*Research Journal of Recent Sciences* ISSN 2277-2502
Vol. **4(4),** 32-37, April **(2015)** *Res.J.Recent Sci.*

# Alzheimer's disease Detection using Data Mining Techniques, MRI Imaging, Blood-Based Biomarkers and Neuropsychological tests

**Alipour Aghdam Pedram[1] and Khademi Maryam[2]**
[1]Department of Computer Engineering, Islamic Azad University South Tehran Branch,P.O. Box 11365/4435, IRAN
[2]Department of Applied Mathematics, Islamic Azad University South Tehran Branch,P.O. Box 11365/4435, IRAN

**Available online at: www.isca.in, www.isca.me**
Received 25th January 2014, revised 26th May 2014, accepted 15th November 2014

## Abstract

*Finding a cure for Alzheimer's disease has been facing many challenges due to the lack of reliable biomarkers for detection and prediction of risk. Fluid based biomarkers provide some criteria for identification of the disease's current stage in patients. But these markers are not reliable predictors for disease progression or response to treatment; also most of these markers are tested in cerebrospinal fluid which reduces the applicability of the method, significantly. The main purpose of this paper is to describe research surveys in effects of blood-based biomarkers and diagnostic imaging in AD, using data mining techniques.*

**Keywords:** AD (Alzheimer's disease), CSF (Cerebrospinal Fluid), blood-based biomarker, mild cognitive impairment, data mining, SVM (support vector machine), feature vector.

## Introduction

Despite many attempts, so far no effective cure has been approved for AD and numerous failed trials has taken over news headlines. There are many reasons for failed attempts and few are possibly as follows:

Study design (e.g. short term studies, choosing subjects in advanced stages of disease which lowers the treatment response). Study components (low efficiency, choosing wrong goals or mechanisms). Lack of reliable biomarkers for choosing "right" patients in the trial (e.g. choosing patients in early stages of AD which affecting the disease is still possible).

The item 3 is based on data showing that pathology and causal factors are present few years prior to emergence of clinical symptoms and studies have shown even in those patients considered as reaching the end point of disease growth, have many differentiation and variations[1]. Thus the efforts for identifying biomarkers of early pathology changes and also biomarkers of nervous system cells, has been increased significantly. It has been predicted that development and verification of biomarkers will facilitate identification of new treatment methods and prevention strategies to detect and treat people with destructive disease.

In the last decade, the focus on blood-based biomarkers of AD has been increased. CSF-based and neuroimaging based biomarkers have a high level of precision, but their application has many limitations. Amyloid Beta peptide 42 protein, total tau protein and hyper phosphorylated tau protein levels, are well known AD biomarkers and can be used as detection markers with high sensitivity and precision and enable differentiation of AD patients and ordinary aged subjects.

Extracting cerebrospinal fluid is considered an aggressive method in many countries and there is negative attitude towards it. Moreover, the samplings are done in various time periods to examine treatment efficiency, risk levels or starting point of disease in a limited manner and can affect the marker levels. In other words markers are gathered from patients in different stages of disease which can affect the marker levels and result in imprecise outputs of data mining techniques.

Many studies have been done on neuroimaging methods like structured MRI of specific brain parts (e.g. Hippocampus) and Amyloid tracing imaging like PBI. Neuroimaging based markers have predictive value in conversion of MCI to AD, since the disease in MCI patients with positive Amyloid beta protein, most likely will progress which is not the case for patients with negative Amyloid beta protein[1]. Similar limitations exist for neuroimaging methods. In general, CSF based and neuroimaging based biomarkers are close to wide application, but there are still limitations and challenges to be dealt with. On the other hand, these markers provide useful information about the stage of the disease and type of dementia and using them alongside blood-based biomarkers will definitely increase their efficiency; especially if the blood-based indicators provide the information about the progression rate of the disease.

