Predicting the Conversion from Mild Cognitive Impairment (MCI) to Alzheimer´s Disease using Evolution Patterns

Andreia Liliana Duarte Fernandes Ferreira
Instituto Superior Técnico, Lisboa, Portugal
December 2014

Abstract: Declines in cognitive functions, together with other evidences of neurological degeneration, become increasingly likely as healthy people age [1]. Alzheimer´s Disease (AD) is a neurodegenerative disease characterized by progressive deterioration of cognitive function and is the most common cause of dementia in elderly people. Mild Cognitive Impairment is considered a prodromal state that represents the transitional period between normal ageing and dementia. As such is regarded with special attention since it represents higher risk to evolve to dementia. Thus, the definition of this clinical entity is fundamental to the timely administration of pharmaceutics and therapeutic interventions, improving patient's quality of life.

This thesis intends to predict the evolution of MCI patients to AD considering two approaches: the first where all patients are assumed to evolve similarly and the second where patient profiles are considered. Time windows for two to five years are used for prediction. Initially, we used supervised learning methods, using feature selection to effectively decrease the dimensionality of the problem. Then, standard clustering algorithms were applied with the purpose of studying the potential existence of MCI subtypes. The patients were also divided according to their state of depression, based on clinical information.

The results demonstrated the importance of considering longer time interval to predict conversion of MCI patients to AD and that the grouping of patients according to their depressive symptoms influences positively the prognosis results. The clustering analyses validated the importance to study MCI subgroups considering the different characteristics of this clinical entity in the prediction models.

Keywords: Alzheimer´s Disease, MCI, Temporal Window, Prognosis, Classification, Clustering.

1. Introduction

Alzheimer´s Disease (AD) is the most common cause of dementia in elderly people and its specific cause is still unknown [2]. This disease progresses gradually, beginning with early signs and subtle behavioural changes, followed by memory loss, impaired judgment and lower ability to participate in daily activities. In a last phase, the disease evolves to the incapacity of the patient to understand language or even to speak, besides the disability to control his body [3]. The accurate assessment of AD is still a challenge, due to the several symptoms shared with other diseases, related with cognitive decline, and the little existence of available longitudinal studies to draw conclusions about [4, 5].

Mild Cognitive Impairment (MCI) is a prodromal state that represents transitional period between normal changes in cognitive ageing and dementia. Furthermore, these patients represent higher risk of evolving to dementia. The construct of MCI proposes to identify individuals at an earlier point in the cognitive decline, such that if

therapeutic interventions become available, clinicians can intervene at this juncture [1,6].

In this context, the aim of this work is to predict the evolution of MCI patients to AD considering two approaches: first where all patients are assumed to evolve similarly and the second where patient profiles are considered. Time windows from two to five years are used for prediction. We first computed a baseline model to predict conversion of MCI individuals to AD through supervised learning techniques without using explicitly different MCI groups. Then, MCI subgroups were obtained, first by considering important characteristics of the patients and then through unsupervised learning methods. These groups were used to train models to predict the conversion to AD and were expected to outperform the baseline model. Demographic data and longitudinal data from large number of neuropsychological assessments from each MCI individual, provided by the Dementia Group at Instituto de Medicina Molecular (IMM), were used.

2. Background

2.1 Alzheimer's Disease and Mild Cognitive Impairment

Alzheimer´s Disease is a neurodegenerative disease clinically characterized by a progressive dementia. The accurate assessment of AD is still a challenge but numerous efforts have been made in the past decades in order to capture a full spectrum of the disease and apply it to research protocols and clinical trials, directed at early stages of the disease [6, 7].

The MCI stage has been described as prodromal stage of AD and interposes between cognitive changes of healthy aging people, which interfere with the distinction from MCI individuals encountered by normal individuals as they age, to those with very early dementia. The definition of this concept had the purpose of identifying individuals at an early stage in the cognitive decline, such that clinicians could intervene at this point [8]. Most of the studies developed have a maximal follow-up of the patients of three years, presenting unsatisfactory results [1,9] and it would be of extreme importance to increase the temporal windows analysed. The separation of patients according to their state of depression is starting to assume a relevant role in the prognosis prediction of MCI patients. Generally, the test used to study this condition is the Geriatric Depression Scale (GDS) that is a self-report assessment used specially to identify depression [10]. In order to increase the accuracy of the MCI classification, many studies had in consideration alternate clinical subtypes of MCI to reduce heterogeneity in the study groups and to match the timely administration of pharmacologic, tailored for specific targets and populations [9, 10].

