Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

PAPER

PCA and logistic regression in 2-[^{18}F]FDG PET neuroimaging as an interpretable and diagnostic tool for Alzheimer's disease

To cite this article: Carlos Eduardo Gonçalves de Oliveira *et al* 2024 *Phys. Med. Biol.* **69** 025003

IOP Publishing Phys. Med. Biol. 69 (2024) 025003 https://doi.org/10.1088/1361-6560/ad0ddd

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
3 May 2023

REVISED
30 October 2023

ACCEPTED FOR PUBLICATION
17 November 2023

PUBLISHED
4 January 2024

PAPER

PCA and logistic regression in 2-[^{18}F]FDG PET neuroimaging as an interpretable and diagnostic tool for Alzheimer's disease

Carlos Eduardo Gonçalves de Oliveira[1], **Whemberton Martins de Araújo**[2], **Ana Beatriz Marinho de Jesus Teixeira**[2], **Gustavo Lopes Gonçalves**[1], **Emerson Nobuyuki Itikawa**[1,*] **For the Alzheimer's Disease Neuroimaging Initiative**[3]

[1] Institute of Physics, Federal University of Goiás, Goiânia, Goiás, Brazil
[2] Centro de Diagnóstico por Imagem, Goiânia, Goiás, Brazil
[3] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found here.
* Author to whom any correspondence should be addressed.

E-mail: carlosedgonc@gmail.com, whemberton@gmail.com, teixeiramjanabeatriz@gmail.com, gustavolopes41.gl@gmail.com and emersonitikawa@ufg.br

Keywords: principal component analysis, logistic regression, FDG PET, neurodegenerative disease, artificial intelligence

Supplementary material for this article is available online

Abstract

Objective. to develop an optimization and training pipeline for a classification model based on principal component analysis and logistic regression using neuroimages from PET with 2-[^{18}F]fluoro-2-deoxy-D-glucose (FDG PET) for the diagnosis of Alzheimer's disease (AD). *Approach.* as training data, 200 FDG PET neuroimages were used, 100 from the group of patients with AD and 100 from the group of cognitively normal subjects (CN), downloaded from the repository of the Alzheimer's Disease Neuroimaging Initiative (ADNI). Regularization methods L1 and L2 were tested and their respective strength varied by the hyperparameter C. Once the best combination of hyperparameters was determined, it was used to train the final classification model, which was then applied to test data, consisting of 192 FDG PET neuroimages, 100 from subjects with no evidence of AD (nAD) and 92 from the AD group, obtained at the Centro de Diagnóstico por Imagem (CDI). *Main results.* the best combination of hyperparameters was L1 regularization and $C \approx 0.316$. The final results on test data were accuracy = 88.54%, recall = 90.22%, precision = 86.46% and AUC = 94.75%, indicating that there was a good generalization to neuroimages outside the training set. Adjusting each principal component by its respective weight, an interpretable image was obtained that represents the regions of greater or lesser probability for AD given high voxel intensities. The resulting image matches what is expected by the pathophysiology of AD. *Significance.* our classification model was trained on publicly available and robust data and tested, with good results, on clinical routine data. Our study shows that it serves as a powerful and interpretable tool capable of assisting in the diagnosis of AD in the possession of FDG PET neuroimages. The relationship between classification model output scores and AD progression can and should be explored in future studies.

1. Introduction

Neurodegenerative dementia currently affects about 47 million people, a number that is expected to increase to 131 million by the year 2050 (Arvanitakis *et al* 2019). Among the causes of dementia, the main one is Alzheimer's disease (AD), which corresponds between 60% and 80% of the total cases (Brown *et al* 2014, Marcus *et al* 2014). Among the symptoms of AD, what stands out the most is the gradual and increasing loss of memory, which generates a difficulty in learning new information and makes the person affected by the disease repeat questions and conversations frequently and not functionally independent (Arvanitakis *et al* 2019).

AD is pathophysiologically characterized by early neuronal loss and gliosis in the mesiotemporal cortex, with subsequent spread to other brain regions (Brown *et al* 2014). The classic pattern of hypometabolism in the brain involves the posterior cingulate gyrus, precuneus, posterior temporal and parietal lobes, and may include the prefrontal cortex in advanced cases of the disease (Brown *et al* 2014, Marcus *et al* 2014). Histopathological analysis is the reference standard for the diagnosis of AD due to the presence of deposits of abnormally phosphorylated τ proteins and extracellular β-amyloid in the brain (Brown *et al* 2014). As brain biopsies are not easy to perform, positron emission tomography (PET) with 2-[^{18}F]fluoro-2-deoxy-D-glucose (FDG PET) has been shown to be an extremely useful imaging modality for diagnosing AD (Brown *et al* 2014, Marcus *et al* 2014).

Given the intricate nature of AD diagnosis, which necessitates clinical correlations, diverse neuroimaging modalities, and cognitive tests, machine learning (Shinde and Shah 2018) (ML) models have been used to expedite the process. These models are particularly beneficial for early diagnosis, enabling timely intervention and appropriate therapeutic measures. Among the studies involving ML and FDG PET for AD diagnosis, those employing Deep Learning (Hao *et al* 2016) (DL) techniques are particularly noteworthy. For instance, Ding, Y *et al* (Ding *et al* 2019) utilized the Inception V3 convolutional neural network architecture. This study is distinguished by its use of follow-up data from ADNI patients, training the model solely on the final clinical diagnosis, and classifying patients into AD, mild cognitive impairment (MCI), or no evidence of dementia. The study concluded that the DL model outperformed radiology readers in identifying patients who would eventually receive an AD diagnosis. Other studies have expanded on this approach by incorporating additional classes such as Lewy body disease (Etminani *et al* 2022) and subdividing MCI into Early MCI and Late MCI (Singh *et al* 2017), yielding promising results.

Deep learning models, while advantageous in their ability to process image data directly (Lai 2019), present several challenges that limit their reproducibility in clinical settings. These include the necessity for large datasets for model training, extensive training durations, high computational power requirements, and the complexity of model interpretation (Zohuri and Moghaddam 2020). Such constraints hinder the reproducibility of these methods in clinical environments where physicians seek intuitive tools for second opinions on AD diagnoses, without the need for advanced computing resources.

With regard to classic machine learning methods, the disadvantage is that, in order to implement the classification model, a robust feature extraction method is necessary. Some successful examples in the literature include the use of several features typical of radiomics (Nancy Noella and Priyadarshini 2023) (such as contrast, entropy and intensity gradient, for example), the use of VOIs (Dukart *et al* 2013, Lu *et al* 2017) (spatial features) in brain regions known to characterize Alzheimer's disease or covariance patterns extracted by principal component analysis (PCA) (Habeck *et al* 2008, Habeck 2010). Classification models, in turn, tend to involve random forest (RF) (Lu *et al* 2017), logistic regression (LR) (Habeck *et al* 2008, Habeck 2010) and support vector machine (SVM) (Lu *et al* 2017, Nancy Noella and Priyadarshini 2023).