## Methodology

**Blood- based markers:** There few challenges to development of blood-based biomarkers for AD: i.AD has a slow rate of progression, ii. The extent of damage caused to blood-brain barrier's integrity is still unknown. Moreover, even for brain disease with extensive damage or trauma (e.g. multiple sclerosis, tissue damage and stroke) the search for a blood-based biomarker is still going on. Another challenge for AD biomarkers research is the necessity of using a control group which themselves may have the illness factors without the symptoms and face the risk of becoming an AD patient over the years. In addition, comparison between ordinary control subjects and AD patients is affected by the fact that most of the time AD subjects have other problems. So the control group should have the exact same problems with the AD subjects.

Markus Britschgi et al presented an analytical theory in which the levels of connecting factors in CSF and plasma are used to mathematically model the Aβ-42 and tau protein levels[2]. The theory is based on the relation between systematic network of intracellular signaling proteome and AD pathology markers levels and progression indicators. Results can help us identify so far unknown proteins and biological pathways included in AD. They used t-tau, p-tau181 Aβ42, or Aβ42/t-tau or Aβ42/t-tau ratios as continuous outcome variables in linear regression equation and standardized CSF or plasma protein levels alongside sex, APOE gene risk factor and age as descriptive variables. Two thirds of the communicating factors studied in the CSF are traceable. They measured the concentrations of 91 secreted proteins in plasma from 78 AD patients and 118 cognitively normal subjects. They also measured 90 proteins in CSF of 25 AD patients and 18 normal people. Seventy four proteins were detectable in plasma and 73 in CSF, with 60 proteins overlapping between the two biological fluids. Protein concentrations were generally lower in CSF than in plasma (between 5- and 8000-fold) with a few exceptions. Consistent with the E-net analysis, they observed in the connectivity network diagram of the overall 12 selected CSF communication factors strong, positive Spearman rank correlations (RS) between t-tau or p-tau181 and several communication factors in healthy individuals but Aβ42 was only weakly integrated into this network.

**MRI images:** We can use brain images to diagnose AD. Magnetic resonance imaging (MRI) is also used to understand anatomical changes of the brain that can help to diagnosis of Alzheimer's disease. The first symptoms of Alzheimer’s disease is the constriction of the hippocampus ina region of the brain which occurs very early in the AD, long before the development of disease to the cortex and appearance of symptoms in cognitive and memory impairment and its volume change is an important sign to detect Alzheimer’s disease with RI images[3]. Because the perfusion pattern is affected by the disease and these images indicate the local amount of blood flow of the brain; so we can use for the diagnosis of AD. Fung and Stoeckel were looking into the use of cerebral perfusion imaging. Cerebral perfusion acquired by single photon emitting computer tomography (SPECT). The study was comprised of a concurrent observation in order to examine using a single photon emitting computer tomography, as a diagnostic tool. People of four different clinical centers, Edinburgh (Scotland), Nice (France), Genoa (Italy), and Cologne (Germany) were included for this study. Overall158 subjects partaken, including 99 patients with AD, 28 patients suffering from depression (not used in F Fung and Stoeckel's article) and 31 healthy candidates (an instance of this data is shown in figure-1). Verification of Alzheimer's disease was obtained by clinical follow up and there was no statistically important age difference between the patients with AD and the healthy people.

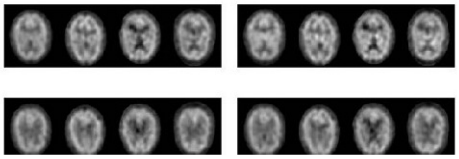

**Figure-1**
**Examples of four MRI sets, after spatial and intensity normalization[3]**

Images were labeled with four categories (very probable, probably, probably not and very unlikely to have AD) they use 16 European expert nuclear medicine physicians. They intended the first two labels as positive and the other two as negative,to be able to compare the data from the experts with that of the automatic methods. Their automatic approach (Fung and Stoeckel) that developed for classification of images, performs at least as well as human observers. Generally, their developed support vector machine (SVM) model is more sensitive and more specific. One would need more data, especially of control subjects to be able to state that automatic methods always significantly outperform human observers in clinical practice. This method was presented in their article[3], is only a general selection based on a specific image, where general information alone, may not be sufficient for clinical purposes.