In opposition to Magnetic Resonance Imaging (MRI) volumetric studies [6] and Positron Emission Tomography (PET) scans [7], neuropsychological tests are relatively inexpensive and noninvasive to the patient, being widely used in the clinical assessment of AD [7, 10]. These tests were established by medical doctors and typically include orientation, new-learning/memory, intelligence, language, visuoperception, executive function, person's psychological, personal, interpersonal and wider contextual circumstances [11].

2.2 Data Mining

Data Mining constitutes an iterative process of data gathering to find regular patterns and extract meaningful information to develop inductive learning models. In medical diagnosis, learning models are an irreplaceable tool for the early detection of diseases using clinical test results [12, 13]. In order to extract useful information different

steps have to be analysed carefully as the data preprocessing, which includes cleaning, integration, selection and transformation, the data mining, where methods are applied to extract relevant patterns from the data, the results evaluation, in order to analysed the discovered patterns through interestingness measures and the knowledge presentation, where the conclusions are presented to the user [12].

2.2.1 Unsupervised learning methods

In unsupervised learning algorithms the class label attribute and the number of classes to learn in advance are unknown. Clustering methods represent a class of models of unsupervised learning, which principle is to maximize the similarity of objects within one class [14]. Regarding the formation MCI subgroups through unsupervised leaning, EM and K-Means were the clustering algorithms applied.

2.2.2 Supervised learning methods

Classification can be divided into two phases, the learning phase and the validation phase. In the learning phase, it is given a dataset containing predictive attributes and a categorical target attribute, called class or label. In our case, the instances were the several observations of each patient, the set of predictive attributes was extracted from the neuropsychological data and the class label was the patient state (evolution or not evolution to AD in a given temporal window). In the validation phase, the learned classifier is applied to an independent set of instances to estimate its predictive performance [13]. The validation set is used to avoid generating overoptimistic and overfitted models.

In this analysis, five classifiers were used, Naïve Bayes, Gaussian SVM, Polynomial SVM, K-Nearest Neighbour and C4.5 Decision Tree, all implemented in WEKA.

2.2.2.1 Learning from imbalanced datasets

Most classical machine learning algorithms perform poorly on imbalanced data, since they are biased toward the majority class, due to the assumption of balanced class distributions or equal misclassifications errors [15]. Medical applications are a common source of imbalanced datasets, where the disease cases are much rarer than the healthy cases. In order to obtain balanced or optimal proportions, the method Synthetic Minority Over-Sampling Technique (SMOTE) was applied [15, 16]. The algorithm takes each minority class sample and introduces artificial minority instances, based on the feature space similarities. Depending upon the amount of over-sampling required, neighbours from the k-nearest neighbours are randomly chosen and the minority class is over-sampled along the line segments, joining any or all of the k minority class nearest neighbours.

2.3 Result Evaluation

One of the methods to evaluate models is the k-fold Cross-Validation, dividing the data into *k* subsets of equal size. In the present thesis, the performance measures were estimated based on the 5-folds cross-validation method, due to the number of the Cognitive Complaints Cohort (CCC) dataset.

The results were analysed from multiple goal perspectives as accuracy, sensitivity, specificity and AUC. Although, these metrics provide a simple way of describing a classifier´s performance, they can be deceiving and are highly sensitive to changes in the data [15]. In order to obtained conclusive evaluations of performance, F-measure is also calculated since it is the ultimate measure of performance of a classifier, not depending on disease prevalence [15, 16].

Accuracy: Accuracy is the ratio between the number of correctly classified instances and the

total number of instances described as $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$.

Sensitivity: Sensitivity measures how many examples of positive class were labelled correctly, represented as the ratio between the correctly classified positive instances over the number of real positive instances. It is described as $Sensitivity = \frac{TP}{TP+FN}$.

Specificity: Specificity is the proportion of negative instances that were correctly classified, described as $Specificity = \frac{TN}{TN+FP}$.