In particular, the combined application of PCA (Abdi and Williams 2010) and LR (Liu *et al* 2009) is promising. PCA, a dimensionality reduction technique (Santo 2012), can be applied to a collection of FDG PET neuroimages to extract covariance patterns or principal components (Habeck 2010, Spetsieris *et al* 2013, Blazhenets *et al* 2019). In turn, LR, a classification model, can be employed on the subject scores (i.e. pattern expression values for each subject) of the principal components. This allows the derivation of linear coefficients that can combine multiple principal components into a singular disease-related spatial covariance pattern (Spetsieris *et al* 2013, Blazhenets *et al* 2019), which corresponds to an interpretable biomarker. The great advantage of using PCA and LR, therefore, lies in the interpretability of the final model. Instead of using complex feature extraction methods and non-linear classification models, PCA can return enlightening and interpretable covariance patterns while the logistic regression model builds a simple and visual decision rule that is easily generalizable.

Two instances of the application of these two techniques for AD prediction using FDG PET neuroimaging can be highlighted. The first study (Habeck *et al* 2008) used only private data, while the second (Habeck and Stern 2010) relied solely on publicly available data obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI). Both studies successfully distinguished healthy subjects from those with AD, but with a limited dataset ($N = 177$ and $N = 80$, respectively) and experimental framework. In addition, these studies did not optimize the classification model. The selection of principal components for training the LR model was limited and could potentially be enhanced through the application of regularization methods (Salehi *et al* 2019).

In an effort to demonstrate the robustness, applicability and predictive capacity of PCA and LR within a realistic experimental framework, the training and optimization of the model were conducted exclusively on publicly available data (ADNI), while testing was carried out on data derived from clinical practice, obtained at Centro de Diagnóstico por Imagem (CDI), reflecting a realistic and practical context. The final outcomes were compared with results derived from DL methodologies in existing literature. For a more direct comparison, three additional models were trained and optimized, namely SVM (Noble 2006), Multi-layer Perceptron

(Popescu *et al* 2009) (MLP) and RF (Breiman 2001). Furthermore, we show how the parameters of the classification model can be intuitively and visually interpreted.

2. Background

2.1. Principal component analysis (PCA)

PCA is a multivariate technique that seeks to explain a set of correlated variables in terms of a reduced number of uncorrelated variables, with greater variance (Abdi and Williams 2010, Santo 2012). Thus, this technique aims to compress and simplify the structure of a dataset. In order to fulfill them, PCA estimates new variables called principal components, which are obtained as linear combinations of the original variables of the dataset (Abdi and Williams 2010, Santo 2012). The first principal component must have the largest possible variance. The second component is computed with the restriction that it must be orthogonal to the first one and have the largest possible variance. The other principal components, in turn, are computed similarly. The corresponding values of each observation for these new variables are called scores, and they can be interpreted geometrically as being the projections of the observations on the principal components (Abdi and Williams 2010, Santo 2012).

In the context of FDG PET neuroimages, it is possible to generate a matrix $M(i,\ j)$ corresponding to all neuroimages of a dataset, so that i represents the subject index and j represents the voxel index. In this way, each neuroimage corresponds to a row vector of the matrix M, with all voxels properly organized horizontally. Thus, after extracting the principal components, we have:

$$M(i,\ j) = SS_1(i)PC_1(\ j) + SS_2(i)PC_2(\ j) + SS_3(i)PC_3(\ j) + ...$$

so that the entire matrix is decomposed in terms of principal components ($PC_x(\ j)$), which explains different and decreasing percentages of the total data variance, and subject scores ($SS_x(i)$), which indicate the projection of the neuroimage of the subject i on the corresponding principal component. Therefore, this is the advantage of using PCA: instead of dealing with the data in terms of voxels, it deals with principal components, where one can take only the ones that add up to the greater part of the total explained variance.

2.2. Logistic regression (LR) and regularization methods

Let $SS_1(i)$, $SS_2(i)$, ..., $SS_p(i)$ be the set of observed subject scores without error, with a total of n observations and p principal components. Thus, the data can be summarized by the matrix $X = (SS_1, SS_2,...,SS_p)$. Furthermore, $Y = (y_1,\ y_2\ ,...,\ y_n)^T$ is considered a random sample of the binary response variable associated with the observations in X, that is, $y_i \in [0,\ 1]$, $i = 1,...,n$ (with 0 indicating the negative group and 1 the positive group). Thus, the LR model is given by (Liu *et al* (2009, Salehi *et al* 2019)

$$y_i = \pi_i + \epsilon_i,\ i = 1\ ,...,\ n,$$

where π_i is the probability of the positive class given ($SS_1(i)$, $SS_2(i)$,...,$SS_p(i)$), calculated as (Liu *et al* 2009, Salehi *et al* 2019)

$$\pi_i = P[(SS_1(i),\ SS_2(i),...,\ SS_p(i))] = \frac{\exp(\beta_0 + \sum_{v=1}^{p} SS_v(i)\beta_v)}{1 + \exp(\beta_0 + \sum_{v=1}^{p} SS_v(i)\beta_v)},$$

where β_v is the weight associated with PC_v. The parameters β_v are determined from the minimization of the cost function given by the expression (Liu *et al* 2009, Salehi *et al* 2019)

$$-\sum_{i=1}^{n}\left[\log(1\ -\ \pi_i) + y_i \log\left(\frac{\pi_i}{1\ -\ \pi_i}\right)\right].$$

In order to increase the generality of the classification model, it is possible to limit its flexibility by penalizing it for high parameter values (Friedman *et al* 2010, Salehi *et al* 2019). One of these methods is the L2 regularization. It is applied by adding a penalty term $\lambda\sum_{v=1}^{p}\beta_v^2$ to the cost function (Friedman *et al* 2010, Salehi *et al* 2019). The λ controls the emphasis that is given to the penalty term: the larger λ, the more the coefficients tend to 0 (in *Python*, the regularization strength is controlled by the inverse of λ, C=1/λ). Although the coefficients tend to 0, few coefficients actually reach zero value (Friedman *et al* 2010, Salehi *et al* 2019). A more aggressive method is the L1-type regularization. The term that is added to the cost function is similar to the method described earlier (Friedman *et al* 2010, Salehi *et al* 2019): $\lambda\sum_{v=1}^{p}|\beta_v|$. With this penalty term, the less important predictor principal components are forced to have a null coefficient (Friedman *et al* 2010, Salehi *et al* 2019).

Table 1. The demographic information for the ADNI and CDI datasets is presented, including the total number of subjects and the distribution of ages, segregated by sex and group. Additionally, the results from the Shapiro-Wilk normality test, represented by p-values, are also provided (non-normal distributions were duly highlighted, p-value <0.05).