Some previous works have focused on the particular problem of automatic recognition of Alzheimer's disease from MRI data. In Kloppel et al. each voxel considered to be a feature in a feature vector, and clasify feature vectors using a support vector machine[4].

discussion and comparing 10 different methods using a large data set from 509 participants is done in Cuingnet et al. work[5].

In contrast the Long and Holder tries to validate a general-purpose classifier on classification of Alzheimer's disease, which is then applied to level of education and gender, and can be applied to other measures as well in the future[6].

The purpose of this work is to discover an applicability of shape of the brain, which is done via a graph and the continuous graph search ability to pinpoint different components of the graph. The ability of this system is classifying individuals based on level of education, gender and level of cognitive impairment due to Alzheimer's disease.

Developing a computational method to rank-order AD-related proteins, based on an initial list of AD related genes and public human protein interaction data was done in Jake Yue Chen et al. work[7].In this method, first an initial seed list of 65 AD-related genes from the OMIM database were collected and mapped to 70 AD seed proteins using HGNC gene mapping table. The slight increase in protein count is due to one-to-many mapping between a gene and its multiple splice variant forms at the protein level. Using OPHID database and nearest-neighbor method, they gathered 775 human protein interactions related to AD. This enriched set of AD interactions, contain 657 human proteins. One the most important proteins in relation to AD, is tau protein which is a known effective factor in brain cell degradation seen in AD. The amyloid beta A4 precursor-protein binding protein, APPB1 (ranked 33), is a well-known interaction partner of APP, but a genetic link to AD was reported in OMIM only for the other member of the family, APPB2 (ranked 32). Nevertheless, method presented by Jake Yue Chen and et al., still predicted that APPB1 also plays some role in AD which was proven in later studies.

**CDR test data analysis:** Significant impairment in verbal memory and functionality (for example the ability to perform consecutive tasks) happen most commonly at the early stages of the AD, but it won't be inferable without formal neuropsychological testing. Reduced independence in daily activities (which is often noticed by the family members) is one of the most powerful predicting factors of the disease. Functionality status can be measured using Clinical Dementia Rating (CDR) which assesses cognitive and functionality capabilities based on a 0 to 3 scale, and higher score means lower capability. This assessment requires a source of information on independent activity capabilities in the patients, but is also useful in preliminary treatments, especially for doctors who don’t have access to neuropsychological tests. Assessment takes 30 to 45 minutes. CDR score was the strongest predictor for AD in a study including volunteers without dementia and practical scoring scale based on CDR, identified early stage AD patients effectively under clinical conditions. Neuropsychological test results, showing significant deterioration in verbal memory and functionality, helps detection of the Alzheimer's disease but execution and interpretation of the test requires a well-trained expert.

In CDR table, there were 41 records with missing values in different fields. Since there were 3038 records in total, the strategy of removing said records was considered. Graphs were produced by Clementine software and AD prediction results were constructed by SVM model. Target variables distribution in figure-2 and first Alzheimer's disease prediction in table-1 is shown.

Data were normalized using formula-1 and maximum and minimum were set as -1 and 1 respectively. This procedure was executed by SQL Server 2012.

$$x'_{ij} = \left(\frac{x_{ij}-x_{min,j}}{x_{max,j}-x_{min,j}}\right)\left(x'_{max,j} - x'_{min,j}\right) + x'_{min,j} \qquad (1)$$

In this formula $x'_{min,j}$ is the minimum (-1), $x'_{max,j}$ is the maximum (+1), $x_{min,j}$ is the minimum of the respective field, $x_{max,j}$ is maximum value of the respective field, $x_{ij}$ is the value of the field in each record, and $x'_{ij}$ is the normalized value.Target variables distribution is shown in figure-2.

