IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Optimized One vs One approach in multiclass classification for early Alzheimer's Disease and Mild Cognitive Impairment diagnosis.

CARMEN JIMÉNEZ-MESA[1], IGNACIO A. ILLAN[1],ALBERTO MARTÍN-MARTÍN[1] , DIEGO CASTILLO-BARNES[1], FRANCISCO JESUS MARTINEZ-MURCIA[1], JAVIER RAMIREZ[1], AND JUAN M. GORRIZ.[1]

[1]Signal Theory and Communications, University of Granada, 18071 Granada, Spain (e-mail: illan@ugr.es)

Corresponding author: (e-mail: gorriz@ugr.es)

This work was supported by the Ministerio de Innovación y Ciencia project DEEP-NEUROMAPS RTI2018-098913-B100, the Consejería de Economía, Innovación, Ciencia, y Empleo of the Junta de Andalucía under the A-TIC-080-UGR18 TIC FRONTERA, and the Ministerio de Universidades under the FPU predoctoral grant (FPU 18/04902)."

ABSTRACT The detection of Alzheimer's Disease in its early stages is crucial for patient care and drugs development. Motivated by this fact, the neuroimaging community has extensively applied machine learning techniques to the early diagnosis problem with promising results. The organization of challenges has helped the community to address different raised problems and to standardize the approaches to the problem. In this work we use the data from international challenge for automated prediction of MCI from MRI data to address the multiclass classification problem. We propose a novel multiclass classification approach that addresses the outlier detection problem, uses pairwise t-test feature selection, project the selected features onto a Partial-Least-Squares multiclass subspace, and applies one-versus-one error correction output codes classification. The proposed method yields to an accuracy of 67 % in the multiclass classification, outperforming all the proposals of the competition.

INDEX TERMS Alzheimer's Disease, CAD, error correcting output codes, Mild Cognitive Impairment, Multiclass classification, One versus One, Partial least squares, Random forests, Support Vector Machines.

I. INTRODUCTION

ASSISTANCE in diagnosis of Alzheimer's Disease (AD) in it's early stages has received a constant interest from the medical imaging community in the past decades due to its medical importance and societal implications [8]. Distinguishing between AD and its related neurological disorders, including its prodromal stage Mild Cognitive Impairment (MCI), is very challenging from the clinical evaluation point of view, predominantly in the early stages of the disease. Medical imaging techniques, such as Magnetic Resonance Imaging (MRI) or Positron Emission Tomography (PET), have provided new tools to assess the subject conditions in a non-invasive way. However, both the subtle changes produced in the brain and the lack of a complete understanding of the disease development still pose challenges to the diagnosis assistance through brain images [45], [47]. Concretely, the discrimination between MCI and AD has been shown to be a difficult task [24], [26], [30], [41], [42]. Machine learning applications in neuroimaging have become an indispensable tool for brain image analysis and computer aided diagnosis (CAD) systems, producing a prolific area of research [8]. However, the lack of standardized datasets hinders direct comparisons of approaches, and the identification of their virtues.

The open data policy and the creation of big databases have facilitated the organization of competitions for improving the CAD systems for AD diagnosis, such as CADDementia [3] and TADPOLE (https://tadpole.grand-challenge.org/), among others. The availability of the data for posterior analysis allows for retrospective analysis of the competitions and new submission proposals. It also facilitates the reproducibility of results, a problem that is becoming of central interest in neuroimaging research [4].

This work makes use of the International challenge for automated prediction of MCI data. The objective of the competition was the development of CAD systems for the multiclass classification of 4 classes: Healthy Controls (HC), Mild Cognitive Impaired (MCI) subjects, Mild cognitive impaired subjects that converted into Alzheimer's Disease during the study (cMCI), and Alzheimer's Disease (AD) patients. The challenge provided with preprocessed MRI data of the different classes to allow participant proposals of optimized CAD systems, based on the finding that combining multiple anatomical measures improves classification of early diagnosis of AD [44]. The results of the challenge were published in the special issue [38] on the Journal of Neuroscience Methods. The winner proposal used a random forest ensemble with feature extraction methods [10], yielding to a 61 % accuracy in the multiclass classification problem.

Ensemble methods have been successfully applied to neuroimaging problem [7], [20], [32], [33]. It has also been proven that multiclass approaches using binary classifiers can be a competitive solution to the problem, such as those based on one-versus-one approaches, one-versus-rest in the context of Error Output Correcting codes (ECOC) [6], [9], [13], [46]. This study uses and compares ensemble methods and aggregation methods by binary classifiers, as those of highest rated approaches in the challenge [5].

To optimize the multiclass classification through combination of anatomical features, not only classifier aggregations are necessary, but also feature extraction techniques [36]. Different approaches to feature selection and extraction reported high relevance in the literature, with high accuracy and also a correspondence between automatic cortical and subcortical region selection and clinical findings [7]. Concretely, brain atrophy has been found to be relevant for AD diagnosis in white matter cortical and subcortical regions, as well as hippocampal volume, cortical thickness, and grey matter density, thus making feature extraction a reasonable preprocessing step (see [44] and references therein). One of such successful methods for feature selection and combination is Partial Least Squares, linearly transforming the data into a space maximal separation between classes [25], [35], [39], [40].

Through the extensive use of feature extraction and one-vs-one feature selection and classification, we propose and study a CAD system for identification of early stages of AD and MCI that optimizes the combination multiple anatomical measures of atrophy in the brain to improve classification performance.

II. METHODS

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database. 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 MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. For up-to-date information, see www.adni-info.org.

N = 240	Male/female	Age	MMSE
HC	30/30	72.34 [5.67]	29.15 [1.11]
MCI	28/32	72.19 [7.42]	28.32 [1.56]
cMCI	35/25	72.96 [7.20]	27.18 [1.87]
AD	29/31	74.75 [7.31]	23.43 [2.11]

TABLE 1. Training dataset (sociodemographic data and MMSE for each group). $X[Y]$ denotes the mean X and standard deviation Y for each group.

N = 160	Male/female	Age	MMSE
HC	18/22	74.88 [5.48]	29.00 [1.10]
MCI	23/17	72.40 [8.04]	27.65 [1.87]
cMCI	25/15	71.75 [6.23]	27.58 [1.80]
AD	23/17	73.11 [8.05]	22.68 [1.98]

TABLE 2. Real data in testing dataset (sociodemographic data and MMSE for each group). $X[Y]$ denotes the mean $X$and standard deviation Y for each group.

A. DATASETS

This section shows the datasets that were provided for the International challenge for automated prediction of MCI from MRI data (https://inclass.kaggle.com/c/mci-prediction). MRIs were selected from the ADNI and preprocessed by Freesurfer (v5.3) [14], [15]. In total 429 demographical, clinical as well as cortical and subcortical MRI features were available for each subject. Fig. 2 shows the average values for different regions across the brain for the four available classes.

Two different datasets were provided for training and testing the proposed methods for automated prediction of MCI from MRI data. According to their diagnosis, patients were grouped into four classes: healthy control (HC) subjects, AD patients, MCI subjects whose diagnosis did not change in the follow-up (MCI) and converter MCI (cMCI) subjects that progressed from MCI to AD in the follow-up of the disease. The training dataset consisted of 240 ADNI real subjects (60 HC, 60 MCI, 60 cMCI and 60 AD). Demographic information is shown in Tables 1 and 2. The testing dataset consisted of 500 subjects. 160 out of them were real subjects, whereas the 340 remaining subjects were artificially generated from the real data. Table 2 shows demographic information of only the 160 real patients excluding 340 dummy subjects in the testing dataset. No information about the class labels of the test set was available during the competition. The test set was half split into public and private test sets and only the accuracy score on the public dataset was available for competitors until the challenge ended. Once the challenge finished, class labels for the subjects on the test set were provided to the competitors. The accuracy score on the real subjects of the testing set was used as the figure of merit in the competition.