			Age distribution							
Dataset	Group	Sex	Min.[a]	Mean	Max.[b]	S.D.[c]	*p*-value[d]	Median	IQR[e]	N[f]
ADNI	AD[g]	F[j]	56	71.3	85	5.6	<0.001[l]	73	3.5	47
		M[k]	57	71.5	85	5.8	0.155	72	5	53
	CN[h]	F	56	70.7	93	5.0	<0.001[l]	71	5	57
		M	62	72.2	75	2.5	<0.001[l]	73	2	43
CDI	AD	F	53	73.7	89	8.0	0.658	74	11	60
		M	60	73.6	85	7.8	0.063	72	11.5	32
	nAD[i]	F	50	63.6	89	10.0	<0.001[l]	60	15	70
		M	50	66.1	84	9.4	0.602	65.5	13	30

Notes.
[a] Minimum age.
[b] Maximum age.
[c] Standard deviation.
[d] Shapiro-Wilk test.
[e] Interquantile range.
[f] Number of subjects.
[g] Alzheimer's disease group.
[h] Cognitively normal group.
[i] Without evidence of any neurodegenerative disease group.
[j] Female.
[k] Male.
[l] Non-normal distribution, p-value < 0.05.

3. Methods

3.1. Data collection

3.1.1. FDG PET neuroimages—ADNI

The publicly available data were downloaded from the ADNI repository. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org.

The search filters used referred to the imaging modality, set to FDG PET, and to the group of subjects, set to *Cognitively Normal* (CN group), corresponding to the healthy subjects, without evidence of cognitive problems of any kind or neurodegenerative diseases, and *Alzheimer*'s *disease* (AD group), corresponding to subjects with AD, appropriately diagnosed and identified according to the ADNI criteria.

In total, 100 neuroimages were downloaded for the CN group and 100 neuroimages for the AD group. The acquisition dates for these neuroimages were between the years 2006 and 2020. The protocol consisted of 30 min 3-dimensional brain scans 30-60 min after the injection of 185 $\pm$ 18.5 MBq of 18F-FDG (more details can be found here). The minimum age of subjects whose neuroimages were collected was set at 50 years. Demographic data regarding this dataset are available in table 1.

3.1.2. FDG PET neuroimages—CDI

The neuroimages from the clinical routine were collected from a folder organized by date and by patients, which is kept as backup. For there to be correspondence between the clinic patients and the AD and CN groups of the ADNI, only neuroimages that, according to the medical report, referred to subjects without evidence of any neurodegenerative disease, but with possible microangiopathy (characteristic of aging) (nAD group) or subjects whose neuroimage was suggestive of AD, without the possibility of other neurodegenerative diseases (AD group) were collected.

The FDG PET neuroimages were acquired at resting-state, in fasting subjects for at least 4 h with a normal glycemic level, using an integrated PET/CT Biograph Siemens camera (Erlangen, Germany), after an intravenous administration of 100 MBq, per 10 min acquisition at 45 min post-injection. PET image reconstruction was performed using the attenuation-weighted ordered subsets expectation maximization (4 iterations, 21 subsets, 4 mm Gaussian postfilter). PET image matrix size was 336 $\times$ 336 $\times$ 110 (1.0182 mm $\times$ 1.0182 mm $\times$ 1.5000 mm spacing) voxels, with Gaussian post-reconstruction filter and

corrected for attenuation using a CT transmission scan. Thus, 100 neuroimages were collected for the nAD group and 92 for the AD group. The minimum age of the subjects whose neuroimages were collected was set at 50 years. Demographic data regarding this dataset are available in table 1.

3.2. Data processing

All neuroimages were converted to *NIfTI* format, allowing them to be manipulated by *SPM12*. The neuroimages were reoriented in order to standardize the relative positioning of the brain and then spatial normalization was performed. The spatial normalization template was the default for *SPM12* (MNI space). After spatial normalization, all neuroimages were smoothed using the *smoothing* filter from *SPM12*, also with the default settings.

A mask was generated from the neuroimages of the CN group using ScanVP (Spetsieris *et al* 2013) (with a *threshold* of 20%) and later applied to all neuroimages.

3.3. Data organization

All neuroimages were manipulated in *Python* so that each voxel of the same neuroimage was a distinct column of a matrix row. Then, all elements of the same row were normalized between 0 and 1 using the maximum-minimum (linear scaling) method, so that voxels from different neuroimages have the same maximum and minimum range. Finally, the data were organized as follows:

- Training/validation data: referring to ADNI data. This dataset was used for the optimization and training of the classification model.
- Test data: referring to data from CDI. After adjusting the final classification model, it was tested on the data obtained from the clinical routine.

3.4. Optimization pipeline and model training

In *Python*, it is possible to estimate the best combination of hyperparameters using *GridSearchCV* , which is a function from the *scikit-learn* library (Pedregosa *et al* 2011) that trains and tests, for different combinations of hyperparameters, the classification model, returning its final performance in several metrics (accuracy, recall, precision and F1-score, for example). The metrics are estimated using the *K-Fold cross validation* method, where, in this work, 5 *folds* of identical size were used. In this way, through the automation allowed by the *GridSearchCV* function, we explored the metrics for all combinations of the hyperparameter space. The combination with the best results were selected.

Briefly, for each hyperparameter combination, the following steps were executed:

(i) *K-Fold cross validation* divides the training data into 5 training and test folds;

(ii) PCA is applied to the training folds, extracting the principal components that add up to 80% of the total variance explained in descending order of importance;

(iii) In principal components space, the training folds are used to train a LR model with the combinations of hyperparameters under consideration;

(iv) With the model trained, it is applied over the test fold;

(v) The final metric is estimated as the average of the calculated metrics for all test folds.

The estimated metrics were accuracy, recall, precision and F1-score, which is defined as the harmonic mean of precision and recall. The hyperparameter space used to generate the combinations was:

- penalty (type of regularization): 'l1', 'l2' and 'none';
- C (inverse of λ, which indicates the emphasis of the regularization method): a logarithmic sequence with a total of 25 terms, from 0.01 to 100;

For PCA, whiten = True was used, so that, at each iteration of *GridSearchCV* , the normalization over each predictor variable by the mean and standard deviation was carried out; for LR, solver = 'saga' was used, as it is compatible with all types of regularization tested.

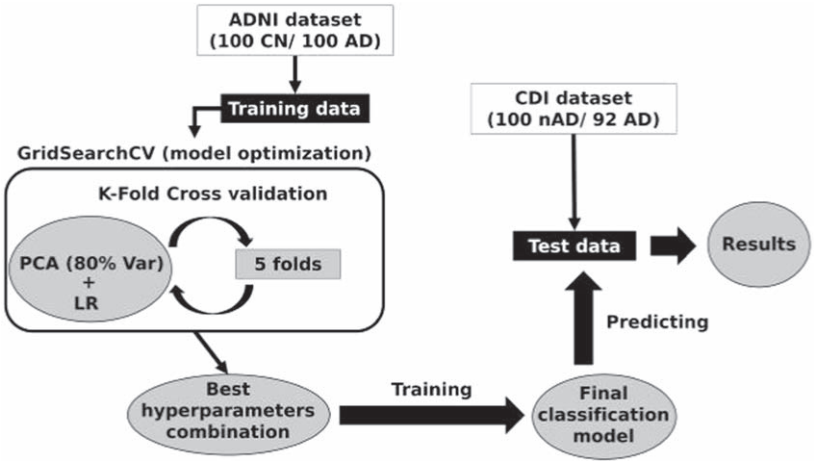

Figure 1. Flowchart summarizing the methodology. First, training and test data are separated. Then the training data is used for hyperparameter optimization. Once the best combination of hyperparameters is found, it is used to train the final model on all the training data and finally the model is applied on the test data.