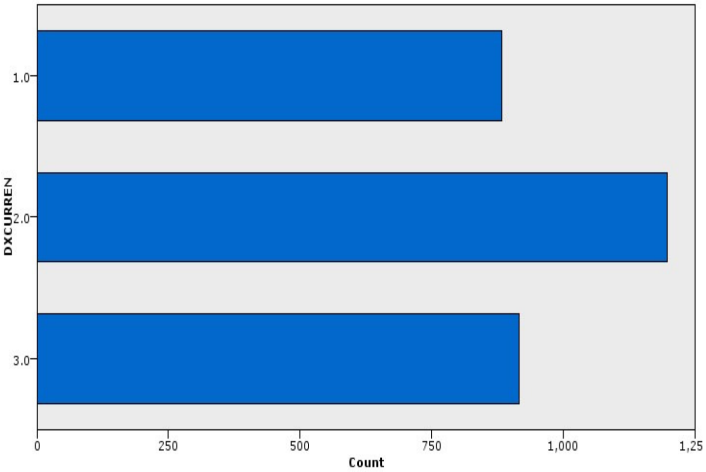

**Figure-2**
**Target variables distribution in CDR table (1=Normal; 2=MCI;3=AD)**

**Table-1**
**Alzheimer's disease prediction using CDR test data (data is not normal)**

| Sets | Learning Set | Learning Set | Test set | Test set |
|---|---|---|---|---|
| Correct | 2,367 | 87.76% | 249 | 83% |
| Wrong | 330 | 12.24% | 51 | 17% |
| Total | 2,697 | | 300 | |

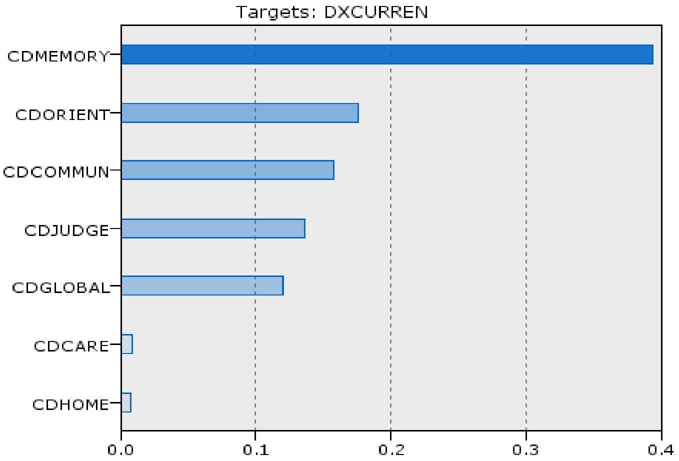

**Figure-3**
**Target variables distribution in CDR table**

## Results and Discussion

General feature selection methods can be classified in three categories: Forward Selection, Backward Selection and Step-by-step selection. Forward selection starts with one (best) feature and other features are added. Backward selection starts with selecting all features, then features with less effect on classification are removed. Step-by-step selection is a combination of two previous methods: after removing n features, m feature will be added to the set.

First and foremost criteria for selecting removed or added features, is classification error in test data set in the literature survey. The set of features producing the least amount of error in classification will be chosen for the further consideration. Next criteria in choosing features, is minimum redundancy-maximum relation (mRmR), which tries to maximize the feature relations within a class, and minimize feature redundancy in analysis.