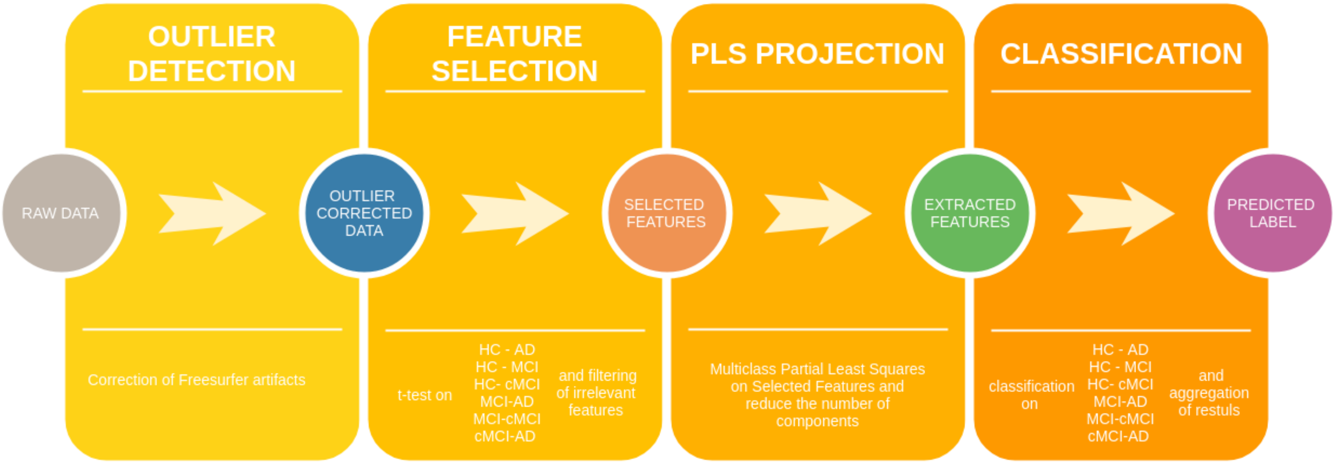


FIGURE 1. Flowchart of the proposed method.

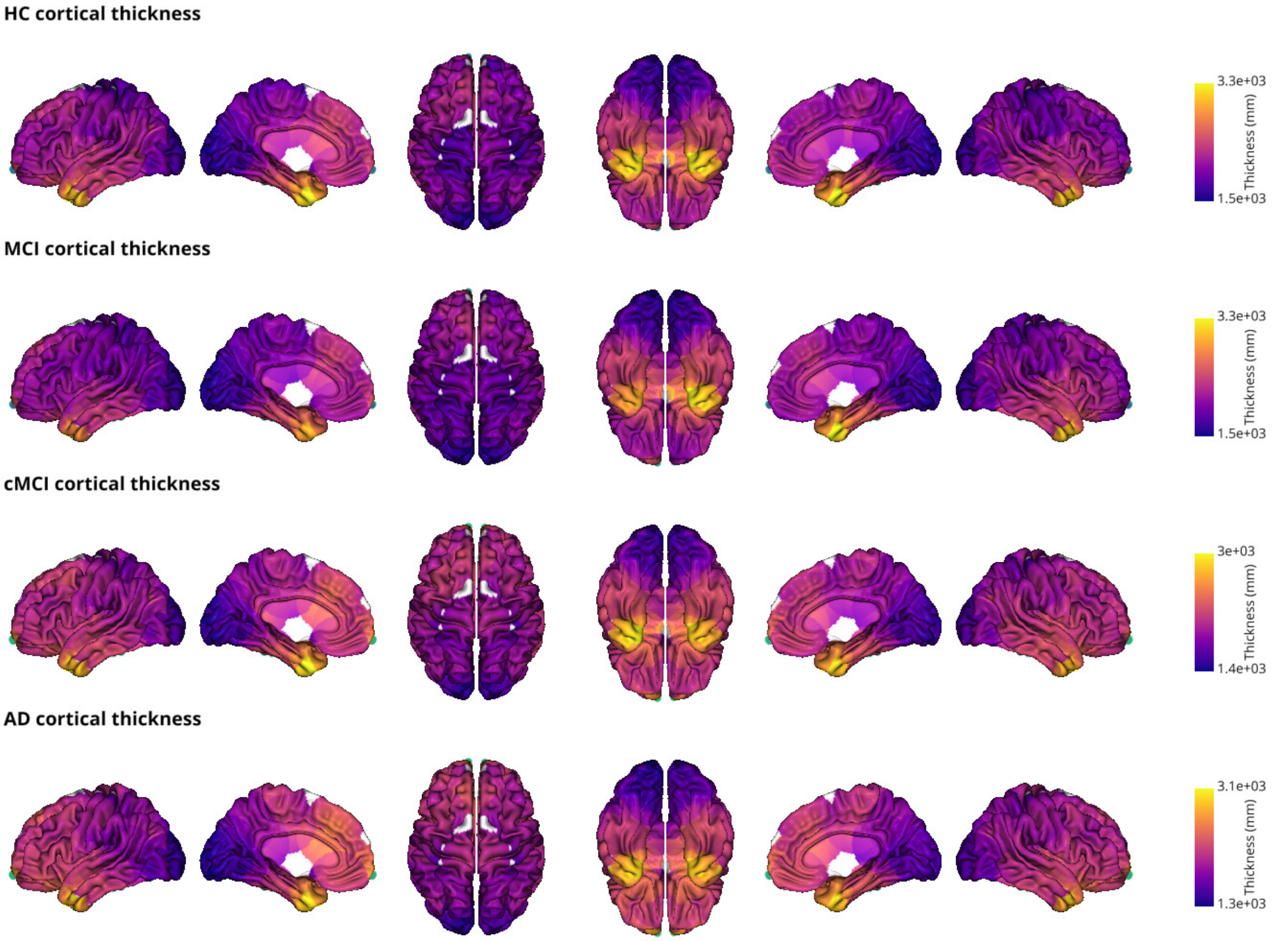


FIGURE 2. Mean cortical thickness of the database for each class.

B. WORKFLOW

The methodology followed in this work aims at optimizing the binary classification of the different classes, HC, MCI, cMCI and AD in the multiclass classification problem, so that the overall classification performance is increased. To that aim, the process is divided in four steps as depicted in Fig. 1. A first preprocessing step is applied to discard outliers and standardize the data. Secondly, based on the observation of the existence of irrelevant or redundant data, a filter is applied to eliminate unimportant features. Once a set of features is selected, a combination of statistical tests and Partial Least Squares (PLS) techniques are used to extract features at binary level for each one vs. one classification (HC vs MCI, HC vs. cMCI, etc..). Finally, binary classifiers are trained on these data, and an aggregation method is proposed for achieving a final multiclass decision.

1) Preprocessing

The presence of outliers is usually an undesirable source of instabilities for machine learning applications. In neuroimaging, outliers are specially challenging as they are frequently found due to acquisition, scanner differences, preprocessing artefacts or resulting from large intrinsic inter-subject variability, having a dramatic effect on the statistical based analysis [17].

A carefully analysis of the data reveals a high abundance of outliers on each of the 429 data features. In Fig. 3, the presence of outliers is depicted for selected features as measured in standard deviations, with values exceeding 8 times the standard deviation.