3.5. Testing of the classification model

The best combination of hyperparameters was determined using the F1-score (as it expresses a balance between recall and precision) and then used to train the classification model over all training data. As in the optimization pipeline, the classification model consisted of applying PCA, extracting the principal components that add up to 80% of the total explained variance in descending order of importance, followed by training a LR model.

For testing, the results obtained from the final classification model on the test data were analyzed. As a brief summary, the entire methodology is shown as a flowchart in figure 1.

Please be aware that, for the three additional models trained for the head-to-head comparison, we adhered to an identical procedural methodology. This entailed the optimization of hyperparameters through *GridSearchCV* and training on the ADNI dataset (principal components as features). Subsequently, the models were applied to the test data. For comprehensive details regarding the hyperparameter space explored for each model and supplementary findings, we direct readers to consult the supplementary materials.

4. Results

Figure 2 shows how the mean values for each metric evolved in the hyperparameters optimization step for the two types of regularization tested. In the absence of regularization parameters, the mean achieved for each metric was: accuracy = 85.00%, sensitivity = 84.00%, precision = 85.46% and F1-score = 84.62%. Thus, of all 51 combinations of hyperparameters tested, the best in terms of F1-score was $C \approx 0.316$ and penalty = 'L1'. In table 2 the metrics are shown for each test fold for the best combination of hyperparameters.

With the classification model trained on the entire training dataset using the best combination of hyperparameters, many principal components had zero weight. By figure 3, it is possible to account for 23 principal components, out of a total of 50, with weight equal to zero.

Taking the linear combination of all principal components adjusted by their respective weights, it is possible to interpret the decision rule of the classification model. The result of this linear combination is shown in figure 4. A reasonable way to read figure 4 is that voxels with high intensity (higher brain metabolism) in the regions in red are associated with a higher probability for the AD group, while the regions in blue are associated with a lower probability.

The confusion matrix referring to the application of the classification model on the test data is shown in figure 5. In it, it is possible to notice that there were 170 true positives, 13 false positives, and 9 false negatives on the test data, which leads to a total accuracy of 88.54%. Regarding the more specific metrics, a recall of 90.22%, a precision of 86.46%, and a F1-score of 88.30% were obtained. The ROC curve of the model over the test data is shown in figure 6 and resulted in an AUC = 94.75%.

The comparison of the final metrics on the test data with those obtained by the additional trained models (SVM, MLP and RF) is shown in table 3. The statistical difference between the prediction accuracy of the additional models and the LR was verified via the McNemar test (Japkowicz and Shah 2011) (95% confidence level: $\alpha < 0.05$ for significance). Further details and results regarding the additional models can be found in the supplementary materials.

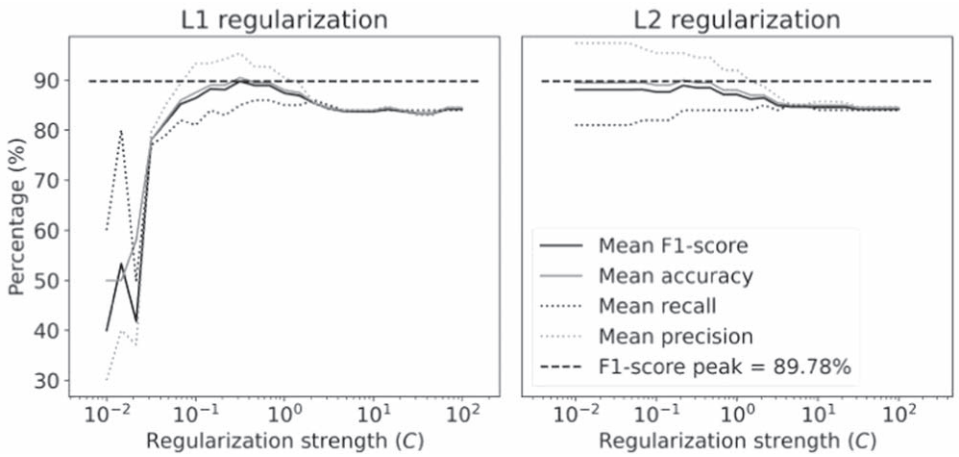

Figure 2. Evolution of the classification model metrics in the hyperparameters optimization step for the two tested regularization methods and for different regularization strengths. Note that, although the L2-type regularization was more stable for the accuracy and F1-score, the L1-type regularization showed a higher peak for both.

Table 2. All metrics referring to the best combination of hyperparameters for the test folds. The mean and standard deviation for each metric for all folds were also calculated.

	Results on test folds						
Metric	1° (%)	2° (%)	3° (%)	4° (%)	5° (%)	Mean (%)	S.D.[a] (%)
Accuracy	92.50	85.00	87.50	87.50	100	90.50	5.34
F1-score	91.89	83.33	86.49	87.18	100	89.78	5.79
Recall	85.00	75.00	80.00	85.00	100	85.00	8.37
Precision	100	93.75	94.12	89.47	100	95.47	4.04

[a] Standard deviation.

As the LR model predicts probability values, it is possible to construct a scatter plot indicating the log-transformed odds (standardized by the mean and standard deviation of the CN group) of each neuroimage as a more intuitive method of analyzing the results. We have the distribution of points shown in figure 7.

As output scores are directly related to the degree of expression of the AD biomarker pattern (figure 4), it is possible that they also indicate disease progression. For this purpose, four FDG PET neuroimages were selected: one from a healthy subject while the others show different stages of AD (early, moderate, and advanced). In figure 8, axial slices of these spatially normalized neuroimages are shown, as well as their respective output scores. As a second example, in figure 9 it is shown how the score of a patient with AD increases as the disease progresses after one year and three months.

5. Discussion

Regarding the hyperparameter optimization step, it is possible to notice in figure 2 that the L2 regularization showed greater stability within the space of tested values. This was to be expected since such a regularization method is conservative in the selection of predictor variables. The best results, however, only appeared for the L1-type regularization, which is more aggressive in the selection of predictor variables. In this last case, there is a sharp peak for the accuracy and F1-score within a very defined range of values for the hyperparameter *C*.

For the best combination of hyperparameters, the mean accuracy for the 5 test folds was greater than 90%, a result that indicates good generalization of the model. Note, however, that for the 5th fold the accuracy was 100% (not just the accuracy, but all other metrics calculated). As the separation of folds is random, this result is explained simply by luck, although such results on test data are uncommon. It is also worth remembering that each test fold consists of only 40 neuroimages (there are 5 folds in total for the 200 neuroimages provided), which increases the probability of such results. The average for the recall was 85.00%, with a standard deviation close to 10%, suggesting an acceptable sensitivity for the AD group. For precision, the average was close to 95%, indicating a very low false positive rate. Finally, the average for the F1-score resulted in 89.78%, a value close to 90%, which is a strong indication of an optimal balance between recall and precision.

