A Comparative Study of Neural Network Techniques to Perform Early Diagnosis of Alzheimer's Disease

Lara Dantas and Mêuser Valença
Polytechnic School of Pernambuco
University of Pernambuco
Recife, Brazil
Email: {ldc, meuser}@ecomp.poli.br

***Abstract*—The life expectancy of the population in most developed countries is growing every day and hence there is an increase of several age-related diseases. In Brazil, about 1.2 million people have Alzheimer's disease (AD), now considered the most common type of dementia in the population. Although it is a degenerative and irreversible disease, if diagnosed early, medications may be administered to slow the progression of symptoms and provide a better quality of life for the patient. Previous studies with classifiers contained in the software Weka using a database with values of 120 blood proteins, and they noticed that they could classify the patient may or may not be diagnosed with AD with an accuracy rate of 94% and 85%, respectively. Thus, this study aims to use a new connectionist approach called Reservoir Computing (RC) to perform early diagnosis of a patient with or without AD, also compare these results with those obtained using a Neural Network Multi-Layer Perceptron (MLP). This article also envisions to utilize the Random Forest Algorithm to select proteins from the original set and, thus, create a new protein signature.**

Keywords–Reservoir Computing; Alzheimer's Disease; Neural Network.

I. INTRODUCTION

More developed countries are undergoing a major demographic shift. The older segments of the population are growing at a faster rate, and therefore, there is an increase in age-related diseases, especially progressive dementia disorders. First described by psychiatrist Alois Alzheimer in 1907, Alzheimer's Disease (AD) is today the most common cause of dementia in the elderly population.

According to the Brazilian Institute of Geography and Statistics (IBGE) and the World Health Organization (WHO), there are 1.2 million people with AD in Brazil. It is believed that only 5% of the patients developed the disease at an early stage, i.e., before 65 years of age. In patients where the AD started after 65 years old, it is estimated that between 10% and 30% of cases started after 85 years old [1].

AD is a degenerative disease that causes irreversible death of several brain cells, the neurons. The patient suffering from this disease has a brain with microscopic pathologic lesions, known as neuritic plaques, and neurofibrillary tangles [2]. In addition, the brain of a person with Alzheimer's is much smaller than the brain of a healthy person.

This disease develops in each patient in a unique way; however, there are several symptoms common to all of them, for example, loss of memory, language disorders, depression, aggression, among others.

Initially, the patient loses episodic memory, i.e., memory that holds information of events and their spatio-temporal relations. Thus, the old facts and the facts that just happened are easily forgotten.

With the progress of the disease, semantic memory is also lost, i.e., lexical knowledge, rules, symbols are forgotten and the patient begins to lose its cultural identity [3].

Although it is an irreversible disease, if it is discovered in its early stage, medications may be administered to slow the progression of symptoms and prolong the patient's welfare [4]. Thus, it is extremely important that mechanisms are developed for the prediction of AD in the whole population.

In the literature, Herbert et al. [5] conducted a study using a database of 120 samples of proteins contained in plasma of several patients. He concluded in his research that a combination of 18 of the 120 available proteins enabled the realization of early diagnosis of AD with a accuracy rate of 91% using a set of tests with data from 92 patients who were diagnosed with AD or not.

In addition, he also used another set of tests containing data from 47 patients diagnosed with Mild Cognitive Impairment (MCI). For this set, the accuracy rate was 81%. These values were calculated from the average success rates found for all classifiers used in clinical trials for both sets [6].

Afterwards, Gòmez [6] conducted a study using 20 different classifiers available in the software Weka and set various signatures with 18, 10, 6 and 5 proteins. These proteins were all contained in the set described by Herbert et al. The 10 protein signature reached a accuracy rate of 89% using the AD test set and of 66% for the MCI test set. These success rates were also calculated from the average values of the 20 classifiers used [6].

Since the results available in the literature use only classification techniques provided by the software Weka [7], this work will use a new connectionist approach called Reservoir Computing (RC) [15] to perform the classification of a patient, given the set of 5 and 10 proteins defined by Gòmez et al.