A common preprocessing step in machine learning consists of centering the data to zero mean and one standard deviation values, usually known as z-score values [19]. However, as Fig. 4a) shows, the z-score values of the data contain high salt-and-pepper type noise. We attributed this effect to a miss-transformation of the data format, coming from freesurfer software, as described by the challenge organizers. The outlier correction algorithm is described in box 1. The algorithm results are shown in Fig. 4b). Correcting this format defect reveals a different data structure, with high redundancy, justifying posterior steps. Concretely, the features sets 1-35, 45-73, 71-139, 140-277, 278-347, 348-413, 413-429, seem to contain very low inter-patient variability, suggesting that the feature space dimension can be highly reduced. In posterior sections, we show that feature space dimension can be optimally reduced to a value below 20.

```
Data: Raw data matrix D of r features and s subjects
Result: Clear outlier values
Compute median values $M(s)$ of D for each feature ;
for i from 1 to r do
    for j from 1 to s do
        if $D(i,j) > 50 * M(j)$ then
            Replace outlier value by $D(i,j)/1000$;
        else if $D(i,j) < 50 * M(j)$ then
            Replace outlier value by $D(i,j) * 1000$;
        end
    end
end
```

Algorithm 1: Outlier elimination algorithm

2) Feature selection and extraction

The preprocessing step is followed by the elimination of irrelevant features and the extraction of features for classification. The former is a filter, in a one vs. one approach. The features are sorted according to a specific criteria, thus eliminating the features with the lowest relevance. The latter is achieved under a multiclass PLS transformation of the selected features, reducing the feature space dimension [21].

The sorting criteria is based on binary comparisons between classes: HC vs. MCI, HC vs. cMCI, HC vs. AD, MCI vs. cMCI, MCI vs. AD, and cMCI vs. AD. For each binary comparison a t-test is performed for each feature, and the features f_i are sorted according to their value of the t statistic, t_i, $i = 1, 2, ..., 429$. A 6x429 matrix S of sorted features is generated in this process. From this matrix S, a submatrix T is constructed by eliminating the n last columns, ordered by decreasing value of the t statistic. The number of times m a feature appeared in the matrix T was calculated for each feature. The parameter m is a significance measure for each feature and is constrained: $0 \leq m \leq 6$. All the features with a value of m under a fixed threshold R where filtered out, resulting in a feature selected set $\mathcal{S}$ containing the most relevant features for all the individual comparisons:

$$\mathcal{S}_R = \{f_i : f_i \in T \,\&\, m_i > R\} \quad i = 1, 2, ..., 429 - n \tag{1}$$

The parameter n was fixed by cross validation, and the parameter R has six possible values, being $R = 3$ a reasonable compromise between very restrictive and non-existent filter.

The feature set $\mathcal{S}_R$ selection is followed by a PLS-based feature extraction. Following the development presented in [37], we will consider the problem of modelling the relationship between two sets of data using PLS. Let $\mathcal{X} \in IR^N$ and $\mathcal{Y} \in IR^M$ be two multidimensional spaces of variables, PLS models the relationship between them by score vectors. After making n observations of each space, PLS decomposes the matrix $\mathbf{X}(n \times N) \in \mathcal{X}$ of zero-mean variables and matrix $\mathbf{Y}(n \times M) \in \mathcal{Y}$ of zero-mean variables as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{2}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{3}$$

where $\mathbf{X}$ is the training data matrix, $\mathbf{Y}$ is a labels matrix, and $\mathbf{T}$ and $\mathbf{U}$ are matrices $(n \times p)$ formed by the p score vectors extracted (components, latent vectors). The matrix $\mathbf{P}(N \times p)$ and the matrix $\mathbf{Q}(M \times p)$ correspond to the weight matrices. Finally, the $\mathbf{E}(n \times N)$ and the matrix $\mathbf{F}(n \times M)$ are identified as residual values matrices. PLS calculates the vectors of weights $\mathbf{w}, \mathbf{c}$ that form the respective weight matrices mentioned above, as follows:

$$[\mathrm{Cov}(\mathbf{t}, \mathbf{u})]^2 = [\mathrm{Cov}(\mathbf{Xw}, \mathbf{Yc})]^2 = \max_{|\mathbf{x}|=|\mathbf{s}|=1} [\mathrm{Cov}(\mathbf{Xr}, \mathbf{Ys})]^2 \tag{4}$$

where $\mathrm{Cov}(\mathbf{t}, \mathbf{u}) = \mathbf{t}^t\mathbf{u}/n$ denotes the covariance sample of the score vectors $\mathbf{t}$ and $\mathbf{u}$.

This last feature extraction step produces a transformed data matrix D_t. The PLS transformation maximizes the separation between classes in the new space, and can also be used

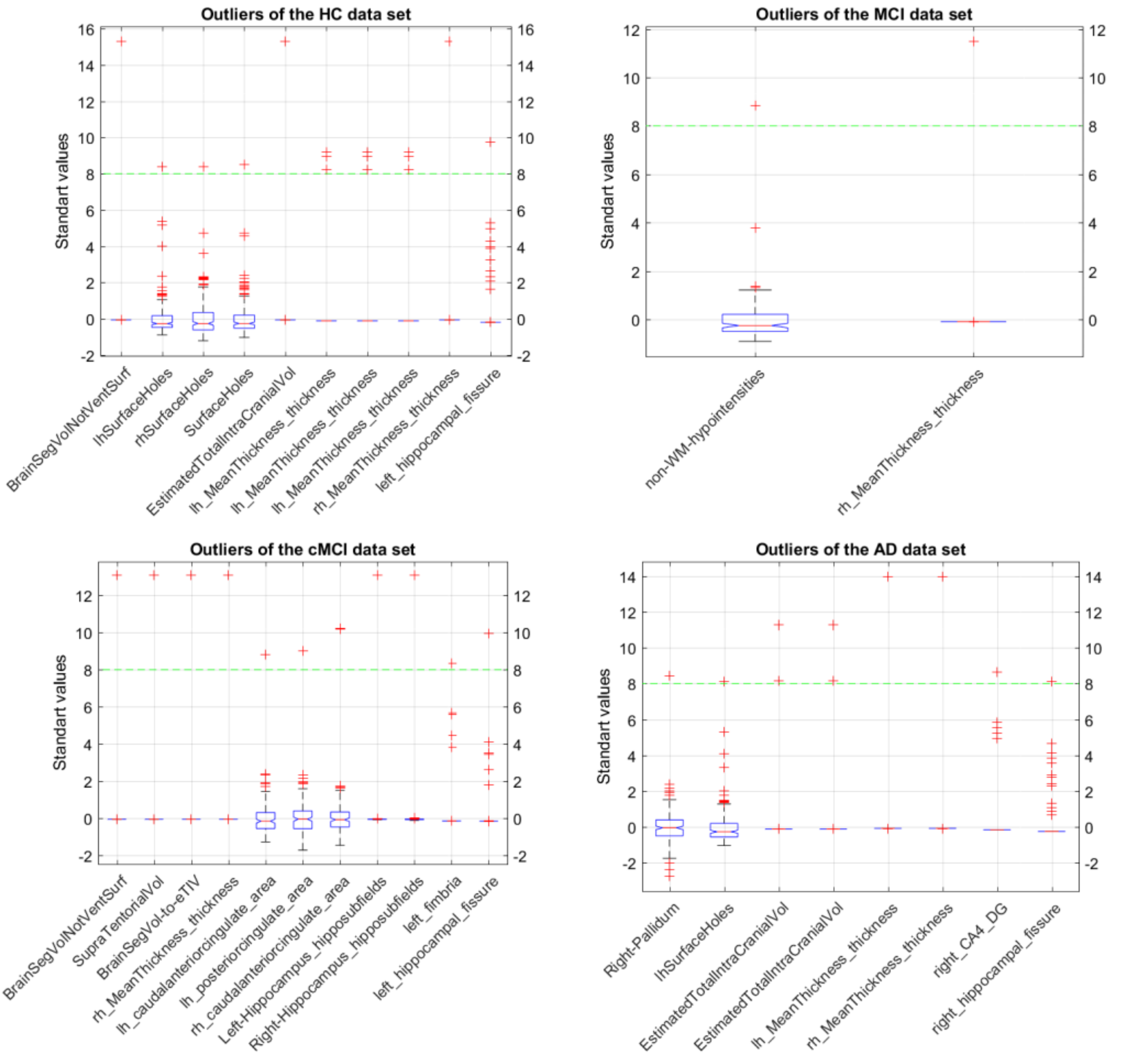


FIGURE 3. Distribution of outliers per class above 8 standard deviations.