F-measure: F-measure metric is a trade-off between sensitivity and specificity. It is defined as $F - measure\ weighted = \left(\frac{TP+FN}{TP+FN+FP+TN}\right)\left(\frac{2TP}{2TP+FP+FN}\right) + \left(\frac{FP+TN}{TP+FN+FP+TN}\right)\left(\frac{2TN}{2TN+FP+FN}\right)$. This metric is motivated in this thesis since in many studies this is the ultimate measure of performance of a classifier, not depending on disease prevalence.

Receiver Operating Characteristic (ROC) curves: ROC curves are useful to express the information content of a sequence of confusion matrices and provides a visual representation of the trade-offs between the benefits, reflected by the true positive rate (TPR), and costs, reflected by the false positive rate (FPR), of a given classifier model.

3. Experimental Methodology

3.1 Description of the Data

The dataset analysed in this work is composed of 1827 instances consisting of individual evaluations of 990 patients, categorized as Normal, Pre-MCI, MCI, AD and Without Diagnosis, considering the patient's cognitive condition in their follow-up at IMM. The dataset has 192 attributes, both categorical and numerical, consisting in the results of the neuropsychological evaluation procedures. Each patient has several evaluations (considered as different instances) where he/she was diagnosed with different stages of MCI or as AD, during the follow-up.

3.2 Dataset Preprocessing

The first step was the exclusion of instances where the patients, in the different phases of the follow-up, are labelled as normal, as pre-MCI and as without diagnosis, resulting in the removal of 383 observations. Later, the patients with only one evaluation corresponding to 384 (near 48.6% of the remaining patients) were eliminated leading to a significant reduction of number of patients analysed. The remaining patients constitute MCI and AD patients with two or more observations (Table 1). Since we want to predict conversion from MCI to AD, we only consider patients whose diagnosis in the first evaluation is MCI. In the following observations the patient can remain MCI or can evolve to AD.

	MCI	AD
Patients in the first evaluation (%)	407(100%)	0(0%)
Observations (%)	878(82.8%)	182(17.2%)

Table 1 Composition of the analysed patients.

3.3 Creating learning examples

The dataset is splitting in 75% for training and 25% for validating. This stratified partition had in consideration the distribution of variables such the number of evaluations, the age, the sex, the schooling years and the class, keeping them constant in the training and validation sets.

In order to predict the conversion of MCI patients, two approaches were used. The first, which constitutes the baseline of this work, looks at the first and last evaluations of the patient in the dataset to see if the patient will ever convert from MCI to AD. A new set of learning examples is then created, as evolution (Evol) or no evolution (noEvol) instances. The second approach is related to the temporal window prognosis problem:

predicting if a patients converts to AD in a given temporal window The choice of the temporal window had in consideration the instances distribution between classes (Evol/noEvol) and the medical relevance at the problem, defined by consulting the medical partners of the NEUROCLINOMICS project. In both approaches the Evol class is considered the positive class.

Instances with undefined class were excluded and non-informative attributes were not used in this analysis, yielding 182 attributes.

3.4 Feature Selection

The Feature Selection improves the classification accuracy substantially or equivalently and increases the speed and the accuracy of the learning process [13, 14]. We used correlation-based feature subset selection and a greedy search, which evaluates the value of the attribute subsets considering their individual predictive ability and the redundancy among them. After the feature selection there were still instances with high percentages of missing values so, the next step was to exclude instances with more than 50% missing values.

3.5 Handling missing values

From the remaining missing values, each classifier has an internal way of dealing with it. The Naïve Bayes classifier excludes the missing values from the calculations and the SVMs, the KNN and the C4.5 Decision Tree classifier use the internal way of median/ mode imputation, in case of numerical/categorical values, respectively. Tables 2 and 3 show the composition of the data after the preprocessing steps referred above.

	noEvol	Evol
FL Evaluation	66 (63.5%)	38 (36.5%)
2Y	61 (66.3%)	31 (33.7%)
3Y	40 (49.4%)	41 (50.6%)
4Y	24 (33.3%)	48 (66.7%)
5Y	16 (23.9%)	51 (76.1%)

Table 2 Train dataset details after preprocessing steps.

	noEvol	Evol
FL Evaluation	171(58.8%)	120 (41.2%)
2Y	161 (68.8%)	73 (31.2%)
3Y	112 (53.1%)	99 (46.9%)
4Y	66 (35.7%)	119 (64.3%)
5Y	42 (23.3%)	138 (76.7%)

Table 3 Validation dataset details after preprocessing steps.