In order to compare the results obtained using the RC, the classification was also performed using a Multiple-Layer Perceptron neural network (MLP). This topology is widely used and it has a good performance for classification problems [8].

In addition to that, a study using the Random Forest Algorithm [9] was conducted in order to select a new signature of 10 proteins for the prediction of the Alzheimer's Disease and another for the diagnosis of MCI. This step was performed to reduce costs for the diagnosis since Gòmez used paid sofware and other techniques for variable selection.

 ISBN: 978-1-61208-340-7

This article is organized into several sections. The first section contains information about the operation, structure and simulation of Reservoir Computing. The next section describes the methodology used throughout this work, i.e., what is the database used and how it is organized, the experiments and statistical analysis that was performed. Finally, there is a section that displays the results and the last one with the conclusions obtained in this work.

II. *Reservoir Computing*

Recurrent Neural Networks (RNN) were created to enable the solution of dynamic problems. This is accomplished through a feedback of a neuron in a layer i to that found in some previous layer, $i - j$. This neural network topology has a better resemblance to the operation and behaviour of the human brain [10].

In 2001, a new approach for the design of the training of a RNN was proposed by Wolfgang Mass called Liquid State Machine (LSM) [11]. At the same time, but independently, the same approach was described by Herbert Jaeger and called Echo State Machine (ESN) [12].

Both ESN and LSM networks have the Echo State Property (ESP) [13], i.e., due to the recurrent network connections, information from previous entries are stored. However, these data are not stored for an infinite period of time, and as well as the human brain, old information must be forgotten over time. Thus, the neural network has a rich set of information from the past and present therefore enhancing its applicability to dynamic systems [14].

In 2007, Verstraeten coined the term Reservoir Computing (RC) that unified the concepts described in ESN and LSM. Since then, this term is used in literature to illustrate learning systems which are represented by a dynamic recurrent neural network [15].

The RC is composed of three parts: an input layer, which as the MLP, represents the input variables of the problem, a reservoir, which can be seen as a large distributed and dynamic RNN with fixed weights, and a linear output layer called readout.

Figure 1 represents the RC topology with two neurons in the input layer, three in the reservoir and one neuron in the output layer.

A. *Construction and Simulation of RC*

The RC used in this work was developed in the Java programming language to make use of the object-oriented paradigm. This framework was created in order to solve classification and prediction problems and it was validated by three Benchmarks available: Iris species, Thyreoide cancer and diabetes.

The first step to be taken in order to prepare the RC and perform the data set classification is configuring its architecture. Thus, it is necessary the amount of neurons that will be used in input and output layers and the reservoir itself is defined.

Furthermore, several RC parameters should also be determined. Being a recent methodology, there are no studies that

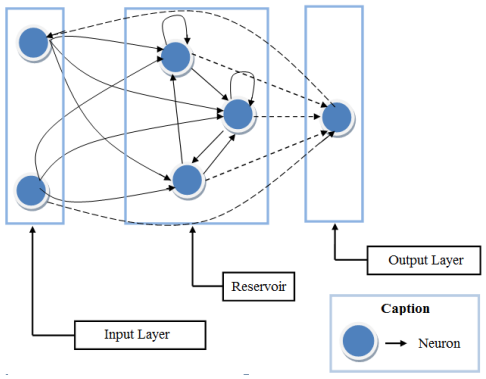

Figure 1: RC architecture. The dashed lines represent the weights that should be adjusted during the training of the network.

prove how many neurons in the reservoir are necessary so that the neural network has better performance, or the rate of connectivity between these neurons. Therefore, for this work, values were defined based on some empirical tests performed.

Once the architecture of the neural network is determined, the next step is to generate the weight matrices connecting the input layer to the reservoir, W_{in}, and the matrix with the weights between neurons in the reservoir, W_{res}. Both matrices are generated with random values between -1 and 1.

Studies claim that the matrix W_{res} must have a spectral radius equal to 1 to provide a more numerical stability [16], i.e., when W_{res} is initialized, it must have its values changed as follows:

- Initially it must be decomposed into singular values;
- Then, W_{res} should have its values changed until the maximum value of the main diagonal of the eigenvalues matrix is less than or equal to 1.