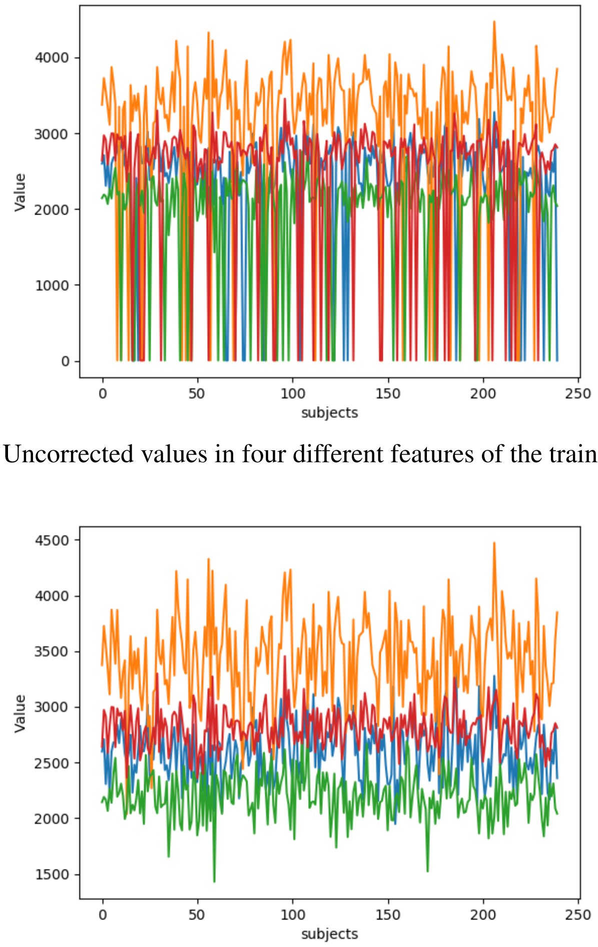
to reduce the feature space dimension by selecting a reduced number of PLS components.

Apart from this last technique, the use of an autoencoder [23] was tested for the same purpose as PLS, in order to use the data generated at the encoder output as low-dimensional data input in the classifier. However, its explanation will not be extended since it is not finally used in the CAD pipeline due to worse results than PLS for this dataset.

3) Classification

A simple solution to the multiclass classification problem is to build binary classifiers and combine them. Classical aggregation techniques of binary classifiers in multiclass problems are usually based on the error correcting output codes (ECOC). Given the multiclass classification problem on N classes, the simplest example is the one-vs-rest model, where the output code is generated by N binary classifiers that exhaust all possible one class versus the $N-1$ rest of the classes classifications. After that, a decoding algorithm is used to assign a final class to each generated output code. Considering the output as a length N codeword, the decoding algorithm can be modelled as a communication problem, where the class information is being transmitted [13].

We consider here the following optimized approach to the



a) Uncorrected values in four different features of the training set

b) Corrected values in four different features of the training set

FIGURE 4. Training data visualization using four features for corrected and uncorrected values.

multiclass classification: a binary classifier is trained on the $K_s = N(N-1)/2$ one-vs-one individual classification tasks using the transformed data matrix D_t, producing a six-bit codeword output for each sample. This output is aggregated to produce a final prediction on the test sample, defined as a decoding process in ternary ECOC algorithms, taking values on the four possible classes: HC, MCI, cMCI and AD. The Hamming decoding [9] is used to map each possible codeword into a single output class as $HD(x, y_i) = \sum_{j=1}^{K_s} 1/2(1 - \text{sign}(x^j y_i^j))$. The justification of this choice lies on the fact that the classes are nested. Concretely, the MCI class is considered as an early stage of AD, although not free of controversy [11], [31]. In any case, the class cMCI is an early stage of AD, and thus can be considered as a subclass of the AD class. For this reason, a ternary ECOC with three possible symbols allows for reduction of the non-relevant class influence in the codeword coding and decoding, and thus managing the possible errors arising from the difference on binary classification accuracies.

Different classifiers are used to perform the individual binary tasks: support vector machine (SVM), including the use of kernel methods [43], nearest neighbours (NN) and decision trees, using different ensemble techniques: bagging [1], boosting [16] and random forest [2]. Moreover, deep learning techniques are also used, such as multilayer perceptron (MLP) and convolution neural network (CNN) [28], for reference and comparison to other published results of the challenge [10], [34]. For evaluation purposes and following the results found in [22], an upper bound can be set on the actual risk based on the re-substitution estimation of the empirical error $|P_{act}(f(x)) - P_{emp}(f(x))| \leq \gamma_{emp}$ for any classifier $f(x)$ at a confidence level η given by:

$$\gamma_{emp} \leq \left(\frac{1}{2l} \ln \frac{2 \sum_{k=0}^{Z-1} \begin{pmatrix} l-1 \\ k \end{pmatrix}}{\eta} \right)^{\frac{1}{2}} \tag{5}$$

III. RESULTS

To estimate the performance of the proposed method, together with the parameter fitting, two strategies were employed in this paper. A 10-fold cross validation strategy and the re-substitution estimation of the actual error on the training set. Once the parameters were optimized, the test set was used to estimate the accuracy, recall and F1-score.

Regarding the parameters of (1) and the number of PLS components, a grid search strategy was employed. Fig. 5 shows the accuracy results on the training set for each pair *(number of PLS components, number of features)*, where the number of features is selected by order from the pool of $\mathcal{S}_R$, affected by the value of n. It can be claimed that a wide range of values around 10 PLS components and 10 selected features produce competitive classification results, whereas a choice of PLS components above 3 and below 20 is also a good compromise independently from the number of features selected. This can be related to the robustness of the method.

The grid search results indicate that a reduced number of selected components, around 23, is optimum for classification results, in combination with a reduced number of PLS components (around 13). The regions that were selected are depicted in Fig. 6, excluding the MMSE score and age, also selected. This optimum set of features was processed as described in section II-B2, providing the training and test data for classification.

Table 4 summarizes the classification results on the training and testing sets for the different classifiers. The results are also summarized for the test set excluding the dummy subjects. SVM outperforms every other classifier, and linear kernel provides slightly better performance than non-linear kernels. However, even the simplest 1-NN classifier provides very competitive results if related to the challenge results. Challenge results are summarized in tables 5 and 6. The competitive performance of every classifier is a sign of the preprocessing importance, revealing that the feature selection

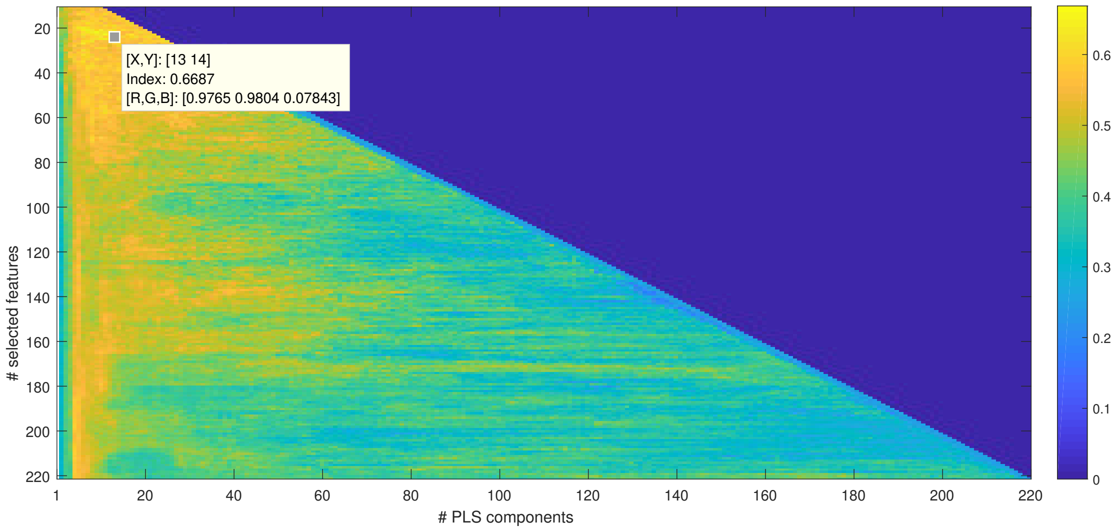


FIGURE 5. Accuracy values for each pair number of PLS components, final number of selected features.

