Predicting conversion from Mild Cognitive Impairment to Alzheimer's Disease using a Temporal Mining approach

Rita Pissarra Levy

Under supervision of Sara Madeira INESC-ID/Técnico Lisboa, Universidade de Lisboa

December 14, 2016

Abstract

Alzheimer's Disease(AD) is neurodegenerative disease and the most common form of dementia. AD prevalence increases each year and there is no efficient and universal treatment. Data mining methods comprise pattern recognition from medical databases consisting on neuropsychological data that provide information to the medical doctor, facilitating the prognosis from a Mild Cognitive Impairment (MCI) state to the conversion to AD. Previous studies use independent classifiers to this prognosis problem, ignoring any temporal information present in the dataset. This thesis focus on contributing to the early detection of the conversion to AD prognosis with two approaches: preprocessing techniques that transform the original dataset in order to capture this temporal information and using a temporal classifier able to deal with this temporal information.

The first approach consists on feature extraction where temporal features that calculate the progression, define temporal patterns and statistically sum up the progression between two time-points are derived from the original values. This prepocessing worflow is followed by a classification task carried out using the Naïve Bayes classifier. The second one relies on Hidden Markov Models (HMM) to process internally the temporal information on the original dataset. First approach shows promising results, where many models outperformed the baseline one. However, HMM were not able to provide a good alternative as the environment was not able to successfully process the original data.

Overall, this exploratory work shows a future path to better understand the underlying AD mechanisms and to improve the AD prognosis with Data Mining methodologies.

Keywords: Alzheimer's Disease, Mild Cognitive Impairment, Temporal Mining, Data Mining

1 Introduction

Dementia is a wide term for several progressive diseases affecting memory, other cognitive abilities and behavior, that interferes significantly with a person's ability to engage in everyday activities of daily living [1]. In 2015, dementia affected 47 million people worldwide (or roughly 5% of the world's elderly population), a figure that is estimated to increase to 75 million in 2030 and 132 million by 2050 [2]. Alzheimer's disease (AD) is the most common form of dementia and may contribute with up to 60 to 70% of cases. Individuals with AD have trouble remembering recent events, become confused and forgetful, often repeating questions and getting lost in familiar places. As the disease evolves, the ability to remember past events is lost while disorientation and violent mood swings increase. Although dementia mainly affects elderly people, it is not a normal part of aging and it is one of the major causes of disability and dependency among elderly people worldwide [3]. Mild Cognitive Impairment (MCI) was once considered an initial stage of AD, as the symptomatic profile is similar but less severe. However, MCI may originate from a variety of different etiologies and pathologies, since there are cases where MCI subjects do not progress to AD, and other where there is a reversion back to health. These cases may suggest that the clinical symptoms of MCI can occur due to causes other than underlying AD pathology [4,5]. Nonetheless, the risk of dementia due to AD in MCI subjects is higher when compared to cognitively normal subjects. The annual incidence rate of healthy subjects that develop AD is 1% to 2%, while the conversion rate from MCI to AD is reported to be approximately 10% to 15% per year [6]. Therefore, is important to know if a subject presenting signals and symptoms from the MCI state will evolve or not to AD state.

The progression of a disease is usually accompanied by a progression of the symptoms. Hence, instead of only looking to test results of one single medical observation, with this work we are attempting to find the best way to combine all the medical appointments of a patient, in order to answer two important clinical questions: Does the outcome of previous neuropsychological tests and their temporal relation help predict the progression from MCI to AD? And if so, what is the best way to look at them? To find an optimal solution, the question was tackled in three distinct but related problems: (i) Find the best way to evaluate the temporal information in various neuropsychological tests of the same patient, by defining a temporal mining model; (ii) Find the best temporal window with new temporal datasets; (iii) Find the best way to model the progression of MCI; The main goal of this work is to explore the usefulness of temporal information in the prediction of the prognosis of Alzheimer's Disease, by means of (i) Developing a framework of preprocessing techniques that lead to better results when analyzing datasets with temporal data; (ii) Introducing the creation of temporal features into the Alzheimer's Disease prognosis problem; (iii) Exploring these same temporal features to find the ones which yield better results; (iv) Introducing the use of Hidden Markov Models to the Alzheimer's Disease prognosis problem.