To perform the simulation of the RC, the database is divided into three sets: training, used to perform the update of the states of the neurons of the reservoir, cross-validation, used to stop the training of the neural network, and test set, used to calculate the RC classification rate [17].

The states of the neurons in the reservoir must be initialized to zero. Since this is a recurrent network and RC stores its states (M_{est}) in a matrix, it is necessary that the final values found by the network are not so influenced by this initialization. Therefore, the literature suggests that before start training, a set of cycles called warm up is executed in order to perform updates in the states of the neurons in the reservoir and overlook the influence of the initial value. The states are updated according to (1) [15]:

$$x[k+1] = f(W_{res}x[k] + W_{in}u[k]) \quad (1)$$

where, $W_{in}u[k]$ represents the matrix containing the result of the product of the values derived from the input layer by the weights connecting these neurons to the reservoir at a time k and $W_{res}x[k]$ is the matrix with the states of the neurons from

 ISBN: 978-1-61208-340-7

the reservoir at the same time k. The result will be assigned to $x[k+1]$, i.e., the state of the neuron RC in an instant forward will be the result of calculating the activation function of the neuron from the sum of the two parcels described above. In this work, the activation function used was the hyperbolic tangent according to 2.

$$f(net_i) = \frac{e^{net_i} - e^{-net_i}}{e^{net_i} + e^{-net_i}} \quad (2)$$

Once the period of warm up is over, the training of the RC can be initialized. The first step should be to load the training set and perform the update of the states of the reservoir, noting that the matrices W_{in} and W_{res} should not be changed. They are randomly generated during construction of the RC, as described in the previous section, and should not be adjusted.

Still during training, the weights matrix that connects the neurons of the input layer to the output (W_{inout}) and the one that connects the reservoir to the output must be calculated by the pseudo-inverse of Moore-Penrose. As they are non-square matrices and their determinants can approach zero, it is necessary to calculate the pseudo-inverse.

At the end of each training cycle, a cross validation cycle should be initiated. This process should be repeated until the stopping criteria is reached and the training set is finalized. During the process of cross-validation, the matrices W_{inout} and W_{out} should remain being readjusted.

When the process of training is finished, the testing process begins. The set of tests is presented to the RC and at this time, all the weights matrices, W_{in}, W_{res}, W_{inout} and W_{out}, should remain unchanged, as the matrix M_{est}. At this point, the classification error is calculated. These values will be used in the future to make the necessary comparisons.

The behaviour of the RC can be best viewed through the algorithm described in Figure 2.

Pseudocode of RC

```
1  Set the number of neurons in the input layer ;
2  Set the number of neurons in the reservoir layer ;
3  Set the number of neurons in the output layer ;
4  Randomly generate the weights of Win matrix between
   -1 e 1;
5  Randomly generate the weights of Wres matrix between
   -1 e 1;
6  Normalize the weights of Wres matrix so that the
   spectral radius of the matrix is smaller than or equal to
   1;
7  while until the end of the number of warm up cycles do
8  |  updates the states of the neurons of the RC;
9  end
10 while until the stopping criterion is reached do
11 |  for each value of the input set do
12 |  |  updates the states of the neurons of the RC;
13 |  end
14 |  Calculates the Moore-Penrose inverse matrix to find
   |  the weights connecting the RC to the output layer;
15 |  Calculates the Moore-Penrose inverse matrix to find
   |  the weights connecting the input layer to the output
   |  layer;
16 |  for each value of the cross-validation set do
17 |  |  updates the states of the neurons of the RC;
18 |  end
19 |  Calculates the output values of the RC;
20 |  Calculates the RMSE;
21 |  Checks if the stopping criterion has been reached;
22 end
23 for each value in the set of tests do
24 |  updates the states of the neurons of the RC;
25 end
26 Calculates the output values of the RC;
27 Calculate the accuracy rate;
```

Figure 2: Reservoir Computing pseudocode

III. METHODOLOGY

A. Database

The database used in the development of this work was the same used by Gòmez et al. in his publication. It has values of 120 proteins found by analysis of blood samples from different patients. The ultimate goal of the database is to classify whether a patient can be diagnosed or not with AD or MCI [6].