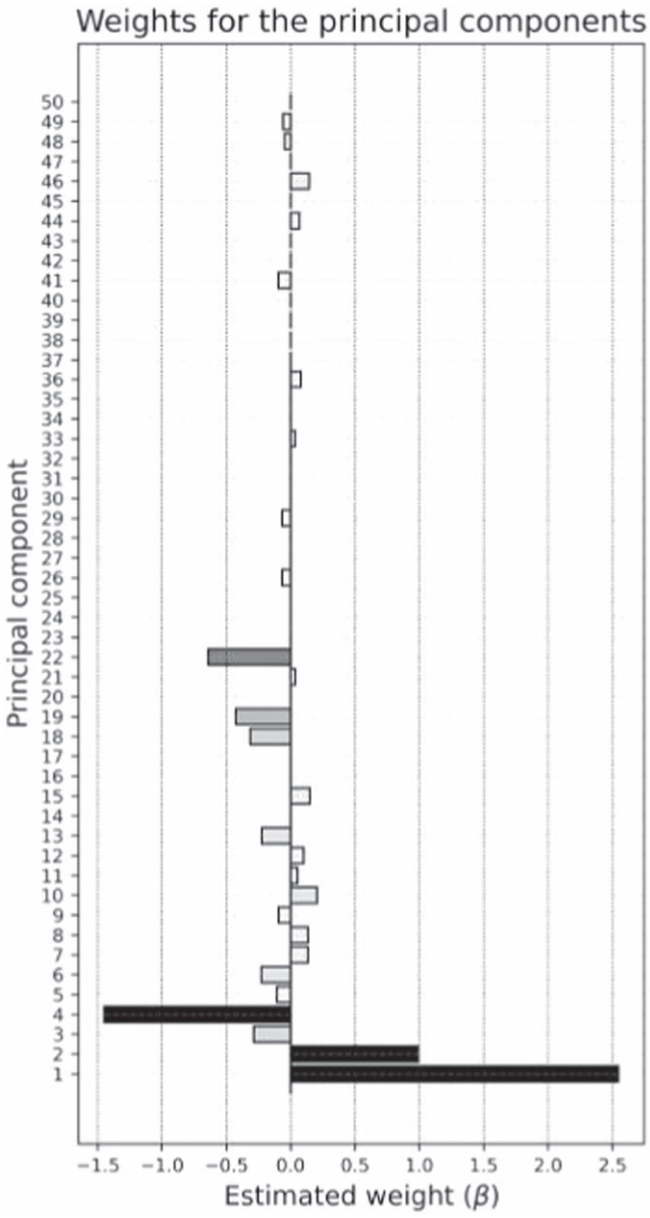

Figure 3. Principal component weights after training the classification model over all training data. The more positive the weight, the greater the increase in probability for AD. The more negative, the greater the decrease in probability. Both cases reflect a high predictive power (darker colors in the figure). Null values indicate no predictive power. The index of each principal component indicate its position in the order of total explained variance.

After adjusting the classification model on all training data using the best combination of hyperparameters and applying it to the test data, accuracy and F1-score (figure 5) had values similar to those obtained in the hyperparameter optimization step (table 2). Bearing in mind that the model was trained on ADNI data (publicly available) and the testing was performed on clinical routine data from CDI, those results are very good, although they exceeded 90% only for recall.

The specific classification metrics for the AD group (recall = 90.22%, precision = 86.46%, F1-score = 88.30%, AUC = 94.75%) can be used to compare with similar works in the literature involving neural networks. In Ding *et al* (2019) (Inception V3 architecture), it was used for training a total of 1921 imaging studies (ADNI) and 188 (ADNI) for testing. The results for AD group (against MCI and nAD groups) were recall = 81%, precision = 76%, F1-score = 78%, AUC = 92%. When the model were tested outside the ADNI set, the results were recall = 100%, precision = 54% and F1-score = 70% for a total of 40 imaging studies. In Singh Singh *et al* (2017), data from ADNI (186 CN, 178 early MCI, 158 late MCI, 146 AD) were used to train and test a Multilayer Perceptron model after applying dimensionality reduction by probabilistic PCA. The results for the AD group (against CN group) were recall = 96.32%, precision = 98.39%, F1-score = 97.34% and AUC = 95%, all of them calculated as the mean for 10 folds by cross validation. Although a detailed comparison cannot be performed due

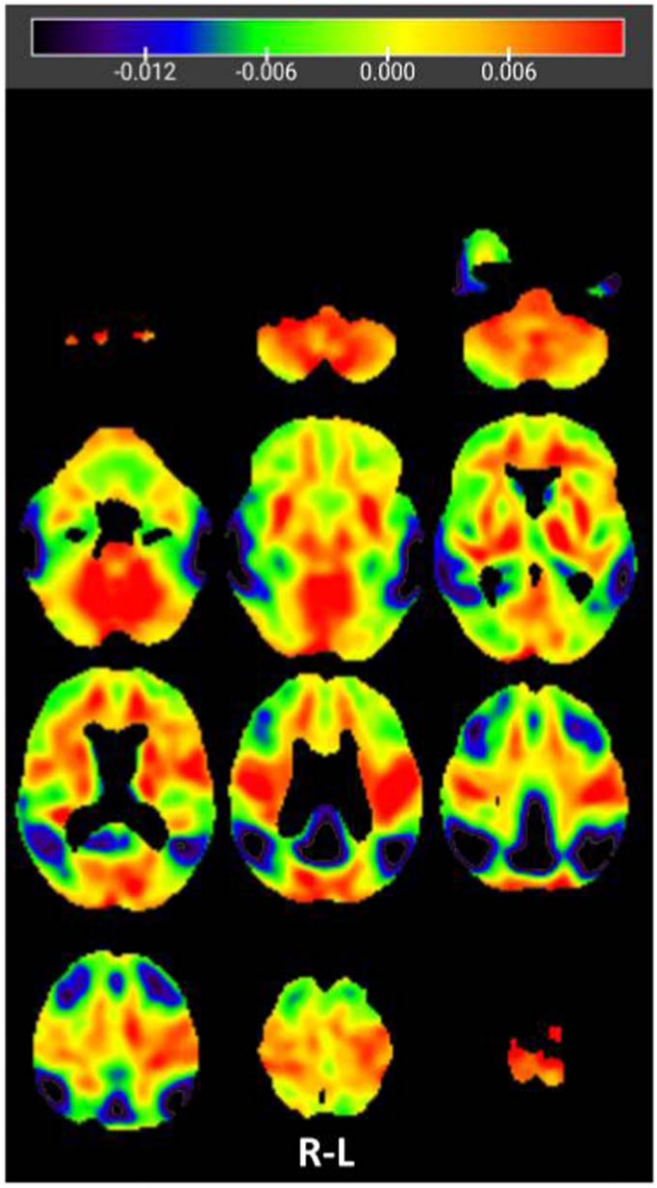

Figure 4. Resulting linear combination of principal components adjusted by their respective weights in a sequence of axial slices. Regions in red are associated with a higher probability for AD given high intensities in the voxels while regions in blue are associated with a lower probability for AD given high intensities in the voxels.

to the difference in the number of classes used, the specific results of this work for the AD group compare well to results in the literature using neural networks. Even though it has not been tested on data outside the ADNI set, Singh *et al* (2017) shows that the classification performance of neural networks can be extremely good when taking the MCI class and its stages into account. It also shows that there is an improvement in model performance when demographic variables are included (Singh *et al* 2017).