2 Background

2.1 Alzheimer's Disease and Mild Cognitive Impairment

Alzheimer's Disease is a chronic neurodegenerative disorder type of dementia, defined as a premature aging of the brain. Alzheimer's Disease affects each person in different ways, but there are symptoms that are consistent in everyone affected: (i) Temporal and spatial disorientation; (ii) Amnesic type of memory loss; (iii) Deterioration of language; (iv) Experiencing behavior changes including violent and aggressive ones; (v) Loss of ability to care for oneself [7].

Mild Cognitive Impairment is a common disorder, characterized as cognitive decline, greater than that expected for an individual's age and education level, but that does not interfere notably with activities of daily life [8]. Although a consensual definition to describe MCI is yet to be achieved in the research and clinical community, according to the European Consortium on Alzheimer's Disease [8], the MCI diagnose is based on the following criteria: (i) Cognitive complaints coming from patients or their caregivers; (ii) The reporting of a decline in cognitive functioning relative to previous abilities during the past year by the patient or caregiver; (iii) Presence of cognitive impairment (1.5 standard deviations below the reference mean) in at least one neuropsychological test; (iv) Absence of major repercussions on daily life (the patient may report difficulties concerning complex day-to-day activities).

Neuropsychological tests are a tool used to assess the state of a patient's cognitive impairment, by using a set of questions and tasks explicitly designed to test specific cognitive domains. Different cognitive domains and functions are affected in dementia and the symptomatic profile of patients' reports the impairments in these different domains with different aggravations [1]. The ability to acquire and remember new information, changes in personality or behavior, visuospatial abilities, language functions and reasoning and handling of complex tasks are assessed according to different cognitive domains, including: Executive functions, Memory, Attention, Language, Conceptual and Abstract Thinking, Visual and spatial perception and Orientation. The outcome of a test depends on the patient's willingness to do it, the judgment of the Medical Doctor (MD) and, when it relies on information given by the caregiver, it depends on the caregiver's perception of the patient's ability to have an independent life.

3 Methods

The dataset used throughout this thesis consists of information from the Cognitive Complaints Cohort (CCC), a database containing information from elderly and non-demented patients with cognitive complaints,from 1999 to 2015. These patients were referred for neuropsychological evaluation at at three institutions,the Laboratory of Language Studies at Santa Maria Hospital, and Memoclínica (a Memory Clinic), both in Lisbon, and the Neurology Department, University Hospital, in Coimbra. In the original database we consider the main unit of information to be an observation, corresponding to an instance with a time-stamp. An observation is a medical appointment where the state of a patient is assessed and categorized into a class. In the original dataset, the observation can be classified as Normal, pre-MCI, MCI and AD. There are 616 patients that resolve into a total of 1604 instances. There are 102 attributes, 59 being various tasks of neuropsychological tests ans scales applied and 29 the correspondent Z-Score value. Three demographic attributes and 11 auxiliary ones are present in the database and all the attributes are numeric. In terms of data structure, a patient is defined as a set of observations that belong to a same person and it is this relation that allows us to have temporal information of the state of the disease (or the state of the patient).

This database was the foundation of every dataset present to the classifier. On this database were applied different preprocessing techniques in order to obtain a good dataset that resolves into a good model. The present database is build with a generous amount of information with all of the neuropsychological results and the consequent observation diagnosis, which can be one of four categories: Normal, pre-MCI, MCI, and AD. However, we are interested in predicting the patient's prognosis (the probability of the patients to evolve to a certain outcome) instead of the diagnosis. Due to this fact, a data transformation and reclassification process are necessary, considering the independent dataset. Hence, the creation of learning examples where s patient's information is compiled into an instance in order to obtain one that can be categorized into the prognostic classes. The prognostic task is set within a time frame that we define as a time window. A patient is considered to be progressing to AD if the progression happens inside the time window, rather than considering if a patient is going to progress to AD or not in an undefined time frame.

