamine transporter imaging agent. ChemMedChem: Chemistry Enabling Drug Discovery 2009;4(9):1480–7.

[38] Riss PJ, Roesch F. Efficient microwave-assisted direct radiosynthesis of [18F] PR04. MZ and [18F] LBT999: selective dopamine transporter ligands for quantitative molecular imaging by means of PET. Bioorg Med Chem 2009;17(22):7630–4.

[39] Juri C, Chana P, Kramer V, Pruzzo R, Amaral H, Riss PJ, Rösch F. Imaging nigrostriatal dopaminergic deficit in holmes tremor with 18F-PR04. MZ-PET/CT. Clin Nucl Med 2015;40(9):740–1.

[40] Kramer V, Juri C, Riss PJ, Pruzzo R, Soza-Ried C, Flores J, Hurtado A, Rösch F, Chana-Cuevas P, Amaral H. Pharmacokinetic evaluation of [18 F] PR04. MZ for PET/CT imaging and quantification of dopamine transporters in the human brain. Eur J Nucl Med Mol Imag 2020;47:1927–37.

[41] Juri C, Kramer V, Riss PJ, Soza-Ried C, Haeger A, Pruzzo R, Rösch F, Amaral H, Chana-Cuevas P. [18F] PR04. MZ PET/CT imaging for evaluation of nigrostriatal neuron integrity in patients with Parkinson disease. Clin Nucl Med 2021;46(2):119.

[42] Lehnert W, Riss PJ, Hurtado de Mendoza A, Lopez S, Fernandez G, Ilheu M, Amaral H, Kramer V. Whole-body biodistribution and radiation dosimetry of [18 F] PR04. MZ: a new PET radiotracer for clinical management of patients with movement disorders. EJNMMI Res 2022;12:1–9.

[43] Suzuki K. Overview of deep learning in medical imaging. Radiol Phys Technol 2017;10(3):257–73.

[44] Prashanth R, Roy SD. Early detection of Parkinson's disease through patient questionnaire and predictive modelling. Int J Med Inf 2018;119:75–87.

[45] Latif J, Xiao C, Imran A, Tu S. Medical imaging using machine learning and deep learning algorithms: a review. In: 2019 2nd International conference on computing, mathematics and engineering technologies (iCoMET). IEEE; 2019, January. p. 1–5.

[46] Khachnaoui H, Mabrouk R, Khlifa N. Machine learning and deep learning for clinical data and PET/SPECT imaging in Parkinson's disease: a review. IET Image Process 2020;14(16):4013–26.

[47] Mei J, Desrosiers C, Frasnelli J. Machine learning for the diagnosis of Parkinson's disease: a review of literature. Front Aging Neurosci 2021;13:633752.

[48] Tăuţan AM, Ionescu B, Santarnecchi E. Artificial intelligence in neurodegenerative diseases: a review of available tools with a focus on machine learning techniques. Artif Intell Med 2021;117:102081.

[49] Muschelli J. Recommendations for processing head CT data. Front Neuroinf 2019;13:61.

[50] Vapnik V. The nature of statistical learning theory. Springer science & business media; 1999.

[51] Illán I, Górriz J, Ramírez J, Segovia F, Jiménez-Hoyuela J, Ortega Lozano SJ. Automatic assistance to Parkinson's disease diagnosis in DaTSCAN SPECT imaging. Med Phys 2012;39(10):5971–80.

[52] Oliveira FP, Faria DB, Costa DC, Castelo-Branco M, Tavares JMR. Extraction, selection and comparison of features for an effective automated computer-aided diagnosis of Parkinson's disease based on [123 I] FP-CIT SPECT images. Eur J Nucl Med Mol Imag 2018;45:1052–62.

[53] Ho TK. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition, vol. 1. IEEE; 1995, August. p. 278–82.

[54] Patra AK, Ray R, Abdullah AA, Dash SR. Prediction of Parkinson's disease using Ensemble Machine Learning classification from acoustic analysis. Journal of physics: conference series, vol. 1372. IOP Publishing; 2019, November, 012041. 1.

[55] Hammers A, Allom R, Koepp MJ, et al. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. Hum Brain Mapp 2003;19(4):224–47.

[56] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012;22(3):276–82. 2012.

[57] Parnetti L, Gaetani L, Eusebi P, Paciotti S, Hansson O, El-Agnaf O, Mollenhauer B, Blennow K, Calabresi P. CSF and blood biomarkers for Parkinson's disease. Lancet Neurol 2019;18(6):573–86.

[58] Tufail AB, Anwar N, Othman MTB, Ullah I, Khan RA, Ma YK, Adhikari D, Rehman AU, Shafiq M, Hamam H. Early-stage Alzheimer's disease categorization using PET neuroimaging modality and convolutional neural networks in the 2D and 3D domains. Sensors 2022;22(12):4609.

[59] Wen J, Thibeau-Sutre E, Diaz-Melo M, Samper-González J, Routier A, Bottani S, Dormont D, Durrleman S, Burgos N, Colliot O, Alzheimer's Disease Neuroimaging

Initiative. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med Image Anal 2020;63:101694.
[60] Avesta A, Hossain S, Lin M, Aboian M, Krumholz HM, Aneja S. Comparing 3D, 2.5 D, and 2D approaches to brain image auto-segmentation. Bioengineering 2023;10 (2):181.
[61] Magesh PR, Myloth RD, Tom RJ. An explainable machine learning model for early detection of Parkinson's disease using LIME on DaTSCAN imagery. Comput Biol Med 2020;126:104041.
[62] Alam MS, Wang D, Sowmya A. Image data augmentation for improving performance of deep learning-based model in pathological lung segmentation. In: 2021 digital image computing: techniques and applications (DICTA). IEEE; 2021, November. p. 1–5.
[63] El Jiani L, El Filali S. Overcome medical image data scarcity by data augmentation techniques: a review. In: 2022 international conference on microelectronics (ICM). IEEE; 2022, December. p. 21–4.
[64] Martins R, Oliveira F, Moreira F, Moreira AP, Abrunhosa A, Januário C, Castelo-Branco M. Automatic classification of idiopathic Parkinson's disease and atypical Parkinsonian syndromes combining [11C] raclopride PET uptake and MRI grey matter morphometry. J Neural Eng 2021;18(4):046037.