Class	Training Accuracy	recall	F1-score	Test Accuracy	recall	F1-score	Test (without dummies) Accuracy	recall	F1-score
HC	0.43	0.41	0.42	0.43	0.39	0.41	0.80	0.67	0.72
MCI	0.23	0.27	0.25	0.23	0.39	0.29	0.45	0.62	0.52
cMCI	0.50	0.48	0.49	0.46	0.31	0.37	0.62	0.59	0.61
AD	0.75	0.70	0.73	0.38	0.44	0.41	0.80	0.78	0.79
mean	0.48	0.47	0.47	0.38	0.39	0.37	**0.67**	**0.67**	**0.66**

TABLE 3. Performance results by class using the selected and extracted features and the one-vs-one classification scheme with SVM.

Ensemble	Clasifier	Training Accuracy	Recall	F1-score	Test (without dummies) Accuracy	Recall	F1-score
-	SVM lineal	0.48	0.47	0.47	**0.67**	**0.52**	**0.66**
-	SVM RBF	0.47	0.47	0.48	0.67	0.52	0.63
LogitBoost	Decision Tree	0.48	0.44	0.51	0.64	0.46	0.60
Random forest	Decision Tree	0.48	0.44	0.52	0.64	0.51	0.57
AdaBoost	Decision Tree	0.50	0.42	0.47	0.60	0.43	0.52
-	5-NN	0.52	0.43	0.44	0.60	0.44	0.46
-	1-NN	0.52	0.39	0.42	0.58	0.39	0.44
-	3-NN	0.51	0.40	0.39	0.58	0.41	0.40
-	MLP	0.56	0.57	0.56	0.60	0.60	0.59
-	CNN	0.55	0.55	0.54	0.48	0.47	0.48

TABLE 4. Performance results for selected features using different classifiers.

and extraction provide a very relevant set of features. Furthermore, the fact that more complex techniques are the ones with the lowest performance, such as CNN and deep learning algorithms in general, is consistent. It is mainly due to the use of such a low number of subjects and features, since these techniques are especially focused on problems with a large and high-dimensional dataset.

Table 3 summarizes the linear SVM classification results obtained following the proposed aggregation method. F1-score and recall are also reported during the training and test phases. The results are detailed for each class: HC, MCI, cMCI and AD. As expected, AD and cMCI are the classes with highest recognition values, whereas MCI report recognition rates slightly over random classification during training, but improved values on test. Overall, recognition rates are several percentage points over the challenge winner approach, outperforming every proposal of the challenge in the partial ranking (Table 5) and in the final ranking (Table 6).

A study about the control of the family-wise error (FWE) rate in our CAD system was performed based on the re-substitution estimation. A dataset of HCs containing 100 samples, 60 from the training set and 40 from the test set (without dummies), was used. The dataset was randomly divided into two subsets of 50 subjects each throughout 1000 iterations. Then, the re-substitution estimation was evaluated under the null hypothesis that the actual risk was equal to 0.50 (no group difference in the feature set should be true), where the number of PLS components (dimensions) was chosen equal to 1. The re-substitution accuracy obtained was equal to 0.612 with a standard deviation of 0.037. The upper bound associated to this configuration is equal to 0.136, with a significance level $\eta = 0.05$ as shown in equation 5, thus the actual risk is then at most 0.523 ± 0.037. As a conclusion, we cannot reject the null-hypothesis in the test.

On the other hand, If a 10-fold cross validation strategy is tested instead of re-substitution, an accuracy of 0.583 ± 0.06 is obtained with 13 PLS components. Although we can reject the null hypothesis with the current test, it is possible to not rejecting the null hypothesis with a confidence interval of 0.10, a value higher than the usual one of 0.05, but interesting in the neuroimaging field [12].

The last verification of the significance of the selected features is assessed using the actual risk [18]. Following (5), we calculated the upper bound considering the final 13-dimensional dataset in each one-vs-one classification. The sample size in each comparison is 200 that is, by combining

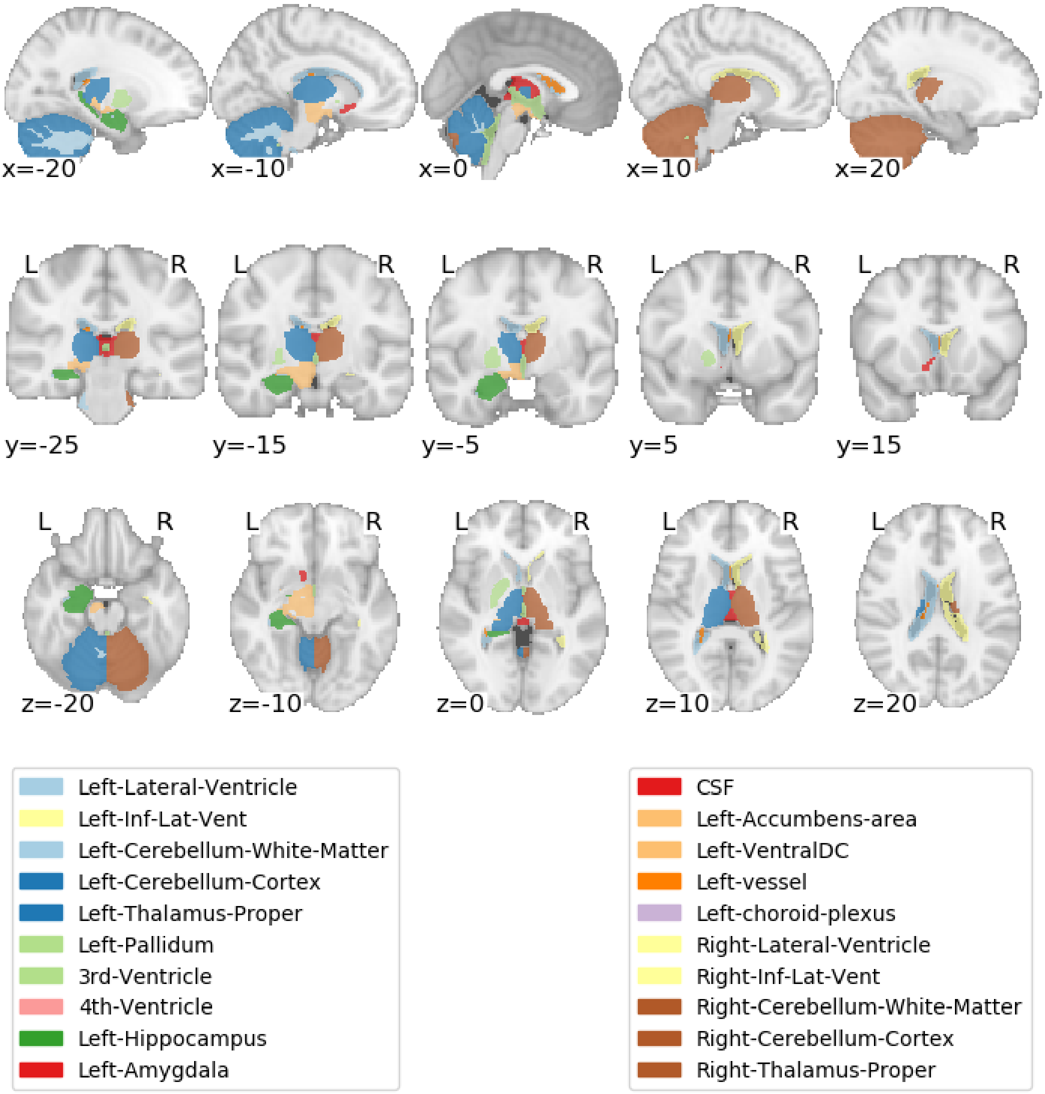


FIGURE 6. Selected regions after one vs. one t-test feature selection.

both training and test (without dummies) sets, a total of 200 subjects are considered in a two-class analysis. With $\eta = 0.05$, the upper bound is equal to 0.343. Table 7 shows empirical errors and actual risks of each one-vs-one comparison according to the upper bound. HCvsMCI and MCIvscMCI actual risks are above 0.50 at the worst case, which means that the selected features cannot be accepted as significant to classify these conditions at the given significance level. Nevertheless, the difficulty of separating these conditions is well-known, thus in general terms a high relevance of the selected features is observed.