The cleaning process consists in a series of methods that alter or delete instances that are not suitable to the data mining problem here addressed. Some attributes were also deleted as they contained discriminatory information about the class of the patients. Normal and pre-MCI instances were deleted as they were not part of this work scope.

To study the clinical history of a patient it is necessary to have temporal information, that is, more than one observations for the same patient. The creation of learning examples makes it possible to assess the evolution of the tests results of the same patient from one observation to another. A minimum of three observations of the same patient are necessary. The first two observations (as they are form two different time points) make us capable of studding the temporal progression. The third observation sets the class of the learning example. Therefore, we have deleted all patient records that have less than three observations. After this cleaning task, the number of patients in the database is 179, with a total of 672 instances, showing a great decrease in the number of patients and instances comparing to the original dataset.

To consider a patient stable, we have to define a time period during which the conversion to AD does not occur. A stable MCI (sMCI) learning example consists of two MCI observations within the time window defined and a third observation outside this time window, that vouches for the stable state of the patient. On the other hand, to state that a learning example is converter MCI (cMCI), the conversion must happen inside the time window. So we need two MCI observations as well as a third one diagnosed AD inside this window, assuring that the conversion occurred in the time period considered.

As mentioned before, within the scope if this work we consider a patient as a group of medical observations of one individual in different time points. The creation of learning examples reconstructs a patient as a single instance instead of an observation as was in the original data. This transformation implies a loss of useful information as a patient is not only two observations. Additionally, the temporal relation of this attribute is also lost, because this learning examples are fed to a non temporal classifier that considers all attributes to be independent. Trying to save as much information as possible we proceeded to a temporal processing stage creating attributes that held values relating two different time points.

Different temporal features were created. The first dataset created did not contain temporal features and was constructed by simply add the information of two different time points as unrelated information.

Next, we considered what we named as a Progression Features . This progression feature was a simple arithmetic operations between each attribute considering different time points, as all attributes are numeric. This attribute is normalized with the time (in days) between the two time points.

We related two attributes defining a simple temporal pattern that specified if an attributes increased (code letter U for Up), decreased (code letter D for Down) or stayed the same(code letter S for Stable). This creates a nominal attribute for every two numeric ones.

The last temporal feature created was a statistic measure of all the values of an attribute for a patient. This means was calculated using three different formulas: Arithmetic mean, Geometric mean and Harmonic mean as well as Variance.

Hidden Markov Models (HMM) classifier supports temporal information. Despite the different proprocessing workflow, HMM datasets were constructed under the same train of thought that independent datasets, as is still need learning examples comprising patients with temporal information, from at least two observations.

Feature Selection was applied to almost every dataset to tackle the overfitting of the model to the high variance dataset and Synthetic Minority Over-sampling Technique (SMOTE) was used to overcame class imbalance.

The Naive Bayes classifier was used to classify and build models for the independent datasets and HMM classifier to the relational datasets built. Accuracy, Area Under the receiver operating characteristic (ROC) curve, Specificity and Senstivity metric were drawn for every dataset.

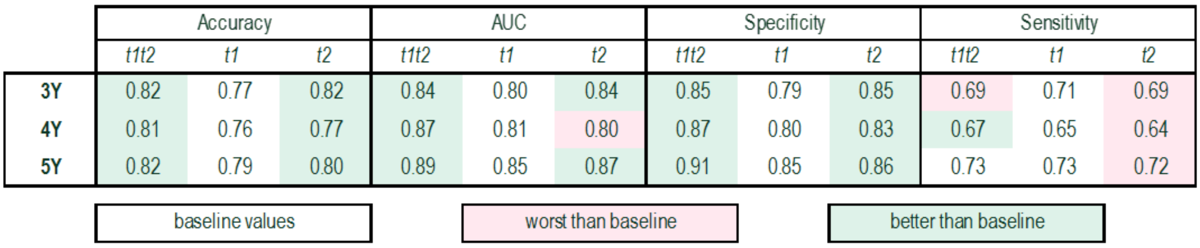
4 Results

The first step taken in order to perceive if information from more than one observation was in fact useful was creating learning examples with two observation values of different timepoints for each test: the value for the baseline and the value for a second observation close to the end of the time window.