**Table-2**
**Results after removing CDGLOBAL feature in prediction of target value**

| Sets | Learning Set | Learning Set | Test set | Test set |
|---|---|---|---|---|
| Correct | 2,342 | 86.84% | 252 | 84% |
| Wrong | 355 | 13.16% | 48 | 16% |
| Total | 2,697 | | 300 | |

**Table-3**
**Results after removing CDCARE feature in prediction of target value**

| Sets | Learning Set | Learning Set | Test set | Test set |
|---|---|---|---|---|
| Correct | 2,341 | 86.8% | 252 | 84% |
| Wrong | 356 | 13.2% | 48 | 16% |
| Total | 2,697 | | 300 | |

**Table-4**
**Results after removing CDHOME feature in prediction of target value**

| Sets | Learning Set | Learning Set | Test set | Test set |
|---|---|---|---|---|
| Correct | 2,341 | 86.8% | 252 | 84% |
| Wrong | 356 | 13.2% | 48 | 16% |
| Total | 2,697 | | 300 | |

Table-2 is shown the backward selection has been used for feature selection and the results in this analysis. We remove CDGLOBAL feature and the result is shown in table-2. Precision is increased by 1%; so the removal step will take place. Results after removing CDCARE feature and CDHOME feature is shown in table-3 and table-4 respectively, precision is not changed; so the removal step will not take place.

**Table-5**
**Results after removing CDJUDGE in prediction of target value**

| Sets | Learning Set | Learning Set | Test set | Test set |
|---|---|---|---|---|
| Correct | 2,319 | 85.98% | 246 | 82% |
| Wrong | 378 | 14.02% | 54 | 18% |
| Total | 2,697 | | 300 | |

**Table-6**
**Results after removing CDUMMUN feature in prediction of target value**

| Sets | Learning Set | Learning Set | Test set | Test set |
|---|---|---|---|---|
| Correct | 2,289 | 84.87% | 247 | 82.33% |
| Wrong | 408 | 15.13% | 53 | 17.67% |
| Total | 2,697 | | 300 | |

**Table-7**
**Results after removing CDMEMORY feature in prediction of target value**

| Sets | Learning Set | Learning Set | Test set | Test set |
|---|---|---|---|---|
| Correct | 2,115 | 78.42% | 228 | 76% |
| Wrong | 582 | 21.58% | 72 | 24% |
| Total | 2,697 | | 300 | |

After removing CDJUDGE, CDUMMUN and CDMEMORY feature precision is reduced; so the removal procedure will not take place, these results are shown in table-5, table-6 and table-7 respectively.

**Table-8**
**Results after removing CDORIENT feature in prediction of target value**

| Sets | Learning Set | Learning Set | Test set | Test set |
|---|---|---|---|---|
| Correct | 2,323 | 86.13% | 249 | 83% |
| Wrong | 374 | 13.87% | 51 | 17% |
| Total | 2,697 | | 300 | |

There is only 1 percent drop in precision when we remove CDORIENT feature (table-8), so the feature is removed. Using a plot diagram on three remaining features, outlier records are removed. For example the plot diagram of CDCOMMUN feature is shown figure-4.

As shown in the figure-4, in DXCURREN = 1 and 2, there are outlier records relative to rest of DXCURREN values. So in DXCURREN = 2, two records, and in DXCURREN = 1 two more records were removed. In total, 12 records were recognized as outlier, and were removed.1.18% increase in precision is achieved by removing outlier records (table-9).

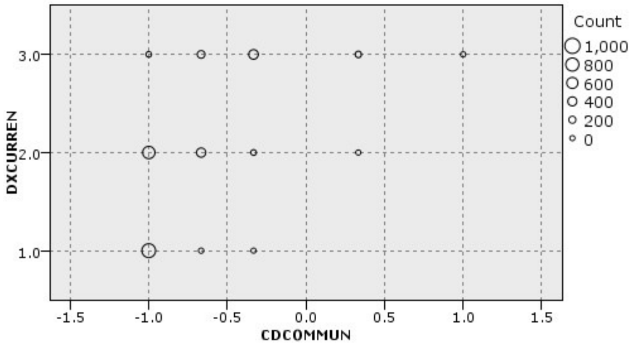

**Figure-4**
**Plot diagram of CDCOMMUN, with CDCURREN as the target feature**

**Table-9**
**Results after removing outlier records**

| Sets | Learning Set | Learning Set | Test set | Test set |
|---|---|---|---|---|
| Correct | 2,317 | 86.13% | 47 | 84.18% |
| Wrong | 373 | 13.87% | 51 | 15.82% |
| Total | 2,690 | | 297 | |

## Conclusion

Overall there are many promising methods for blood-based markers and other types of biomarkers for Alzheimer's disease and these methods are getting closer to general practicality. However, there are still multiple gaps in the related literature which were pointed out in this paper.