3.6 Classification

3.6.1 Training Model

In order to determine the best SMOTE percentage and the best classifier parameters, it was necessary to cross all tested SMOTE percentages with all tested parameters sets in a grid search of a parameterized model. The classification model used in this thesis was the automated model created by the previous work of Lemos et. al [17] with slight changes, in order to optimize the grid search process. The metric used to compare models for each SMOTE and parameter set was the F-Measure. This choice relates to the fact that the other metrics are highly sensitive to imbalanced data and the F-measure metric is a trade-off between the sensitivity and specificity, not depending on disease prevalence. Both the parameters sets and the SMOTE percentages are tested with 5-fold cross-validation, using 11 different values for each parameter combination.

The grid search was performed in all datasets and classifier models. After the search, five best triples are determined {*Classifier, Parameters, SMOTE percentage*}, one for each classifier and were tested in 30 repetitions, using different seeds in the 5-fold cross validation for each repetition. In the present study, a paired t-test using the 30

repetitions was used and the t-test was only applied if the ANOVA test with 95% confidence level confirmed the existence of a significant difference [18]. The feature selection is accomplished outside the cross-validation.

3.6.2 Validation model

The aim of the validation set is to evaluate the final models created during the training phase, by analysing the behaviour of the trained models in a real-world simulation, since the model has never been in contact with any instance of the validation patients. The preprocessing steps were applied in an equivalent way to training and validation sets, except for the attribute subset selection. The attributes obtained in the training set after applying the filter were extrapolated directly to the validation sets.

Figure 1 presents the data stream in the grid search.

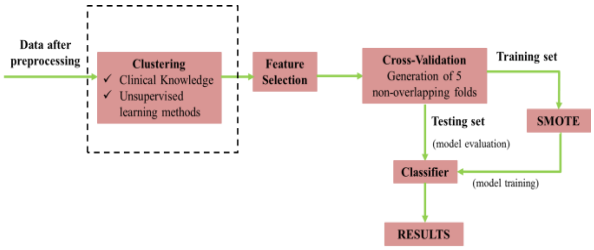

Figure 1 Data flow used in the parameter grid search for finding the classifiers parameters [48].

4. Predict conversion from MCI to AD

In a first phase, the classification model explained already was applied to the First and Last Evaluation Prognosis Problem, with the purpose of constructing a baseline comparison for further analyses. Afterwards the different temporal windows containing all patients was analysed regarding the same principles, expecting to outperform the baseline. The first problem to address is the class imbalanced found in the datasets used for prognosis. After comparing the results with and without contemplating the SMOTE algorithm, we decided that the application of the SMOTE algorithm is only consider if the proportion of the majority class is above 70%. With proportions higher than that, the use of the algorithm may not be beneficial and may lead to the overfitting of the data and bias the results. The best results were obtained with three and four temporal window with feature selection and in this section only the respective results will be presented.

4.1 Three Years Temporal Window

The train dataset size relative to three years temporal window is described in Table 4. The SMOTE algorithm was not applied due the similar proportions of classes in this dataset.

3Years	noEvol	Evol
Size (%)	112 (53.1%)	99 (46.9%)

Table 4 Size of the train dataset after applying feature selection using three years temporal window.

SVM Poly is the best model in the reduced dataset and the DT4.5 got the worst results. The SVM Poly has an accuracy of 82%, a sensitivity of 78%, a specificity of 86%, an AUC of 0.82 and F-measure of 0.82. The trade-off between sensitivity and specificity is preferable than in other temporal windows, reflected by the higher values of F-measure. From Figure 2 we note that the choice of feature selection has a great effect on the final outcome.

In the validation set (Table 5), the SVM Poly achieved the Area under the ROC curve near to 0.7 and performed with values of F-measure close to 0.7 and accuracy of 69.1%, obtaining an overall good performance in this set.