Exploring values presented in Table 1, we can see better values in almost every time window and evaluation metric for t_2 and t_1t_2 than t_1. Specificity values for the t_2 dataset are worse than the baselines for every time window. Comparing t_1 with t_2 it is possible to see that only t_2 gives better results, probably because it is an observation closer to the observation that defines the progression state. However, t_2 model do not outperform the t_1t_2 model. It is interesting to see that t_2 and t_1t_2 have the same evaluation metric values in the 3-year window, yet, when the time window starts to increase the t_2 values start to drop while t_1t_2 actually increase. This is the start point for the sub sequential datasets as they all combined information from exactly those two timepoints in various ways, aiming to explore ways to combine the information from the two timepoints in order to obtained better models. Baseline values are now considered to be t_1t_2.

Progression Feature are attributes calculated from two timepoints within the time window considered for that database, with simple arithmetic operations: PR_+, PR_-, $PR_\times$ and $PR_\div$ and normalized in terms of days between these two timepoints. These results are present in Table 2.

Accuracy and AUC results for all progression Features show us the overall results of progression Features attributes are better than the values for the baseline datasets. Progression Feature models surpasses the baseline model for the 3-year window and the 4-year window. The 5-year window progression feature model does not outperform the baseline model yet it is not significantly worst. Moreover, if instead of comparing the progression Feature models with the baseline yielding two time points, we considered the results for just one time point, for the 5-year window, the progression Feature model outputs better results in almost every operation and evaluation metric (with exception of accuracy and sensitivity values for PR_+ and specificity values for PR_-, $PR_\div$). Comparing the different progression feature models there is one that clearly surpasses the others. $PR_\times$ has basically every value better that the baseline (with exception of sensitivity values for the 4 and 5-year window) where other do not behave as well as this. However, $PR_\times$ does not yield the better value comparing with other in the same time window and for the same metric.



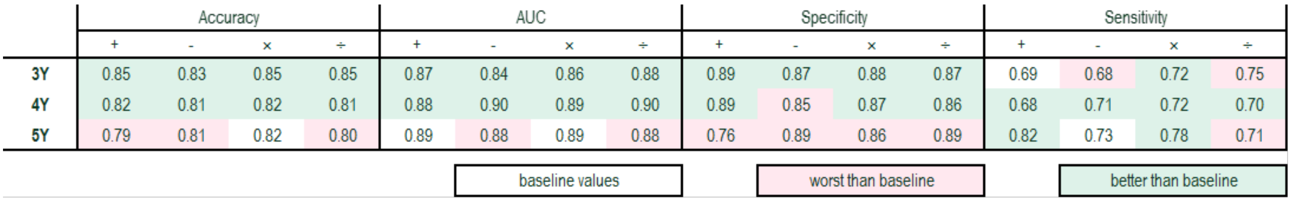
	Accuracy			AUC			Specificity			Sensitivity		
	t1t2	*t1*	*t2*	*t1t2*	*t1*	*t2*	*t1t2*	*t1*	*t2*	*t1t2*	*t1*	*t2*
3Y	0.82	0.77	0.82	0.84	0.80	0.84	0.85	0.79	0.85	0.69	0.71	0.69
4Y	0.81	0.76	0.77	0.87	0.81	0.80	0.87	0.80	0.83	0.67	0.65	0.64
5Y	0.82	0.79	0.80	0.89	0.85	0.87	0.91	0.85	0.86	0.73	0.73	0.72

baseline values

worst than baseline

better than baseline

Figure 1: Cross Validation results for datasets without temporal features.