In his work, Gòmez et al. [6] subdivided the database in 2 sets. The first set contained the results of blood samples of 83 patients. Of these 83 patients, 68 were allocated to the training process of the chosen classifier. The data for the remaining 15 patients were used in the process of cross-validation of the classifier, i.e., a process that determines the optimal point to stop its training [8].

The second set, used in the testing process of the classifier, has two options. It could be used to diagnosis AD and in this case, this set will contain the samples related to the 92 patients that could be diagnosed with AD. The second option is use this set to perform diagnosis of MCI. In this other case, the test set will contain blood samples related to 47 patients with a possible diagnosis of MCI.

Aiming a comparison with the signatures previously defined by Gòmez et al. [6], two new signatures of 10 proteins were proposed. One of them used to perform the diagnosis of Alzheimer's Disease and another for MCI.

The Random Forest algorithm was executed 30 times and in each of the simulations, a signature of 10 proteins was defined. To choose the best signature among 30 proposals, a new classifier was used to determine which one got a better classification rate.

Thus, the Support Vector Machine (SVM) was chosen to perform 30 executions with each signature. Thus, an average of the classification rates was calculated, besides the analysis of the standard deviations. The signature that obtained the best average classification rate was chosen to be used later by the RC and the MLP.

It is important to mention that Gòmez defined the same signature for both cases, that is, the signature composed by 10 proteins is used for the AD and MCI testing sets.

Table I shows the signatures of proteins that are contained in Gòmez et al. work and which are used in this study.

In this work, 4 databases were prepared in order to reproduce the experiments described by Gòmez et al., using the MLP and the RC. They were:

- 1 database for testing the signature of 10 proteins defined by Gòmez et al. with the DA set of tests, called

TABLE I: Representation of the proteins contained in each one of the signatures used.

Abbreviation	Signature	Proteins
S1	**10 proteins signature defined by Gòmez et al.**	CCL7/MCP-3, CCL15/MIP-1d, EGF, G-CSF, IL-1a, IL-3, IL-6, IL-11, PDGF-BB, TNF-a [6]
S2	**10 proteins signature defined by the Random Forest for AD test set**	IL-1a, TNF-a, G-CSF ,PDGF-BB, IGFBP-6, M-CSF, EGF, IL-3, GDNF, Eotaxin-3
S3	**10 proteins signature defined by the Random Forest for MCI test set**	IL-1a, PDGF-BB, EGF, TNF-a, RANTES, FAS, GCSF, MIP-1d, FGF-6, IL-11

now by **Database 1**;

- 1 database for testing the signature of 10 proteins defined by Gòmez et al. with the MCI set of tests, called now by **Database 2**;
- 1 database for testing the signature of 10 proteins defined by Random Forest Algorithm with the DA set of tests, called now by **Database 3**;
- 1 database for testing the signature of 10 proteins defined by Random Forest Algorithm with the MCI set of tests, called now by **Database 4**;

All databases described above maintained the organization used by Gòmez et al. regarding the division of values for the training, cross validation and testing set.

1) Pre-processing of data: To properly execute the training of the neural network it is necessary that your data is normalized, i.e., the input values of the neural network must be contained in the same numerical range. This is important since very different values can influence the training and generate a loss in the generalization ability of the neural network [8].

One of the most commonly used normalization techniques in literature is the linear transformation and it was the one chosen for this work. Equation (3) is the formula used to normalize the values of the database.

$$y = ((b - a) \times (\frac{x - x_{min}}{x_{max} - x_{min}})) + a \quad (3)$$

In (3), a and b represent the maximum and minimum values that the data should take. In this work, it was used the value of -0.85 for a and 0.85 for b, as the activation function chosen for this neural network is the Hyperbolic Tangent. Therefore, the values contained in the database must be between -1 and 1.

B. Simulations

As described above, in order to compare the results obtained with RC, it was used a neural network MLP. Table II shows which parameters were chosen to perform the simulations with the RC and MLP. They were obtained through empirical testing and the settings that showed the lowest mean squared errors in the cross-validation procces were chosen.