3Years	noEvol	Evol
Size (%)	40 (49.4%)	41 (50.6%)

Table 5 Size of the validation dataset after applying Feature Selection using three years temporal window.

4.2 Four Years Temporal Window

The train dataset size relative to five years temporal is described in Table 6.

4Years	noEvol	Evol
Size (%)	66 (35.7%)	119 (64.3%)

Table 6 Size of the train dataset after applying Feature Selection using four years temporal window

Naïve Bayes, SVM Poly and SVM RBF were the best models, with no statistical difference between them and the DT4.5 model got the worst results. Naïve Bayes had an accuracy of 82%, a sensitivity of 82%, a specificity of 81%, an AUC of 0.88 and a F-measure of 0.82. The increase of the temporal window leads to better results in the cross-validation without the application of feature selection. The use of correlated features even improves such good results (Figure 2).

In the validation set (Table 7), the Naïve Bayes achieved an accuracy of 72%, sensitivity of 83%, specificity of 50%, the Area under the ROC curve near to 0.8 and performed with values of F-measure of 0.71. The specificity has low values in contrast with the high values of sensitivity, probably due to the lower number of examples of noEvol.

4Years	noEvol	Evol
Size (%)	24 (33.3%)	48 (66.7%)

Table 7 Size of the validation dataset using four years temporal.

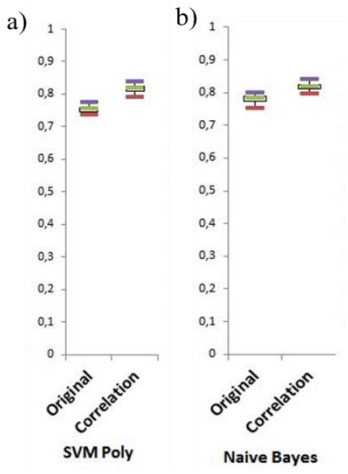

Figure 2 Train results of the F-measure metric using: a) three years temporal window; b) four years temporal window.

5. Predicting conversion from MCI to AD based on different MCI characteristics

The knowledge acquired so far that the separation of patients with different characteristics would improve the classification of the MCI either motivated us to develop a new approach predicting models trained with different MCI groups, based on clinical criteria or results of unsupervised learning methods. Ideally, this approach should outperform the classification approach where all MCI are assumed to evolve similarly.

5.1 Prognosis prediction based on clinical criteria: depressed/not depressed

Previously in this work, we described the attribute that measures the state of depression of a patient named as GDS, Geriatric Depression Scale [10]. With the purpose of studying the influence of the depressive symptoms in classification, a new analysis was performed based on the GDS attribute. For this study, a short form of the self-report instrument was used, comprising score values between 0 and 14. Looking to Table 8, it is noticeable the occurrence of more than 50% of missing values. Because of that those instances were not taken in consideration.

	GDS score 0-4	GDS score 5-14	GDS value missing
FL	72(23.8%)	69(22.8%)	162(53.5%)
2Y	72(25.3%)	63(22.1%)	150(52.6%)
3Y	58(22.8%)	56(22.1%)	140(55.1%)
4Y	52(22.8%)	48(21.3%)	128(56.1%)
5Y	46(36.7%)	45(5.6%)	124(57.7%)

Table 8 Composition of instances considering the GDS attribute after preprocessing

In the training phase two approaches were analysed in order to determine if there is a higher capability to predict conversion based on the characteristics of the patients instead of assuming the same profile in the entire set of patients. The datasets used to create these models were separated in three, one considers only the instances with this attribute defined ($\mathfrak{D}_{all}$), other includes only instances with values from 0 to 4 corresponding to the not depressed patients (denominated as $\mathfrak{D}_{0-4}$) and another comprises instances with values from 5 to 14 corresponding to the depressed patients (denominated as $\mathfrak{D}_{5-14}$). The first approach consists on training the models from the entire dataset $\mathfrak{D}_{all}$, to understand how these general models classify the patients according to their state of depression. The best classifier under the cross-validation evaluation is then applied to the restrictive datasets ($\mathfrak{D}_{0-4}$ and $\mathfrak{D}_{5-14}$) and the respective result is generated for each group of patients. In the second approach, the dataset $\mathfrak{D}_{all}$ is divided in the two restrictive datasets and for each one of them is determined the best classifier under the cross-validation evaluation and from the particular models obtained from each group of patients is determined the output result.