As a better way to evaluate the performance of the LR model, it is convenient to train additional classification models to perform a head-to-head comparison over the test data. In this sense, the SVM, MLP and RF models were trained and optimized over the ADNI data, and the final metrics obtained by them over the test set are shown in table 3. Note that, in terms of total accuracy, F1-score and AUC, the LR model was better than other models. However, the RF and SVM models were better than the LR in terms of recall and precision, respectively. Despite the differences in the metrics values, it is important to highlight that there was no statistically significant difference between the predictive accuracy of the additional models and the LR, as indicated by the McNemar test. This result suggests that the LR model performs at least as good as the additional models.

An advantage of using PCA and LR as a classification model lies in the interpretability of the decision rule and the output. As depicted in figure 3, 23 out of 50 principal components were found to have zero weight, indicating their lack of predictive power. This suggests that these covariance patterns are likely attributable to noise. The use of regularization methods, particularly those of the L1 type, offers the benefit of automatic feature selection

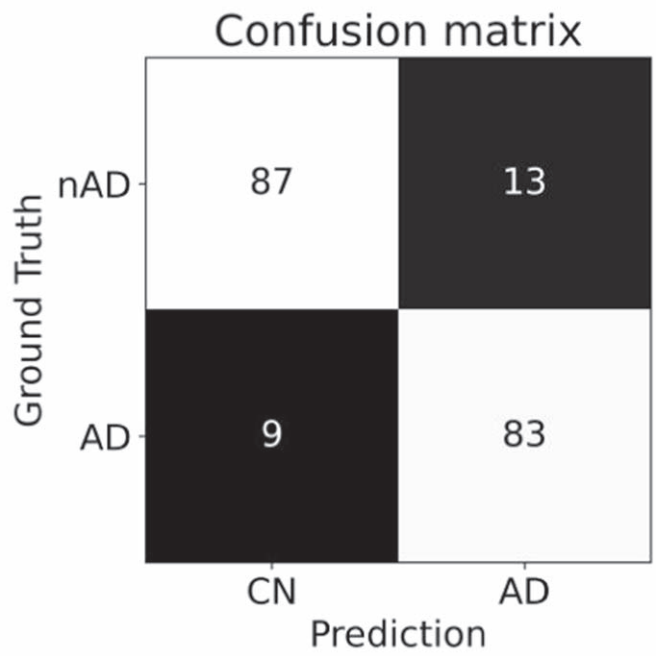

Figure 5. Confusion matrix of the classification model on the test data. There were 170 true positives, 13 false positives, and 9 false negatives. The resulting metrics were: total accuracy = 88.54%, recall = 90.22%, precision = 86.46%, and F1-score = 88.30%.

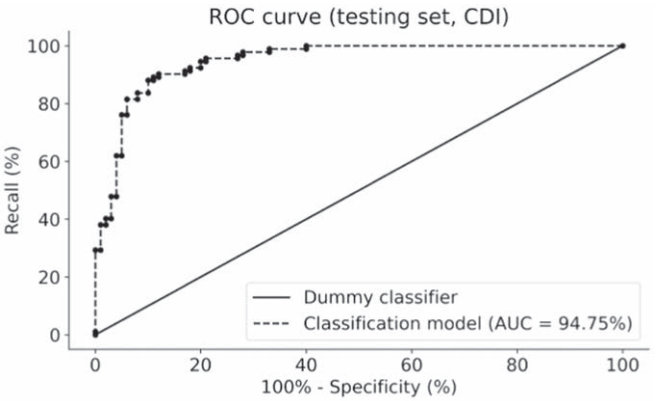

Figure 6. ROC curve of the classification model on the test data, with an AUC = 94.75%.

Table 3. Head-to-head comparison between the specific metrics of the LR model and the additional models over the test set. McNemar test against LR predictions was performed for the additional models (95% confidence).

Model	Acc[a] (%)	Rec[b] (%)	Prec[c] (%)	F1[d] (%)	AUC (%)	*p*-value[e]
LR	88.54	90.22	86.46	88.30	94.75	—
SVM	84.90	78.26	88.89	83.24	93.60	0.286
RF	88.02	92.39	84.16	88.08	93.59	1
MLP	84.38	77.17	88.75	82.56	92.95	0.210

[a] Total accuracy.
[b] Recall.
[c] Precision.
[d] F1-score.
[e] McNemar test against LR.

(Miao and Niu 2016), forcing the weight of noisy principal components to zero. The principal components with non-zero weight probably have direct relation to AD and can be more effectively interpreted when analyzed collectively through linear combination, as is shown in figure 4. The resultant biomarker, as it relates directly to

Figure 7. Scatter plot comparing the log-transformed odds of non-AD (white color) and AD (black color) groups for the two datasets (training and testing sets). The values were standardized by the mean and standard deviation of the CN group.