	Accuracy				AUC				Specificity				Sensitivity			
	+	-	×	÷	+	-	×	÷	+	-	×	÷	+	-	×	÷
3Y	0.85	0.83	0.85	0.85	0.87	0.84	0.86	0.88	0.89	0.87	0.88	0.87	0.69	0.68	0.72	0.75
4Y	0.82	0.81	0.82	0.81	0.88	0.90	0.89	0.90	0.89	0.85	0.87	0.86	0.68	0.71	0.72	0.70
5Y	0.79	0.81	0.82	0.80	0.89	0.88	0.89	0.88	0.76	0.89	0.86	0.89	0.82	0.73	0.78	0.71

baseline values

worst than baseline

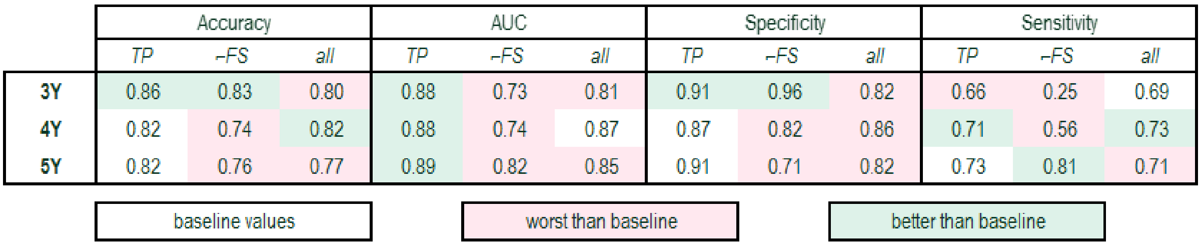
better than baseline

Figure 2: Cross Validation results for the datasets with progression features.

As mention before, $PR_{\times}$ end up accentuating small differences, and considering values that are very similar and this can be an advantage, especially in datasets comprising small time frames as these differences are not accentuated. Hence, this progression feature can contribute to the dataset with useful information and make rise to good classification models. This reasoning is validated with good sensitivity values present for this model, standing out from the overall bad sensitivity results for other models.

As the time window increases, so does the spacing between the two timepoints considered and the progression feature that derived from this information loses predictive power. While a big accentuated difference in two close timepoints might be predicting a fast progression, probably towards AD, the same difference on a broader time frame losses importance and does not stand out. This fact makes the negative instances less distinguishable and this is validated by the poor specificity values.

The temporal pattern results are present in Table 3. Overall, the temporal pattern results do not show the improvement we were looking for when adding the temporal pattern attribute. Of all these three models, the simple Temporal pattern for two time points holds better results. Exploring this match a little further, feature selection results show us that the features chosen by the algorithm are basically the same, adding one temporal pattern attribute for the Graphomotor Initiative's Z score. In fact, feature selection seldom chooses temporal pattern attributes. This can be since the temporal patter represents the behavior between two time points but these two timepoints can have different time intervals and the temporal pattern is not able to distinguish between a slow of fast increase/decrease. Considering two learning examples from the same dataset the first one can have the two timepoints two years apart the second one can have the time points 4 years apart. An attribute that increases within 2 or 4 years can have very distinct mean-



	Accuracy			AUC			Specificity			Sensitivity		
	TP	*¬FS*	*all*	*TP*	*¬FS*	*all*	*TP*	*¬FS*	*all*	*TP*	*¬FS*	*all*
3Y	🟩 0.86	🟩 0.83	🟥 0.80	🟩 0.88	🟥 0.73	🟥 0.81	🟩 0.91	🟩 0.96	🟥 0.82	🟥 0.66	🟥 0.25	⬜ 0.69
4Y	⬜ 0.82	🟥 0.74	🟩 0.82	🟩 0.88	🟥 0.74	⬜ 0.87	⬜ 0.87	🟥 0.82	🟥 0.86	🟩 0.71	🟥 0.56	🟩 0.73
5Y	⬜ 0.82	🟥 0.76	🟥 0.77	🟩 0.89	🟥 0.82	🟥 0.85	⬜ 0.91	🟥 0.71	🟥 0.82	⬜ 0.73	🟩 0.81	🟥 0.71