After defining the settings of the RC and the MLP, 30 simulations were performed with each of the databases in each of the chosen neural network topologies in this work. This number is considered ideal to perform more meaningful statistical comparisons [18].

TABLE II: Representation of the parameters used for the simulations with the RC and MLP

Parameters	RC value	MLP value
RC connectivity	20%	Not applicable
Number of neurons in the input layer	10	10
Number of neurons in the RC	4	Not applicable.
Number of neurons in the hidden layer	Not applicable	20
Number of neurons in the output layer	2	2
Number of warm up cycles	100	Not applicable.
Activation function of neurons in the reservoir or hidden layer	Hyperbolic Tangent	Hyperbolic Tangent
Activation function of neurons in the output layer	Linear	Linear
Learning rate	Not applicable.	0.7
Momentum	Not applicable.	0.4

C. Statistical Analysis

When all the simulations were completed, it was necessary to perform a sequence of statistical tests in order to scientifically validate the results. For this, it was used the R mathematical software, since it contains all the implementations of the tests used. This software uses as default a level of significance (α) previously defined with the value of 0.05.

Before using a parametric test on a data set is necessary to check whether the samples are normally distributed and if they have statistically equal variances. If these two assumptions are validated, one can apply a parametric test, otherwise it must be used a non-parametric test.

Thus, it was applied the Shapiro-Wilk test to verify whether or not the samples were normally distributed and the F-test to verify whether or not the samples were drawn from the same population, i.e., if their variances were statistically equal.

As none of the four datasets met these two premises at the same time, it was not possible to perform the Student's T-test. Thus, it was chosen the Wilcoxon Rank-Sum Test, since it is a non-parametric test, i.e., makes no assumptions about the probability distribution of the samples.

IV. RESULTS

After all simulations were performed with the databases, it was calculated the arithmetic mean for each set of simulations and Table III displays those values found.

When applied the Wilcoxon Rank-Sum Test for each of the four cases, the results found were that the MLP has a statistically better performance than the RC, except for the case 2. Thus, with the 10 proteins signature defined by Gòmez et al. and MCI test set, the test indicated there is not statistical differences between the results obtained with both techniques.

In order to verify the performance of the new proposed signatures, simulations with RC and MLP topologies were

TABLE III: Representation of the average accuracy rates after the 30 experiments

Database	Average accuracy rate with RC / Standard Deviation	Average accuracy rate with MLP / Standard Deviation
Database 1	86.62% / 0.026	93.44% / 0.017
Database 2	69.29% / 0.024	68.15% / 0.018
Database 3	90.57% / 0.022	94.31% / 0.008
Database 4	76.59% / 0.047	78.86% / 0.031

TABLE IV: Comparison of the results with the new protein signature proposal obtained with RC, MLP and the ones available in literature.

Protein Signature	RC - Maximum value	MLP - Maximum value	Results found by Gòmez et al.
New 10-protein signature for AD	96.73%	95.65%	93%
New 10-protein signature for MCI	89.36%	82.97%	66%

performed for both the diagnosis of AD and MCI. The results were compared with those found by the same neural networks when the signatures used were the ones defined by Gòmez et al.

In all cases, the classification rate showed improvement when the new signatures were used for both architectures. After the Wilcoxon Rank-Sum test, this statement was confirmed statistically. The maximum and average values are also bigger than those described by Gòmez et al.

Table IV summarizes the maximum values found in the simulations of the RC and MLP. Those results were found using the new signature proposoal with 10 proteins. The Table IV also display the results obtained in the work of Gòmez et al. using ther own signature [6].

From Table IV, it can be concluded that the RC obtained results consistent with those described by Gòmez et al. and succeeded in reaching a maximum value greater than the average found in the literature.

V. Conclusion and Future Work

Nowadays, Alzheimer's disease is one of the most common diseases in the elderly population. More recently, the number of patients has grown significantly since the life expectancy in most developed countries has increased.

AD is a degenerative disease, i.e., brain cells will deteriorate and there is no way to reverse the disease. However, the earlier the drugs are administered, the better the quality of life of the patient since the medication will slow the progression of symptoms.