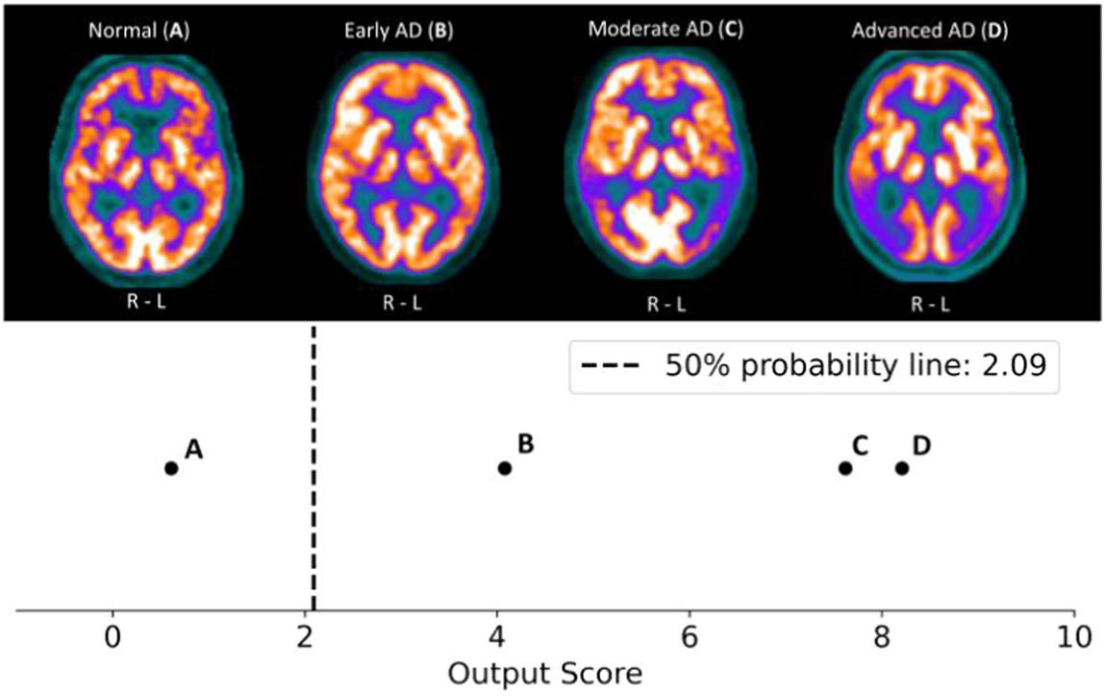

Figure 8. Axial slices of FDG PET neuroimages from the CDI clinic of different patients and their respective output scores from the classification model. All neuroimages were spatially normalized. The scores of patients with labels A, B, C, and D are, respectively, 0.61, 4.08, 7.62, and 8.21.

the probabilites for AD, is consistent with what is expected physiologically, as the dark blue regions (higher probability for AD if there is hypometabolism) are concentrated in the parietotemporal association cortices and the precuneus. This suggests that the classification model was well adjusted and was able to capture the pattern of AD in FDG PET neuroimaging.

Additionaly, the use of log-transformed odds instead of the raw confusion matrix allows a more direct and intuitive notion of the probability for the AD group, as well as the visualization of how the different groups are separated from each other according to the classification model. The log-transformed odds can be seen in figure 7 and show that the points for the nAD group from the test data are more spread out than the CN group from training data. This was to be expected since classification as CN by ADNI takes into account not only FDG

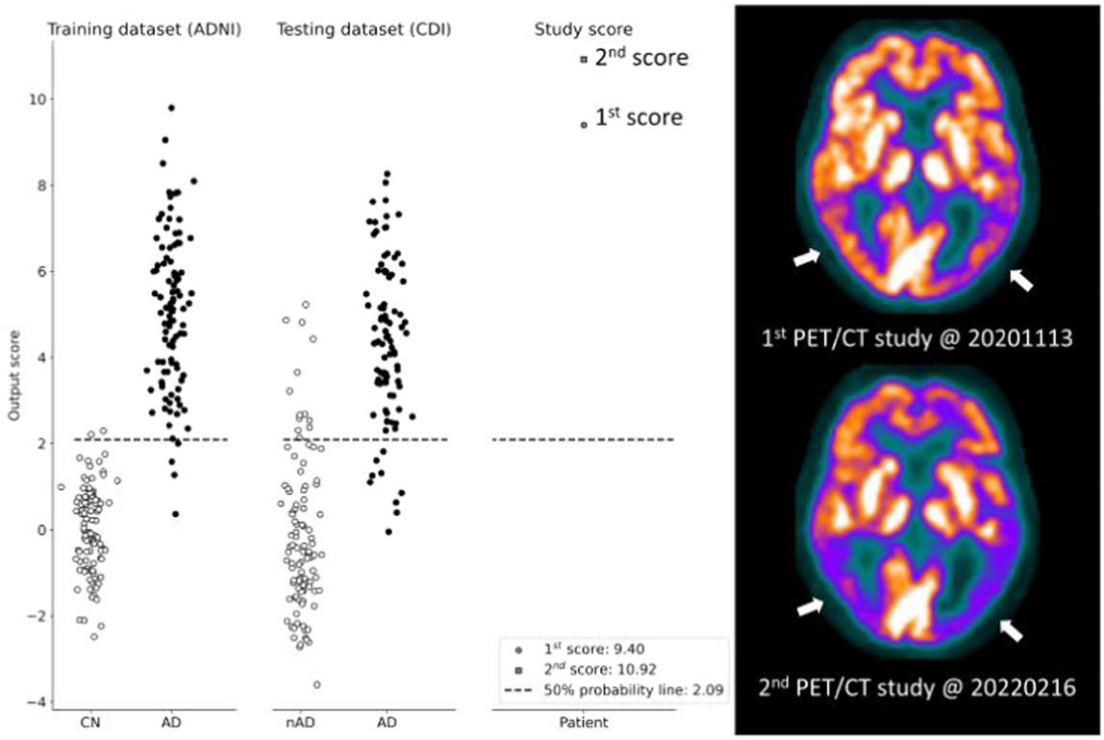

Figure 9. One year and 3 months apart follow-up of a AD patient at the CDI. Note the follow-up scores of 9.40 in the first and 10.92 in the second PET/CT study and how they compare to training and testing scores. In the left, AD and non-AD groups are shown in black and white colors, respectively. The FDG PET neuroimages in the right show the increase of hypometabolism on the bilateral parietal lobes, indicating the progression of AD. Neuroimages were spatially normalized.

PET neuroimages but also cognitive tests, for example. In the case of the CDI clinic, only FDG PET neuroimages were used as reference. Despite this, the non-AD and AD groups are well separated for both datasets.

Beyond its role in classification, the output scores may also reflect the progression of AD. As depicted in figure 8, the score for a healthy patient is near zero and situated to the left of the cutoff line, while the scores for AD patients are to its right, as expected. Among AD patients, the output scores align with the disease's progression. Furthermore, figure 9 illustrates an increase in an AD patient's output score correlating with disease advancement over a period of one year and three months. This is attributed to the output scores quantifying the extent of AD-specific hypometabolism expression (figure 4). However, to establish a definitive correlation between output scores and AD progression, a more rigorous study is recommended. This could involve comparing output scores with cognitive test results such as those from the Mini-Mental State Examination (MMSE), a reliable measure of AD progression.

6. Conclusion

Through optimization of the classification model performed via *GridSearchCV* , it was possible to find the best combination of hyperparameters (penalty = 'l1' and $C \approx 0.316$). With this combination, the model was trained on all the training data and then applied on the test data. The resulting metrics were satisfactory (accuracy = 88.54%, recall = 90.22%, precision = 86.46%, F1-score = 88.30%, AUC = 94.75%), indicating that there was a good fit of the classification model as well as a good generalization to neuroimages outside the training set. Although a detailed comparison with other works in the literature was not possible, the metrics referring to the AD group are comparable to the results of other works (Singh *et al* 2017, Ding *et al* 2019) involving neural networks, indicating that PCA and LR, despite being classic methods, are applicable. A head-to-head comparison with the SVM, MLP and RF classification models showed that LR was better in terms of total accuracy, F1-score and AUC, but it is worth highlighting that there was no statistically significant difference in terms of predictive accuracy between the LR and the additional models. This suggests that the LR performed at least as good as the additional models. It was also shown that the advantages of using PCA and LR consist mainly of interpretability, both in the parameters of the final classification model (which can be summarized as a biomarker, directly related to the pattern of hypometabolism in AD) and in the final output (which can be shown as the log-transformed odds).