An additional study was conducted to detect the relationship between the actual error and dimensionality used in each classifier. Fig.7 shows that the actual risk is never less than 0.50 in the HCvsMCI comparison, whilst in the MCIvscMCI classification the number of PLS components needed for achieving that condition is less or equal to 6. Nevertheless, the use of 6 PLS components would only decrease the overall accuracy down to 60%.

Team	2nd Ranking (Partial)
Stavros Dimitriadis – Dimitris Liparas	0.35999
GRAAL	0.35599
Bari Medical Physics Group	0.34799
BrainE	0.34399
gogogo	0.336
DevinAnnaWilley	0.336
fengxy	0.332
ChaseCowart	0.328
BoyX	0.328
SiPBA-UGR	0.324
Jean-Baptiste SCHIRATTI	0.324
utaphys	0.32
Salvatore C. | Castiglioni I.	0.31199
Sørensen	0.30399
Neuroimage Division – CIFASIS – ARG	0.30399
agrickard	0.30399
JocelynHoye	0.30399
Loris Nanni	0.29199
Webiolab	0.276
Proposed method	**0.38**

TABLE 5. Partial results of the challenge by group, using the whole test set

Team	3rd Ranking (Automatically selected)
Stavros Dimitriadis – Dimitris Liparas	0.61875
SiPBA-UGR	0.5625
Sørensen	0.55
Bari Medical Physics Group	0.55
GRAAL	0.54375
Jean-Baptiste SCHIRATTI	0.54375
Neuroimage Division – CIFASIS – ARG	0.54375
Salvatore C. | Castiglioni I.	0.5375
Loris Nanni	0.53125
BrainE	0.525
utaphys	0.525
gogogo	0.525
ChaseCowart	0.51875
agrickard	0.50625
fengxy	0.5
JocelynHoye	0.5
DevinAnnaWilley	0.46875
BoyX	0.4625
Webiolab	0.2125
Proposed method	**0.67**

TABLE 6. Final results of the challenge by group, using the test set without dummies.

Classifier	Empirical error	Upper bound	Actual risk
HC vs MCI	0.320	0.343	**0.663**
HC vs cMCI	0.135	0.343	0.478
HC vs AD	0	0.343	0.343
MCI vs cMCI	0.240	0.343	**0.583**
MCI vs AD	0.065	0.343	0.408
cMCI vs AD	0.130	0.343	0.473

TABLE 7. Actual risk associated to each one-vs-one classifier using the selected and extracted features (13) and SVM by resubstitution (200 samples).

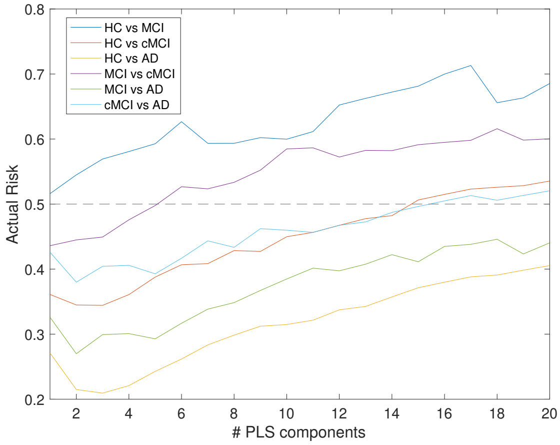


FIGURE 7. Estimates of actual risk in each one-vs-one classifier for several dimensions using a sample size of 200 subjects.

IV. DISCUSSION

The results presented reveal the importance of feature selection and extraction in the CAD pipeline of this challenge problem. Concretely, the default sorting of the 429 cortical thickness values provided by the organizers of the challenge seem to contain already important information for classification. The best results are obtained by removing some of the original set of default sorted features, by applying the proposed minimum filter. After the feature selection by means of the filter in (1), it is relevant to emphasize that there is not an equilibrium between right and left hemisphere regions. Specifically, there is a dominance of left-sided hemisphere regions, which is coherent with recent findings in CAD diagnosis of AD [29]. If compared with other competent CAD proposals of the challenge, such as the winner Dimitriadis-Liparas (DL) proposal [10], there is a significant overlap with the feature extraction results. The DL approach also results in a left-hemisphere regions predominance. Therefore, a successful feature selection and extraction method is critical for optimal performance.

The use of t-tests as sorting criteria for filtering features and the upper-bound tests for assessing the feature relevance are justified on the basis of the Kolmogorov-Smirnov (KS) and the upper bound tests [18]. Moreover, the KS test quantifies the departure of the empirical distribution function of the features from a cumulative distribution function of a particular statistical distribution. In this case, the assumption underlying a t-test implies that the feature values follow a normal distribution, which is an acceptable assumption in the light of the results of the KS test presented in Fig. 8. It is important to stress that direct comparisons between different t_i values at different tests are never performed, but the values of t_i are used for feature sorting. In addition, we employed a novel approach [18] for testing relevance in a set of features based on a data-driven approach (agnostic or free-parameter model). The latter is based on the re-substitution error estimate and the theoretical upper-bound of the empirical errors that provide a confidence interval for performing hypothesis testing.

The present methodology can be applied in other multiclass classification problems, in which there is a hierarchy and overlap between classes. Furthermore, the computation of the final selected algorithm is fast, which is an advantage over tested deep learning techniques, which require a longer processing time associated with network training.

Concerning the limitations on the present work, the preprocessing of outliers reveal a high redundancy on the original data. Therefore, the pre-selection of brain regions for the challenge, and the extraction of cortical thickness affects the maximum achievable performance in several ways. Firstly, the limited number of training samples makes statistical estimations prone to bias, a widely known-problem in medical imaging [4], [22]. The cross validation technique used in this work for performance estimations can be considered as "pessimistic" [27], and therefore some mismatches between training fitting and final test estimations can be expected, limiting the capabilities of the system for reaching its highest performance at test level.

Even though the proposed CAD was evaluated using the test set labels, which were not available during the challenge competition, the robustness of the method would have led to the best competition results with just a few submissions. Table 4 and Fig. 5 illustrate how the method is robust against small variations on the optimal parameters and classifier choice, providing with accuracy values over 60% for a wide range of combinations.

V. CONCLUSION

Using the available data for *A Machine learning neuroimaging challenge for automated diagnosis of Mild Cognitive Impairment*, we developed a post-competition method for multiclass classification. The method is based on a one vs. one approach for feature selection, PLS feature extraction and classification. The presented methodology is capable of identifying the most relevant features for a multiclass classification by a sorting-and-filtering method, and is evaluated using different parameters and classifiers. The results are robust against variations of parameters and classifiers, and they outperform all the proposals submitted to the challenge by more than 5 percentage points in accuracy. The method is also coherent with recent findings in CAD of AD, and can be applied to other multiclass classification problems.

ACKNOWLEDGMENT

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. Hoff-mann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & 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.