⬜ baseline values

🟥 worst than baseline

🟩 better than baseline

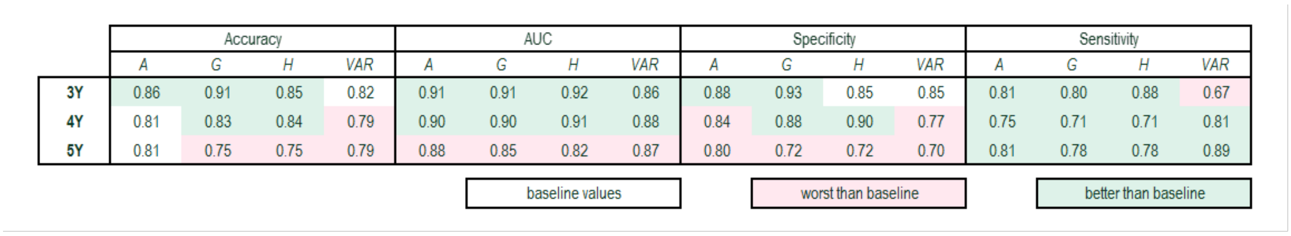
Figure 3: Validation results for the datasets with temporal pattern attribute.

ings and the gradient of this increase can be very different and still be categorized into the same code letter. When the dataset was created using all time points available within a learning examples in order to capture more information of the temporal behavior of an attribute it is created a bigger set of possible nominal values to the same number of instances and the classifier was not able to create rules for every nominal value, leading to worst results. In this dataset, two learning examples with two and three timepoints both with simultaneous end time points, maintaining always the same value for an attribute will have different values within the temporal pattern attribute (S and SS respectively) and the classifier perceives as two completely different values. Concluding, temporal pattern can yield a good representation of a time series yet not when this time series does not have regular temporal intervals.

Lastly, here are presented the results concerning the Statistics-Based Summarization Datasets.

Every mean summarizes the global characteristics of a temporal sequence into a single value. If we consider a temporal sequence with values within a small range, the mean value can represent the sequence truly. However, if a sequence consists of values within a large range and containing outliers, the mean will not characterize well the sequence. Due to these facts, datasets comprising the mean attribute only outperform the original dataset in small time windows and loses quality as we increase the size of the time window. The 3-year window comprise less time points than the 4 and 5-year windows, thus sequences tend to be more similar than the sequences found in the 4 and 5-year windows. Moreover, as the dataset created with 3-year window learning examples is highly imbalanced having a great number of sMCI examples comparing with cMCI. sMCI examples, as they remain stable within the MCI condition, maintain their tests scores stable as well, at least more stable than cMCI examples consequently contributing for the good performance of the model in this time window.

Hidden Markov Models made possible to create evaluation models with datasets comprising temporal information. These information is compiled into relation attributes that the HMM algorithm depicts as temporal relations and classified



	Accuracy				AUC				Specificity				Sensitivity			
	A	G	H	VAR	A	G	H	VAR	A	G	H	VAR	A	G	H	VAR
3Y	0.86	0.91	0.85	0.82	0.91	0.91	0.92	0.86	0.88	0.93	0.85	0.85	0.81	0.80	0.88	0.67
4Y	0.81	0.83	0.84	0.79	0.90	0.90	0.91	0.88	0.84	0.88	0.90	0.77	0.75	0.71	0.71	0.81
5Y	0.81	0.75	0.75	0.79	0.88	0.85	0.82	0.87	0.80	0.72	0.72	0.70	0.81	0.78	0.78	0.89

baseline values

worst than baseline

better than baseline

Figure 4: Cross Validation results for the datasets with statistics-based features.

according this information.

However, the Waikato Environment for Knowledge Analysis (WEKA) used in this work did not made possible the use of the original dataset. Missing values ha to be replaced, all instances values normalized and SMOTE was applied to increase the number of instances even if the dataset was not imbalanced. Nevertheless, results were obtained, showing slightly worse result than the baseline obtained with an independent model.