The exposed methods are perfectly reproducible even in the absence of high computational power and were carried out with a simple experimental setup, relying only on a total of 200 FDG PET neuroimages from ADNI for training and testing the model on 192 clinical routine neuroimages. Because the log-transformed odds directly represent the degree of expression of the hypometabolism pattern for AD, we draw attention to the possibility of future studies involving the relationship between output scores and AD progression. Although in this work we have shown some selected examples in this sense, a more precise and robust study should be carried out.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson and Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck and Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

The authors also thank the technical and medical staff of the CDI clinic for their willingness and for all the support provided.

In the development of this work, the authors contributed as follows:

- Carlos Eduardo Gonçalves de Oliveira: programming and analysis, study execution, manuscript writing and reviewing;
- Whemberton Martins de Araújo: medical PET-CT image evaluation, dataset retrieving, study execution, manuscript reviewing;
- Ana Beatriz Marinho de Jesus Teixeira: medical PET-CT image evaluation, study execution, manuscript reviewing;
- Gustavo Lopes Gonçalves: programming and manuscript reviewing;
- Emerson Nobuyuki Itikawa: study supervision, manuscript writing, manuscript reviewing;

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/cego669/PCA-and-LR-as-a-diagnostic-tool-for-AD. Data will be available from 26 June 2023.

Ethical statement

This research was evaluated by the Ethical Committee of the Federal University of Goiás with the ID number 69048823.9.0000.5083. Our study is exempt of informed consent as it was a transversal, retrospective study. Furthermore, this research is in accordance with the Declaration of Helsinki, and the national regulations involving humans (CNS 466/2012).

Competing interests statement

The authors have no relevant financial or non-financial interests to disclose.

ORCID iDs

Emerson Nobuyuki Itikawa https://orcid.org/0000-0001-5478-6203

References

Abdi H and Williams L J 2010 Principal Component Analysis *Wiley Interdiscip. Rev. Comput. Stat.* **2** 433–59
Arvanitakis Z, Shah R C and Bennett D A 2019 Diagnosis and Management of Dementia: Review *Jama* **322** 1589–99
Blazhenets G, Ma Y, Sörensen A, Rücker G, Schiller F, Eidelberg D, Frings L and Meyer P T 2019 Principal Components Analysis of Brain Metabolism Predicts Development of Alzheimer Dementia *J. Nucl. Med.* **60** 837–43
Breiman L 2001 Random Forests *Mach. Learn.* **45** 5–32
Brown R K, Bohnen N I, Wong K K, Minoshima S and Frey K A 2014 Brain PET in Suspected Dementia: Patterns of Altered FDG Metabolism *Radiographics* **34** 684–701
Ding Y *et al* 2019 A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain *Radiology* **290** 456–64
Dukart J *et al* 2013 Meta-analysis based SVM classification enables accurate detection of Alzheimer's disease across different clinical centers using FDG-PET and MRI *Psychiatry Res.: Neuroimaging* **212** 230–6
Etminani K *et al* 2022 A 3D deep learning model to predict the diagnosis of dementia with Lewy bodies, Alzheimer's disease, and mild cognitive impairment using brain 18F-FDG PET *Eur. J. Nucl. Med. Mol. Imaging* **49** 563–84
Friedman J, Hastie T and Tibshirani R 2010 Regularization Paths for Generalized Linear Models via Coordinate Descent *J. Stat. Softw.* **33** 1–22
Habeck C, Foster N L, Perneczky R, Kurz A, Alexopoulos P, Koeppe R A, Drzezga A and Stern Y 2008 Multivariate and univariate neuroimaging biomarkers of Alzheimer's disease *Neuroimage* **40** 1503–15
Habeck C and Stern Y 2010 Multivariate Data Analysis for Neuroimaging Data: Overview and Application to Alzheimer's Disease *Cell Biochem. Biophys.* **58** 53–67
Habeck C G 2010 Basics of Multivariate Analysis in Neuroimaging Data *J. Vis. Exp.* **41** e1988
Hao X, Zhang G and Ma S 2016 Deep Learning *Int. J. Semant. Comput.* **10** 417–39
Japkowicz N and Shah M 2011 *Evaluating Learning Algorithms: A Classification Perspective* (Cambridge University Press)
Lai Y 2019 A comparison of traditional machine learning and deep learning in image recognition *J. Phys.: Conf. Ser.* **1314** 012148
Liu J, Chen J and Ye J 2009 Large-scale sparse logistic regression *Proc. of the XV ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining* pp 547–56
Lu S *et al* 2017 Early identification of mild cognitive impairment using incomplete random forest-robust support vector machine and FDG-PET imaging *Comput. Med. Imaging Graph* **60** 35–41
Marcus C, Mena E and Subramaniam R M 2014 Brain PET in the Diagnosis of Alzheimer's Disease *Clin. Nucl. Med.* **39** e413–e426
Miao J and Niu L 2016 A Survey on Feature Selection *Proc. Comput. Sci.* **91** 919–26
Nancy Noella R and Priyadarshini J 2023 Machine learning algorithms for the diagnosis of Alzheimer and Parkinson disease *J. Med. Eng. Technol.* **47** 35–43
Noble W S 2006 What is a support vector machine? *Nat. Biotechnol.* **24** 1565–7
Pedregosa F *et al* 2011 Scikit-learn: Machine Learning in Python *J. Mach. Learn. Res.* **12** 2825–30
Popescu M C, Balas V E, Perescu-Popescu L and Mastorakis N 2009 Multilayer Perceptron and Neural Networks *WSEAS Trans. Circ. Syst.* **8** 579–88
Salehi F, Abbasi E and Hassibi B 2019 The Impact of Regularization on High-dimensional Logistic Regression *Adv. Neural Inf. Process. Syst.* **32** 1–116476
Santo R D E 2012 Utilização da Análise de Componentes Principais na compressão de imagens digitais *Einstein (São Paulo)* **10** 135–9
Shinde P P and Shah S 2018 A review of machine learning and deep learning applications *2018 Fourth Int. Conf. on Computing Communication Control and Automation (ICCUBEA)* (IEEE) pp 1–6
Singh S, Srivastava A, Mi L, Caselli R J, Chen K, Goradia D, Reiman E M and Wang Y 2017 Deep-learning-based classification of fdg-pet data for alzheimer's disease categories *13th Int. Conf. on Medical Information Processing and Analysis* vol 10572 (SPIE) pp 143–58
Spetsieris P, Ma Y, Peng S, Ko J H, Dhawan V, Tang C C and Eidelberg D 2013 Identification of Disease-related Spatial Covariance Patterns using Neuroimaging Data *J. Vis. Exp.* **76** e50319
Zohuri B and Moghaddam M 2020 Deep learning limitations and flaws *Mod. Approaches Mater. Sci.* **2** 241–50