Thus, this study aimed to verify the performance of a new connectionist neural network approach called Reservoir Computing to early classify if a patient can be diagnosed with AD or not. Moreover, another goal was to make a comparison of the performance of the RC with the MLP neural network, and also with the results available in the literature.

From the statistical tests and simulations, it can be concluded that the MLP presented a superior performance in most cases, although the RC have obtained maximum values higher than the MLP and available in the literature. This can be explained by the fact that the RC is more suitable for use in dynamic systems as it has a good storage capacity.

It is also possible to conclude that the 2 new signatures proposed achieved better results when compared to those showed by Gòmez et al. Furthermore, they also had better perfomance when compared to the results obtained from the same neural network topologies when the signatures used were the ones proposed by Gòmez et al.

Unlike the RC, the MLP has a better ability to approximate non-linear functions and does not contain the property to store the previous state of its neurons, i.e., does not have recurrence.

As future work, it is intended to conduct a comparative study between the RC non-linear approach capability and storage capacity of the network, in order to assess the appropriate parameters to obtain better results.

It will be also carried out a thorough study on parametrization of the RC, such as the definition of the spectral radius size, how many neurons should be placed in reservoir, the degree of connectivity between these neurons. Furthermore, in order to address dynamic problems, recurrence will be implemented between the neurons of the output layer with the reservoir.

Finally, it is intended to invest in more variable selection techniques in order to further optimize the results and to reduce the number of proteins in the signatures used to perform early diagnosis of Alzheimer and MCI.

References

[1] H. Portal. Portal of the brazilian health [retrieved March, 2014] [Online]. Available: http://portal.saude.gov.br/saude/ (2012)

[2] R. Green, Diagnosis and Treatment of Alzheimer's Disease and Other Dementias. Editora de Publicações Científicas Ltda., 2001.

[3] M. Paulo, Dementia of Alzheimer type: diagnostic, treatment and social aspects. Editora de Publicações Científicas Ltda., 1997.

[4] E. Giusti and V. Surdo, Alzheimer's care and family counseling: psychological needs and treatment of dementia. Gryphus, 2010.

[5] H. Charles, Y. Takeda-Uchimura, B. Adam, S. Ray and B. Markus. "Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins" Nat Med, vol. v13, 2007, pp. 1359 - 1362.

[6] M. Gómez and P. Moscato, "Identification of a 5-protein biomarker molecular signature for predicting alzheimer's disease," PLos One, vol. v3, 2008, p. 12p.

[7] S. David and R. Peter, WEKA Experimenter Tutorial for Version 3-5-8, 2008.

[8] M. Valença, Fundamentals of Neural Networks: Examples in Java, 2nd ed. Livro Rápido, 2009.

[9] StatSoft. Random forests. https://www.statsoft.com/Textbook/Random-Forest. [retrieved: May, 2014]

[10] M. Valença, Applying Neural Networks: A Complete Guide, 1st ed. Livro Rápido, 2005.

[11] W. Maass, Motivation, theory, and applications of liquid state machines. Imperial College Press, 2011.

[12] H. Jaeger, "The echo state approach to analysing and trainning recurrent neural networks," German National Resource Center for Information Technology, Tech. Rep., 2010.

[13] M. Massar and S. Massar, "Mean-field theory of echo state networks," PHYSICAL REVIEW E, vol. 87, 2013.

 ISBN: 978-1-61208-340-7

[14] M. Lukoševičius, A practical guide to applying echo state networks. Springer Berlin Heidelberg, 2012, pp. 659-686.

[15] D. Verstraeten, "Reservoir computing : computation with dynamical systems," Ph.D. dissertation, Ghent University. Faculty of Engineering, Ghent, Belgium, 2009.

[16] A. Ferreira, Araújo, "A method for design and training reservoir computing applied to prediction of time series," Ph.D. dissertation, Universidade de Federal de Pernambuco, 2001.

[17] M. Kulkarni and C. Teuscher, "Memristor-based reservoir computing," in Proceedings of the IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH'12), Amsterdam, 2012, pp. 226-232.

[18] N. Juristo and A. Moreno, M., Basics of Software Engineering Experimentation. Kluwer Academic Publishers, 2001.

 ISBN: 978-1-61208-340-7