REFERENCES

[1] Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (Aug 1996), https://doi.org/10.1007/BF00058655

[2] Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (Oct 2001), https://doi.org/10.1023/A:1010933404324

[3] Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M.E., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J.R., Garrett, C., Bastos-Leite, A.J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., Álvarez Meza, A.M., Dolph, C.V., Iftekharuddin, K.M., Eskildsen, S.F., Coupé, P., Fonov, V.S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K.R., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G.M., Stoyanov, Z.V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J.C., Niessen, W.J., Klein, S.: Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: The CADDementia challenge. NeuroImage 111, 562–579 (May 2015), http://www.sciencedirect.com/science/article/pii/S1053811915000737

[4] Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R.: Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14, 365 (Apr 2013), https://doi.org/10.1038/nrn3475

[5] Castiglioni, I., Salvatore, C., Ramírez, J., Górriz, J.M.: Machine-learning neuroimaging challenge for automated diagnosis of mild cognitive impairment: Lessons learnt. Journal of Neuroscience Methods 302, 10–13 (May 2018), http://www.sciencedirect.com/science/article/pii/S0165027017304375

[6] Chih-Wei Hsu, Chih-Jen Lin: A comparison of methods for multiclass support vector machines. IEEE Transactions on Neural Networks 13(2), 415–425 (Mar 2002)

[7] Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O.: Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. NeuroImage 56(2), 766–781 (May 2011), http://www.sciencedirect.com/science/article/pii/S1053811910008578

[8] Davatzikos, C.: Machine learning in neuroimaging: Progress and challenges. NeuroImage 197, 652–656 (Aug 2019), http://www.sciencedirect.com/science/article/pii/S1053811918319621

[9] Dietterich, T.G., Bakiri, G.: Solving Multiclass Learning Problems via Error-Correcting Output Codes. Journal of Artificial Intelligence Research 2, 263–286 (1994), https://www.jair.org/index.php/jair/article/view/10127

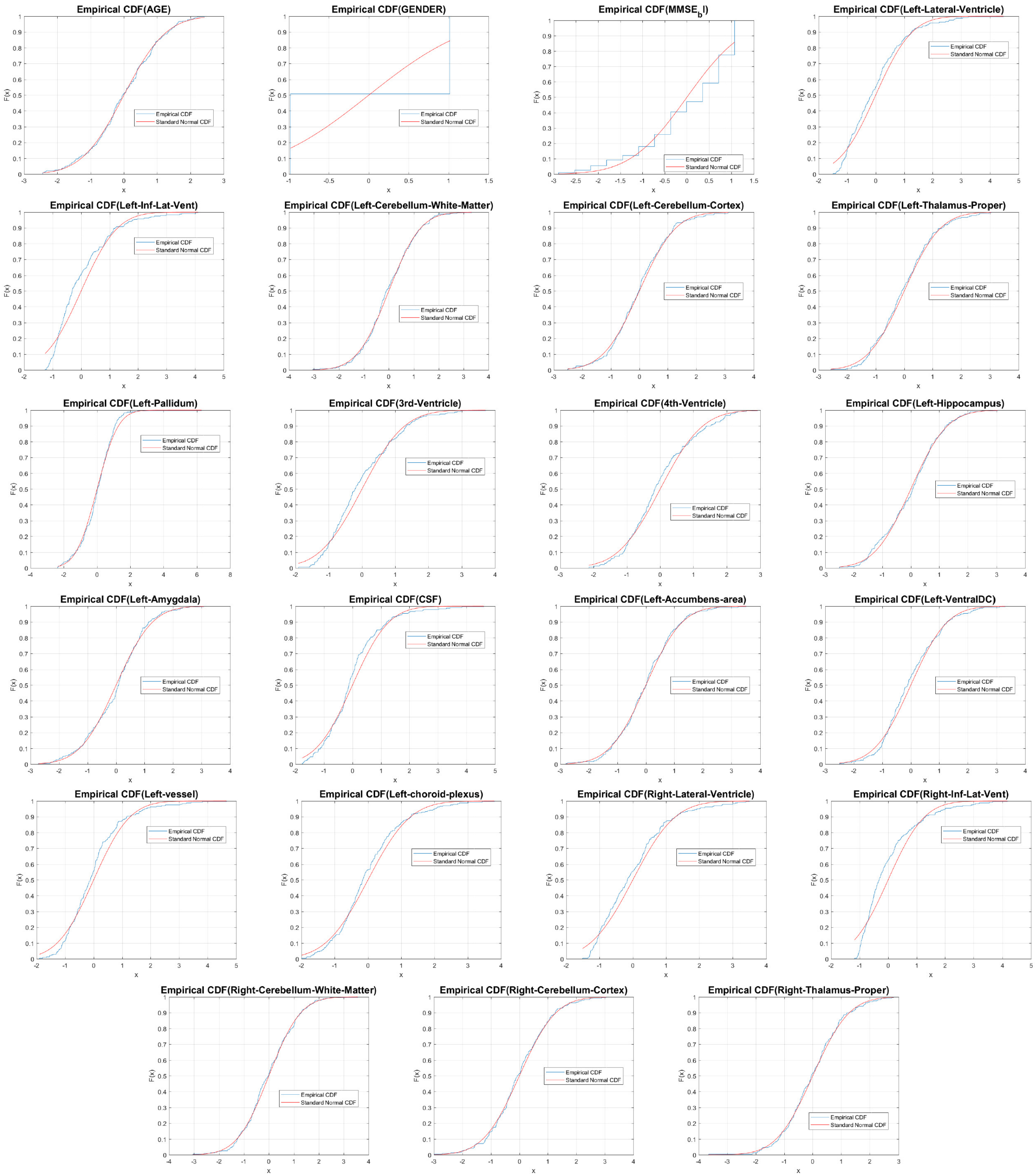


FIGURE 8. Results of the KS test for selected features.

[10] Dimitriadis, S.I., Liparas, D., Tsolaki, M.N.: Random forest feature selection, fusion and ensemble strategy: Combining multiple morphological MRI measures to discriminate among healhy elderly, MCI, cMCI and alzheimer's disease patients: From the alzheimer's disease neuroimaging initiative (ADNI) database. Journal of Neuroscience Methods 302, 14–23 (May 2018), http://www.sciencedirect.com/science/article/pii/S0165027017304272

[11] Dubois, B., Feldman, H.H., Jacova, C., DeKosky, S.T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway,

S., Stern, Y., Visser, P.J., Scheltens, P.: Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. The Lancet Neurology 6(8), 734–746 (Aug 2007), http://www.sciencedirect.com/science/article/pii/S1474442207701783

[12] Eklund, A., Nichols, T.E., Knutsson, H.: Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proceedings of the National Academy of Sciences 113(28), 7900–7905 (jun 2016)

[13] Escalera, S., Pujol, O., Radeva, P.: On the Decoding Process in Ternary Error-Correcting Output Codes. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(1), 120–134 (Jan 2010)

[14] Fischl, B.: FreeSurfer. NeuroImage 62(2), 774–781 (Aug 2012), http://www.sciencedirect.com/science/article/pii/S1053811912000389

[15] Fischl, B., Dale, A.M.: Measuring the thickness of the human cerebral cortex from magnetic resonance images. Proceedings of the National Academy of Sciences 97(20), 11050–11055 (Sep 2000), https://www.pnas.org/content/97/20/11050

[16] Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm (1996)

[17] Fritsch, V., Varoquaux, G., Thyreau, B., Poline, J.B., Thirion, B.: Detecting outliers in high-dimensional neuroimaging datasets with robust covariance estimators. Medical Image Analysis 16(7), 1359–1370 (Oct 2012), http://www.sciencedirect.com/science/article/pii/S1361841512000564

[18] Gorriz, J.M., Group, S., neuroscience, C.: Statistical agnostic mapping: a framework in neuroimaging based on concentration inequalities. arXiv (1912.12274) (2019)

[19] Graf, A.B., Borer, S.: Normalization in Support Vector Machines. In: Radig, B., Florczyk, S. (eds.) Pattern Recognition. pp. 277–282. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2001)

[20] Gray, K.R., Aljabar, P., Heckemann, R.A., Hammers, A., Rueckert, D.: Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. NeuroImage 65, 167–175 (Jan 2013), http://www.sciencedirect.com/science/article/pii/S1053811912009834