The datasets were analysed after determining the set of attributes that allows the achievement of more accurate classifiers. The elimination of instances with more than 50% of missing values after the feature selection was not performed, since such analyses are based on comparisons between datasets, where is important to keep the same number of instances, and the removals are minor.

In the next section only the results related with the three temporal window will be presented.

5.1.1 Three Years Temporal Window

In $\mathfrak{D}_{all}$, the classifier Naïve Bayes is the best model and the DT4.5 got the worst results. The Naïve Bayes model has a similar predictive capability when considers depressed or not depressed patients. Thus, its performance is class independent, with 82% of accuracy, 79% of sensitivity, 87% of specificity and 0.82 of F-measure. In $\mathfrak{D}_{0-4}$, the classifier Naïve Bayes is the best model and the DT4.5 model got the worst results. In $\mathfrak{D}_{5-14}$, the classifier Naïve Bayes is the best model and the classifier *k*NN presents the lower results, with no statistical difference with DT4.5 model. The models $\mathcal{M}_{0-4}$ and $\mathcal{M}_{0-4}$ considering different characteristics of the patients have higher predictive capability with F-measure of 0.881 and 0.876 respectively, resulting in an accuracy of 84%, a sensitivity of 77%, a specificity of 91% and F-measure of 0.84. As regards to the results, the separation of patients concerning their state of depression seems to lead to a better learning of the models and produces valuable results.

In the validation set, in contradiction of what we obtained in the cross-validation results, the best results are achieved with the model trained with all patients.

5.2 Prognosis prediction based on Patient Similarities

The clustering analysis cannot be performed from all attributes together [1], so the attributes were separated in the ones that are not corrected for age and school (called as A attributes) from the ones with a z-score correspondent (called as B and B_Z attributes). One set of features analysed was labelled as A+B attributes and the other was labelled as A+B_Z attributes. It was only presented only the ones regarding the attributes corrected for age and school (A+B_Z attributes).

This section describes the results obtained by means of clustering algorithms, in order to investigate the potential existence of MCI groups. In order to test the possible existence of three MCI groups, this study was performed for different number of clusters, in particular 2, 3 and 4.

In general, the results achieved clusters constituted by a disorganized blend of patients from both classes in every temporal window, demonstrating the difficult task of defining MCI groups. Due to the fact that the number of patients per cluster was reduced, invalidating the learning of the models, it was established to consider two clusters in the classification, in every temporal window.

In order to determine the differences between models, firstly the entire set of features from the dataset $\mathfrak{D}_{all}$ is reduced to the features from the baseline analysis. Then, the dataset comprising all the instances along with the set of attributes selected is divided in the two clusters and the clusters are analysed in a cross-validation evaluation. In every temporal window, the values of assessment metrics were always lower when the learning was performed considering the clusters obtained, comparing with the results from the baseline analysis.

6. Conclusions and Future Work

Our results showed that the temporal window approach leads to better discriminative results. Both in cross-validation and in the validation results, as long as the temporal window increases, the higher prediction capability of the models was noticed. Concerning the cross-validation results, in the five temporal window, even though we obtained the highest values of F-measure and accuracy, the values of specificity were the lowest, due to the reduced number of noEvol examples in this window. The results obtained in the three and four years temporal window are more reliable, being more suitable to predict conversion with higher predictive capability.

A more reliable approach would be to learn an ensemble of models, increasing the predictive performance over any of the constituent learning algorithms. Among these techniques the more prominent schemes are the bagging and the boosting.

Considering the clinical information, the patients were divided according to their state of depression, dictated by the GDS attribute. The prevalence of more than 50% of missing values regarding this attribute was a drawback. The results presented corroborate the hypothesis that the separation of the patients according to their depressive symptoms influences positively the classification and that this data mining approach has potential as a possible strategy for the prognosis prediction of conversion of groups of MCI patients to AD. In future studies, it is expected that the integration of more patients with GDS information will allow for a higher prediction capability separating the patients with different characteristics.

