AlzheTect: Bio-Marker Analysis for Early Alzheimer's Prediction

Raidel Hernandez
Computer Science
Florida State University
Tallahassee, Florida
rh13k@my.fsu.edu

***Abstract* -- One of the most obscure yet significantly debilitating diseases currently facing humanity is alzheimer's disease (AD). In 2010, there were 4.7 million individuals with alzheimer, it is predicted that by 2050 the figure will be 13.8 million [8]. To be able to effectively treat AD, it is important that we are able to identify it early in individuals. This study will focus on identifying biomarkers (features) that are strong indicators of progressive AD disease through supervised machine learning models. Then, the selected features will be used along with deep learning models to properly classify mild cognitively impaired (MCI) patients. In this research we will be able to predict if a patient, believed to have MCI, is still cognitively normal (CN) or may already be showing potential signs of AD. As well as a confidence value that the individual will start showing symptoms of AD within a relatively short period of time (1-5 years).**

Keywords -- mild cognitive impairment (MCI), Alzheimer's Disease (AD), SVM, Deep Learning (DNN), MRI, ADNI, classification.

I. INTRODUCTION

Alzheimer's disease is the leading cause of dementia. Dementia is a term used to describe the set of conditions affecting an individual's memory, thinking, and social abilities. Statistics show dementia affect 7% of individuals over the age of 65, and 20% of individuals above 80%. Alzheimer's is the most common type of dementia seen and contributes 60% to 70% of all cases [1]. Alzheimer's is a progressive neurodegenerative disease which worsens over time. However, alzheimer is not only a disease that affects the older population, it also affects many individuals under the age of 65, approximately 200,000 cases in the U.S alone, these individuals are said to have younger-onset (or early-onset) alzheimer's. It is also a disease for which there is no cure or proper treatment. However, early prediction of alzheimer's allow patients to reap the benefits of resources available to them, and reduce symptoms that affect their everyday life.

Due to this increasingly prevalent disease, being able to identify different biomarkers for early onset diagnosis is a crucial area of research. Many previous works have already shown the effectiveness of using neuroimaging techniques, especially those associated with outlining the structural features of the brain to identify differences between patients with alzheimer's and cognitively normal patients. Other strategies have recently been proposed to help accurately measure certain interest regions of the brain, which typically atrophy as the condition progresses, such as the hippocampus. Providing techniques to measure certain atrophied regions has enabled researchers with the ability to properly classify individuals by severity of dementia, or cognitive normality.

In this research, we aim to study individuals with mild cognitive impairment (MCI), specifically the early (EMCI) to late (LMCI) stages of the disease. The motivation for the research came from a challenge posted on *Kaggle* [2], by a UCL group focusing on Progression Of Neurodegenerative Disease (POND). The challenge is called The *A*lzheimer's *D*isease *P*rediction *O*f *L*ongitudinal *E*volution (TADPOLE)**, which aims at the goal of predicting which subject currently believed to have mild cognitive impairment will likely turn into an alzheimer's patient in a relatively short period of time (1-5 years). The data provided for the challenge contains labels and detailed information for 3 main categories of patients: cognitively normal (CN), several stages of mild cognitive impairment (MCI), and alzheimer's disease (AD).

From the set of data points for each patient, per disease category, a feature selection models such as Support Vector Machine (SVM) or cross validation techniques, will be used to extract a subset of important data points for each category, CN and AD. Once the subset of features is selected, a model will then be trained to classify patients into classes. Particularly, we will be training a deep neural network (DNN), on the subset of features extracted, on patients that are cognitively normal or have alzheimer's disease. Once a

model has been trained we will test the models classification accuracy with unlabeled mild cognitive impaired patient information. The reason for training the model on CN and AD patients is due to the fact that MCI patients may belong to a different classification of dementia that may not necessarily be alzheimer. However, due to the nature of how alzheimer's disease is diagnosed, which in part is typically done by ruling out other diseases or types of dementia, classifying MCI patients as AD subject is not trivial. Therefore, by training a model with CN and AD patients, the deep neural network will be able to have a better classification accuracy on the conversion of patients from MCI to AD.