[21] Górriz, J.M., Ramírez, J., Suckling, J., Illán, I.A., Ortiz, A., Martinez-Murcia, F.J., Segovia, F., Salas-González, D., Wang, S.: Case-based statistical learning: A non-parametric implementation with a conditional-error rate svm. IEEE Access 5, 11468–11478 (2017)

[22] Górriz, J.M., Ramirez, J., Suckling, J.: On the computation of distribution-free performance bounds: Application to small sample sizes in neuroimaging. Pattern Recognition 93, 1–13 (Sep 2019), http://www.sciencedirect.com/science/article/pii/S0031320319301402

[23] Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (jul 2006)

[24] Illan, I.A., Górriz, J.M., Ramírez, J., Meyer-Base, A.: Spatial component analysis of MRI data for Alzheimer's disease diagnosis: a Bayesian network approach. Frontiers in Computational Neuroscience 8, 156 (2014), http://journal.frontiersin.org/Journal/10.3389/fncom.2014.00156/abstract

[25] Khedher, L., Ramírez, J., Górriz, J.M., Brahim, A., Segovia, F.: Early diagnosis of Alzheimer's disease based on partial least squares, principal component analysis and support vector machine using segmented MRI images. Neurocomputing 151, 139–150 (Mar 2015), http://www.sciencedirect.com/science/article/pii/S0925231214013137

[26] Khedher, L., Illán, I.A., Górriz, J.M., Ramírez, J., Brahim, A., Meyer-Baese, A.: Independent Component Analysis-Support Vector Machine-Based Computer-Aided Diagnosis System for Alzheimer's with Visual Support. International Journal of Neural Systems 27(03), 1650050 (Jul 2016), http://www.worldscientific.com/doi/abs/10.1142/S0129065716500507

[27] Kohavi, R.: A Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2. pp. 1137–1143. IJCAI'95, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1995), http://dl.acm.org/citation.cfm?id=1643031.1643047, event-place: Montreal, Quebec, Canada

[28] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (may 2015)

[29] Li, H., Habes, M., Wolk, D.A., Fan, Y.: A deep learning model for early prediction of Alzheimer's disease dementia based on hippocampal magnetic resonance imaging data. Alzheimer's & Dementia 15(8), 1059–1070 (Aug 2019), http://www.sciencedirect.com/science/article/pii/S155252601930086X

[30] Martinez-Murcia, F.J., Górriz, J.M., Ramírez, J., Ortiz, A., for the Alzheimer's Disease Neuroimaging Initiative: A spherical brain mapping of mr images for the detection of alzheimer's disease. Current Alzheimer Research 13(5), 575–588 (2016), http://www.eurekaselect.com/node/140378/article

[31] Modrego, P.J.: Predictors of conversion to dementia of probable Alzheimer type in patients with mild cognitive impairment. Current Alzheimer Research 3(2), 161–170 (Apr 2006)

[32] Ortiz, A., Munilla, J., Górriz, J.M., Ramírez, J.: Ensembles of Deep Learning Architectures for the Early Diagnosis of the Alzheimer's Disease. International Journal of Neural Systems 26(07), 1650025 (Apr 2016), https://www.worldscientific.com/doi/abs/10.1142/S0129065716500258

[33] Ramirez, J., Gorriz, J.M., Chaves, R., Lopez, M., Salas-Gonzalez, D., Alvarez, I., Segovia, F.: Spect image classification using random forests. Electronics Letters 45(12), 604–605 (2009)

[34] Ramírez, J., Górriz, J.M., Ortiz, A., Martínez-Murcia, F.J., Segovia, F., Salas-Gonzalez, D., Castillo-Barnes, D., Illán, I.A., Puntonet, C.G.: Ensemble of random forests One vs. Rest classifiers for MCI and AD prediction using ANOVA cortical and subcortical feature selection and partial least squares. Journal of Neuroscience Methods 302, 47–57 (May 2018), http://www.sciencedirect.com/science/article/pii/S0165027017304223

[35] Ramírez, J., Górriz, J., Segovia, F., Chaves, R., Salas-Gonzalez, D., López, M., Álvarez, I., Padilla, P.: Computer aided diagnosis system for the Alzheimer's disease based on partial least squares and random forest SPECT image classification. Neuroscience Letters 472(2), 99–103 (Mar 2010), http://www.sciencedirect.com/science/article/B6T0G-4Y95RSJ-9/2/64ec495d7589612445949dc79114ed80

[36] Rojas, A., Górriz, J., Ramírez, J., Illán, I., Martínez-Murcia, F., Ortiz, A., Río], M.G., Moreno-Caballero, M.: Application of empirical mode decomposition (emd) on datscan spect images to explore parkinson disease. Expert Systems with Applications 40(7), 2756 – 2766 (2013), http://www.sciencedirect.com/science/article/pii/S0957417412012274

[37] Rosipal, R., Krämer, N.: Overview and Recent Advances in Partial Least Squares. In: Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J. (eds.) Subspace, Latent Structure and Feature Selection. pp. 34–51. Lecture Notes in Computer Science, Springer Berlin Heidelberg (2006)

[38] Sarica, A., Cerasa, A., Quattrone, A., Calhoun, V.: Editorial on special issue: Machine learning on MCI. Journal of Neuroscience Methods 302, 1–2 (May 2018), http://www.sciencedirect.com/science/article/pii/S0165027018300827

[39] Segovia, F., Górriz, J.M., Ramírez, J., Salas-González, D., Álvarez, I.: Early diagnosis of Alzheimer's disease based on Partial Least Squares and Support Vector Machine. Expert Systems with Applications 40(2), 677–683 (Feb 2013), http://www.sciencedirect.com/science/article/pii/S095741741200927X

[40] Segovia, F., Górriz, J.M., Ramírez, J., Álvarez, I., Jiménez-Hoyuela, J.M., Ortega, S.J.: Improved Parkinsonism diagnosis using a partial least squares based approach. Medical Physics 39(7), 4395–4403 (Jun 2012), http://scitation.aip.org/content/aapm/journal/medphys/39/7/10.1118/1.4730289

[41] Segovia, F., Górriz, J., Ramírez, J., Salas-Gonzalez, D., Álvarez, I., López, M., Chaves, R.: A comparative study of feature extraction methods for the diagnosis of Alzheimer's disease using the ADNI database. Neurocomputing 75(1), 64–71 (2012), http://www.sciencedirect.com/science/article/pii/S0925231211003961

[42] Teipel, S.J., Kurth, J., Krause, B., Grothe, M.J.: The relative importance of imaging markers for the prediction of Alzheimer's disease dementia in mild cognitive impairment — Beyond classical regression. NeuroImage: Clinical 8, 583–593 (Jan 2015), http://www.sciencedirect.com/science/article/pii/S2213158215000984

[43] Vapnik, V.N.: Statistical Learning Theory. John Wiley and Sons, Inc., New York (1998)

[44] Vos, F.d., Schouten, T.M., Hafkemeijer, A., Dopper, E.G.P., Swieten, J.C.v., Rooij, M.d., Grond, J.v.d., Rombouts, S.A.R.B.: Combining multiple anatomical MRI measures improves Alzheimer's disease classification. Human Brain Mapping 37(5), 1920–1929 (2016), https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23147

[45] Wang, S.H., et al: Unilateral sensorineural hearing loss identification based on double-density dual-tree complex wavelet transform and multinomial logistic regression. Integrated Computer-Aided Engineering 26, 411–426 (209)

[46] Wu, T.F., Lin, C.J., Weng, R.C.: Probability Estimates for Multi-class Classification by Pairwise Coupling. Journal of Machine Learning Research 5(Aug), 975–1005 (2004), http://www.jmlr.org/papers/v5/wu04a.html?907d3908

[47] Zhang, Y., et al: Multivariate approach for alzheimer's disease detection using stationary wavelet entropy and predator-prey particle swarm optimization. Journal of Alzheimer Disease 1, 855–869 (Jan 2018)