In consideration of the prognosis prediction based on patient similarities, the results did not outperform the classification approach where all MCI are assumed to evolve similarly. Notwithstanding, the diverse experiments performed seem to enhance the importance of the study of MCI subgroups considering their putative differences. Other similarity clustering algorithms, should be applied to corroborate such assumptions. Other strategy to try in future analyses is the feature selection for unsupervised learning [14].

The results demonstrated that future studies should give priority to differentiate MCI subgroups using clinical knowledge.

7. References

[1] Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T. and Jin, J.S. (2011) Identification of Conversion from Mild Cognitive Impairment to Alzheimer's Disease Using Multivariate Predictors, PlosOne, 6(7), e21896.

[2] Seixas, F.L., Zadrozny, B., Laks, J., Conci, A. and Saade, D.C.M, (2014) A Bayesian network decision model for supporting the diagnosis of dementia, Alzheimer's disease and mild cognitive impairment, Elsevier, Computers in Biology and Medicine, 51, 140–158.

[3] Sloane, P. D., Zimmerman, S., Suchindran, C., Reed, P., Wang, L., Boustani, M., and Sudha, S. (2002) The Public Health Impact of Alzheimer's Disease, 2000–2050: Potential Implication of Treatment Advances, Annu. Rev. Public Health, 23, 213–31.

[4] McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D. and Stadlan, E.M. (1984) Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease, Neurology, 34, 939-944.

[5] Petersen, R.C., Caracciolo, B., Brayne, C., Gauthier, S., Jelic, V. and Fratiglioni, L., (2014) Mild cognitive impairment: a concept in evolution, KeySymposium, Journal of International Medicine, 275, 214–228.

[6] Prestia, A., Carolia, A., Herholz, K., Reimand, E., Chend, K., Jaguste, W. J., Frisonia and G. B. Frisoni, (2013) Diagnostic accuracy of markers for prodromal Alzheimer's disease in independent clinical series, Alzheimer's & Dementia, 1-10.

[7] Liu, F., Wee, C., Chen, H. and Shen, D. (2014) Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification, NeuroImage, 84, 466-475.

[8] Peterson, R. C., (2009) Mild Cognitive Impairment, Ten years Later, Arch Neurol, 66, 1447-1455.

[9] Roberts, R. O., Knopman, D. S., Mielke, M. M., Cha,R.H., Pankratz,V.S., Christianson, T.J.H., Geda,Y.E., Boeve,B.F., Ivnik, R.J.,Tangalos, E.G., Rocca,W.A. and Petersen, R.C. (2013) Higher risk of progression to dementia in MCI cases who convert to normal , Neurology, 82, 1-9.

[10] Silva, D., Guerreiro, M., Maroco, J., Santana, I., Rodrigues, A., Marques, J. B. and Mendonça, A. (2012) Comparison of Four Verbal Memory Tests for the Diagnosis and Predictive Value of Mild Cognitive Impairment, Dementia and Geriatric Cognitive Disorders Extra, 2, 120-131.

[11] Mathuranath, P.S., Nestor, P. J., Berrios, G. E., Rakowicz, W., and Hodges, J. R. (2000) A brief cognitive test battery to differentiate Alzheimer's disease and frontotemporal dementia, Neurology, 55, 1613-1620.

[12] He, H. and Garcia, E.A, (2009), Learning from Imbalanced Data, IEEE Transactions on knowledge and Data Engineering, 21(9), 1263-1284.

[13] Singhal, S. and Jena, M., A Study on WEKA Tool for Data Preprocessing, Classification and Clustering, (2013), International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, Volume-2, Issue-6, 2278-3075

[14] Vercellis, C., (2009), Business Intelligence: Data Mining and Optimization for Decision Making, John Wiley & Sons, Ltd.

[15] Rokach, L., Maimon, O., Data Mining and Knowledge Discovery Handbook, (2010) Springer Science+Business Media, 2nd Edition.

[16] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., (2002), SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research, 16, 321–357.

[17] Lemos, L. J. M., A data mining approach to predict conversion from mild cognitive impairment to Alzheimer's Disease (2012), Master degree, Instituto Superior Técnico.

[18] Demsar, J., (2006) Statistical Comparisons of Classifiers over Multiple Data Sets, Journal of Machine Learning Research, 7, 1–30.