II. BACKGROUND

Mild cognitive impairment (MCI) is the transition stage from cognitive decline due to age, and alzheimer's disease (AD). MCI is also the earliest detectable stage of progression todards AD [3]. Annually, around 10% to 15% of MCI patients, referred from a memory clinic, will eventually become AD patients [3]. In this research, data will be obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI)* database. The database contains many sets of data values regarding each patient analyzed, including information such as: patients age, demographics, MRI scans, PET scans, neuropsychological test results, cerebrospinal fluid, cognitive information, and many other data values.

A. Data

The ADNI study provides a data set (CSV) files for the TADPOLE challenge. The data set contains a list of individuals who have been clinically studied, and have agreed to follow-up for the next round of clinical study. Therefore, the data sets contain ADNI rollover individuals, those who have provided data to ADNI in at least two separate visits, for whom they have obtained longitudinal data from in the ADNI databases. Three sets of data are provided, which are obtained from 3 different studies and thus different databases (ADNI-1, ADNI-GO, ADNI-2). Data-set 1 (D1) is a collection of measurements and the outcomes for each patient, whether they are CN, or had MCI or AD. Data-set 2 (D2) is a collection of only baseline measurements, without outcomes for patients, for which we will forecast an outcome. Lastly, data-set 3 (D3) is the outcome for each patient against which we will evaluate the forecasted results. Therefore, D1 will be used as the training data set, D2 will be the prediction data set, and D3 will be the test data set we will use to compare results and obtain the classification accuracy.

B. Data Breakdown

ADNI provides a massive list of relevant biomarkers, for each individual, believed to have some potential influence on AD classification. Two main categories of biomarkers include: measures of amyloid beta protein, measures of damage to nerve cells. Amyloid beta proteins group together in clumps, forming plaque and blocking cell-to-cell signaling at synapses when they mis-fold. Beta-amyloid and tau levels can be measured using cerebrospinal fluid (CSF) or positron emission tomography (PET) scan, both of which we have values for in the data-sets, for each patient. Damaged nerve cells can be measured by quantifying brain metabolism using fluorodeoxyglucose (FDG) PET scan, or brain atrophy using magnetic resonance imaging (MRI), both of which we have values for in the data-sets.

MRI results are obtained after they are preprocessed by a software called FreeSurfer [4], which is used to measure brain structural integrity and calculate measures of volumes, cortical thicknesses, and surface areas [6]. Another type of MRI scan performed was diffusion tensor imaging (DTI), which help obtain values regarding the brains white matter tracts, related to cells and axons. DTI is used to estimate the brains mean diffusivity, axial diffusivity, and radial diffusivity. FDG-PET scans are used to obtain average measurements of cell metabolism, AD affected individuals tend to show reduced metabolism.

Also, a number of neuropsychological tests were conducted, by clinical experts, to assess a variety of skills such as general cognition, language, memory, vision, and many others. The cognition tests are used as an indicator of cognitive decline in AD individuals. If there is no decline in mental health, the next time taking the test, individuals should already remember how to approach it and thus perform better, however the opposite is seen in AD patients as one would expect with dementia. Other data provided include: genetic data, demographics, and medical diagnosis among many others. The results of all tests are included in the data for

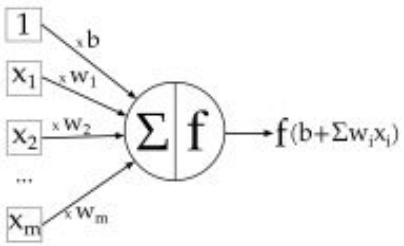

Fig. 1. ANN Neuron

each patient entry, in the data-sets provided, unless otherwise specified (D2 set doesn't include diagnosis).