The main problems can be listed as following: Changes in final samples of blood may not be a sign of pathology in internal units. Blood-based markers for Alzheimer's disease require standard methods of sampling procedure. Tests should include quality control processes and calibrations, especially in proteomic oriented studies which use multiple sampling groups with longitudinal samples. Extreme precision should be introduced in definition of prognosis status (pathology vs. clinical) and the fact that lateral illnesses and medications are of critical importance in interpretation of results, should always be considered. Verification of results in single studies and independent tests are required, hence the availability of samples is very important.

While many efforts are in motion around diagnostic capabilities of blood-based biomarkers, few studies are focused on other possible applications of these biomarkers which may help development of specialized treatments. Also there is no consistency between methodologies used and the differences make the progress much harder and slower. It is still not clear that if any of these markers can provide timely diagnosis (before the damage to neurological cells become extensive) or not, and may they have predictive values or not (e.g. predicting the progress from mild cognitive impairment to AD, or risk of AD among healthy people). Moreover, in pharmaceutical world, some of these markers have shown promise as drug effectiveness measures; this part of their application hasn't been explored enough and more work is required in this area, especially considering whether its drug effectiveness (for example Aβ removal) or preventing the disease progression. The goal of this paper is to shine a light on current gaps in literature, to encourage more efforts and studies bringing the clinical application and implementations of this area, closer to reality.

Also other markers like CSF proteins, neuropsychological tests and brain images can be used and the combination of these markers can help early diagnosis of the illness and improve the precision of data mining models.

## Acknowledgement

The authors are grateful to Alzheimer's disease Neuroimaging Initiative Data Committeefor their help.

## References

**1.** Kim Henriksenemail address, Sid E. O’Bryant, HaraldHampel, John Q. Trojanowski, Thomas J. Montine, Andreas Jeromin, KajBlennow, Anders Lönneborg, Tony Wyss-Coray, Holly Soares, Chantal Bazenet, Magnus Sjögren, William Hu, Simon Lovestone, Morten A. Karsdal and Michael W. Weiner, The future of blood-based biomarkers for Alzheimer's disease, Elsevier Alzheimer's and Dementia, **(2013)**

**2.** Britschgi M, Rufibach K, Huang SL, Clark CM, Kaye JA, Li G, Peskind ER, Quinn JF, Galasko DR and Wyss-Coray T, Modeling of pathological traits in Alzheimer's disease based on systemic extracellular signaling proteome, *Molecular and Cellular Proteomics*, **(2011)**

**3.** Glenn Fung and Jonathan Stoeckel, SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information, Data Mining, Fifth IEEE International Conference, **(2005)**

**4.** Stefan Klöppel, Cynthia M. Stonnington, Carlton Chu1, Bogdan Draganski1, Rachael I. Scahill, Jonathan D. Rohrer, Nick C. Fox, Clifford R. Jack Jr, John Ashburner1 and Richard S.J. Frackowiak1, Automatic classification of MR scans in Alzheimer's disease, *Brain,* **131(3),** 681–689 **(2008)**

**5.** Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, Chupin M, Benali H and Colliot O; Alzheimer's Disease Neuroimaging Initiative, Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database, *Neuro image,* **(2010)**

**6.** Long S.S. and Holder L.B., Graph based MRI brain scan classification and correlation discovery, IEEE, Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), **(2012)**

7. Jake Yue Chen, ChangyuShen and Andrey Y. Sivachenko, Mining Alzheimer disease relevant proteins from integrated protein interactome data, *Pacific Symposium on Biocomputing,* 367-378, World Scientific, **(2006)**