5 Conclusion and Future Work

In this work, we successfully added and treated temporal information and obtained good results considering the prognosis of the evolution to AD. The first dataset created in new leaning examples showed us an important conclusion. More information matters, even as independent features. This is a excellent starting point, especially considering these learning examples contain features from the two time points evenly and give rise to better results than only using the second timepoint. Progression features had better results in a discrete dataset leading to AUC results of 0.97% in a 5-year window with the $PF_{\div}$ feature. Discretization helps the NB algorithm as it works with discrete values by default. Considering the 5-year window, neuropsychological values have a wider range and are better distributed in the discrete bins. Another characteristic of a real word database, namely the one used in this work is the lack of periodic timepoints. Patients do have the same spacing between timepoints and this has consequences in the results obtained. Hence, temporal pattern had to be carried out in this aperiodic set, even with far from perfect conditions. Using all timepoints continues no to be an answer to the aperiodicity problem, as produces different strings to similar progressions, being very difficult for a model to learn for this dataset. As future work, other alphabets could be used to create this temporal pattern attribute and make them more complete. The CAPSUL alphabet [12] for instance, as it comprises more nominal values (up, Up, down, Down, zero and stable) can be a good option. To try and minimize the aperiodicity problem, nominal values could be arranged according to the spacing of attributes, differentiating a stable value where it is for 2 years or 5 years. Lastly, statistical summarization models were obtained through the calculation of different means and the variance. All these models have best results for the 3-year window. This is easy to understand as a mean represents better a sequence of number within a small range than a sequence containing outlier values. However, outlier values are one of the things models should look out for, as they represent behaviors different from the usual ones. Classifications using HMM models brought us very preliminary results as the environment used did not made possible further exploring.

Overall, this work allowed us to have a sense of what temporal information can do for the progression to AD problem, showing interesting results and paths to continuing to exploring.

References

[1] World Health Organization. "World health statistics 2016: Monitoring health for the SDGs, sustainable development goals." (2016).

[2] Prince, M., M. Prina, and M. Guerchet. "World Alzheimer Report 2013: journey of caring analysis of long-term care for dementia. London: Alzheimer's Disease International;" (2015).

[3] Tortora, G. J., and B. Derrickson. "Introduction to the human body: the essentials of anatomy and physiology", ed 9, Hoboken, NJ, 2011.

[4] Ewers, Michael, et al. "Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance." Neurobiology of aging 33.7 (2012): 1203-1214.

[5] Ritchie, K., Artero, S., Touchon, J., 2001. "Classification criteria for mild cognitive impairment: a population-based validation study". Neurology 56, 3742.

[6] Maroco, João, et al. "Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests." BMC research notes 4.1 (2011): 299.

[7] Hall, John E. Guyton and Hall "Textbook of medical physiology". Elsevier Health Sciences, 2015.

[8] Portet F, Ousset PJ, Visser PJ, Frisoni GB, Nobili F, Scheltens P, Vellas B, Touchon J; "MCI Working Group of the European Consortium on Alzheimer's Disease (EADC): Mild cognitive impairment (MCI) in medical practice: a critical review of the concept and new diagnostic procedure. Report of the MCI Working Group of the European Consortium on Alzheimer's Disease". J Neurol Neurosurg Psychiatry 2006;77:714-718.

[9] Pooler, Amy M., Wendy Noble, and Diane P. Hanger. "A role for tau at the synapse in Alzheimer's disease pathogenesis." Neuropharmacology 76 (2014): 1-8.

[10] Hardy, John, and Dennis J. Selkoe. "The amyloid hypothesis of Alzheimer's disease: progress and problems on the road to therapeutics." Science 297.5580 (2002): 353-356.

[11] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.

[12] Antunes, Cláudia M., and Arlindo L. Oliveira. Temporal data mining: An overview. KDD workshop on temporal data mining. Vol. 1. 2001.