III. TECHNICAL PLAN / RELATED WORK

A. Extracting Data

We will begin by extracting the data from the csv file, depending on the size of the file we will be able to either store it in memory or analyze the file through batch processing. Once we have the file we will begin to process of feeding the data through a supervised learning model. I expect to use several models to test and see which one produces better results. Some of the models I expect to use are: SVM, cross-validation, random-forest [2], logistic regression [3]. They will be used to reduce the number of data values provided, and produce a list of correlated values for each patient data-set provided. Feature extraction will be performed to every value in the training data set as well as any data we will be predicting on using this model. The results obtained (relevant fields) will then be used as inputs to a deep neural network, which will be used to classify individuals into one of 2 classes: CN, and AD.

B. Training a Deep Neural Network

An artificial neural network is a mathematical model used to solve problem related to optimization. An artificial neural network is an interconnected network of neurons, which take an input, computes on that input, and produce a value. More specifically, a neuron sums the values of previous neurons with the weight of each connection. Initially a neurons must be initiated with help of an activation function. A very commonly used activation function uses a sigmoid neuron which output a value between [0, 1]. After a neuron calculates all possible values it must add a bias at every layer of a neural network. Fig. 1 (to the left) is an example of a neuron receiving values from previous layers and applying an activation function to produce a single output for the next layer.

The model will be trained using the D1 data-set provided by ADNI for the challenge. Once the features have been extracted from the original file, using a supervised learning classifier, they will be used for training using a Deep neural network. According to a previous study, using the data provided by ADNI, a DNN with around 11 layers, including hidden and visible layers, and reduced number of neurons will perform best [2]. In their research any added layers after 11, or more neurons did not yield much improvement. Amoroso et al. [2] used an input layer consisting of 2056 units, and a 4 unit output, since they were trying to classify 4 different classes: CN, cMCI, MCI, AD. However, for our research and methods the size of the input layer will be determined once the feature classifier, described previously, has selected values of interest. The output layer will consist of 2 outputs and a confidence for the result that was predicted, the possible prediction classes are: CN, or AD. Also, The DNN will produce a weighted graph that may be kept and used for further predictions without the need to re-train, we may even further train the model if we wish as well through the graph produced. Future works will focus on providing prediction of more types of dementia classification.

REFERENCES

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. *http://adni.loni.usc.edu*

** TADPOLE Contest details: *https://tadpole.grand-challenge.org*

[1] "Dementia Fact sheet N°362". World Health Organization. March 2015. Archived from the original on 18 March 2015. Retrieved 13 January 2016.

[2] Nicola Amoroso, Domenico Diacono, Annarita Fanizzi, Marianna La Rocca, Alfonso Monaco, Angela Lombardi, Cataldo Guaragnella, Roberto Bellotti, Sabina Tangaro, Deep learning reveals Alzheimer's disease onset in MCI subjects: Results from an international challenge, Journal of Neuroscience Methods, 2017

[3] Moradi, Elaheh & Pepe, Antonietta & Gaser, Christian & Huttunen, Heikki & Tohka, Jussi. (2014). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. NeuroImage. 104. 10.1016/j.neuroimage.2014.10.002.

[4] Reuter, M., Schmansky, N.J., Rosas, H.D., Fischl, B. 2012. Within-Subject Template Estimation for Unbiased Longitudinal Image Analysis. Neuroimage 61 (4), 1402-1418.

[5] Harper, Lorna et al. "MRI Visual Rating Scales in the Diagnosis of Dementia: Evaluation in 184 Post-Mortem Confirmed Cases." Brain 139.4 (2016): 1211–1225. PMC. Web. 23 Mar. 2018.

[6] Rachael, Scahill, et al. Mapping the evolution of regional atrophy in Alzheimer's disease: Unbiased analysis of fluid-registered serial MRI, Proceedings of the National Academy of Sciences Apr 2002, 99 (7) 4703-4707; DOI:10.1073/pnas.052587399

[7] Yang, E, et al. "Quantifying the Pathophysiological Timeline of Alzheimer's Disease." Journal of Alzheimer's Disease : JAD., U.S. National Library of Medicine,2011,www.ncbi.nlm.nih.gov/pubmed/21694449.

[8] Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. Neurology. 2013;80(19):1778-1783. doi:10.1212/WNL.0b013e31828726f5.