Utilizando Redes Neurais Artificiais para Análise de Sensibilidade da Base de dados utilizada para diagnóstico da Doença de Alzheimer

Trabalho de Conclusão de Curso

Engenharia de Computação

Natália Guimarães Marinheiro

Orientador: Prof. Mêuser Valença

Natália Guimarães Marinheiro

Utilizando Redes Neurais Artificiais para Análise de Sensibilidade da Base de Dados utilizada para Diagnóstico da Doença de Alzheimer

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Engenharia de Computação

Escola Politécnica de Pernambuco

Universidade de Pernambuco

Orientador: Prof. Mêuser Valença

Recife - PE, Brasil

Dezembro de 2017

Este trabalho é dedicado à minha família.

POLI
ESCOLA POLITÉCNICA DE PERNAMBUCO

MONOGRAFIA DE FINAL DE CURSO

Avaliação Final (para o presidente da banca)*

No dia 19 de dezembro de 2017, às 10:00 horas, reuniu-se para deliberar a defesa da monografia de conclusão de curso do discente **NATALIA GUIMARAES MARINHEIRO**, orientado pelo professor **Mêuser Jorge Silva Valença**, sob título **Utilizando redes neurais artificiais para análise de sensibilidade da base de dados utilizada para diagnóstico da doença de Alzheimer**, a banca composta pelos professores:

Sérgio Mário Lins Galdino
Mêuser Jorge Silva Valença

Após a apresentação da monografia e discussão entre os membros da Banca, a mesma foi considerada:

☒ Aprovada ☐ Aprovada com Restrições* ☐ Reprovada

e foi-lhe atribuída nota: 9,5 (nove , cinco)

*(Obrigatório o preenchimento do campo abaixo com comentários para o autor)

O discente terá 07 dias para entrega da versão final da monografia a contar da data deste documento.

Sérgio Galdino

SÉRGIO MÁRIO LINS GALDINO

MÊUSER JORGE SILVA VALENÇA

* Este documento deverá ser encadernado juntamente com a monografia em versão final.

Agradecimentos

Agradeço primeiramente à Deus por me dar forças para vencer minhas batalhas diárias em busca dos meus objetivos. Aos meus pais e todo o corpo docente da Universidade de Pernambuco pela contribuição na minha formação, nos princípios que levarei para toda a vida. E ao meu orientador por toda a paciência e pelo exemplo de profissional que ele é.

"Guardei as coisas mais incríveis,
ao pensar que íamos nos esconder.
Não me arrependo;
recordações valem muito
mais do que vestidos."
Anne Frank

Resumo

A doença de Alzheimer representa atualmente 60% dos casos de demência no mundo. Com o crescimento previsto da população idosa espera-se que o número de casos de idosos no mundo diagnosticados com o Mal aumente significativamente, podendo triplicar até 2050. Apesar da grande quantidade de casos registrados, a ciência ainda desconhece muitos aspectos da doença, tais como sua etiologia; e o seu diagnóstico é ainda é identificado tardiamente, quando a doença já se manifesta de forma significativa no cérebro do paciente. Os benefícios de um diagnóstico precoce do mal de Alzheimer variam desde a contribuição com as pesquisas correntes à facilidade do paciente em obter informações e realizar ações para garantir a melhoria da sua qualidade de vida durante as fases da doença. Existem na literatura trabalhos relacionados à identificação dessa doença de forma mais precoce, dentre eles Ray et Al, Gómez e Moscato e Lara realizaram estudos em uma base de dados contendo 220 amostras do plasma sanguíneo de pacientes contendo 120 valores de concentrações de proteínas do sangue. Essas proteínas são consideradas biomarcadores da presença de DA e, portanto, foram aplicadas em estudos matemáticos na tentativa de prever um diagnóstico para a doença. Os trabalhos realizados pela literatura constituíram e definir subconjuntos dessa base de dados para realizar o diagnóstico de DA com mais eficiência. Este trabalho se propõe a utilizar as redes neurais artificiais MLP e RC para realizar uma análise de sensibilidade na base de dados utilizada pela literatura a fim de determinar o efeito causado pela variação dos conjuntos de entrada das redes sobre o desempenho das assinaturas de base de dados encontradas nos trabalhos anteriores. Como resultado dessa análise foi identificado que a base de dados é sensível às variações durante a classificação, obtendo menores taxas de acertos nos diagnósticos. Desta forma julga-se necessário melhor coleta dos dados associados à os biomarcadores que constituem essa base.

Abstract

Alzheimer's disease currently accounts for 60 % of the world's dementia cases. With the expected growth on the elderly population, it is expected that the number of elderly people diagnosed with this disorder will increase significantly, and may triple by 2050. Despite the large number of registered cases, the science is still unaware of many aspects of the disease, such as its etiology; and its diagnosis is still identified late, when the disease is significantly spread in the patient's brain. The benefits of an early diagnosis of Alzheimer's disease vary from the contribution with current research to the patient's ability to obtain information and take actions to ensure the improvement of their quality of life during the phases of the disease. In the literature, there are few studies related to the identification of this disease, among them Ray et al, Gómez and Moscato and Lara carried out studies in a database containing 220 samples of patient's blood plasma containing 120 values of blood protein concentrations. These proteins are considered biomarkers of the presence of AD and therefore have been applied in mathematical studies in an attempt to predict a diagnosis for the disease. The work carried out by the literature constituted and defined subsets of this database to perform the diagnosis of AD more efficiently. This work proposes to use the artificials neural networks MLP and RC to perform a sensitivity analysis in the database used in the literature to determine the effect caused by the size variation of the networks' input sets on the performance of the database signatures found in previous works. As a result of this analysis it was identified that the database is sensitive to variations during classification, obtaining lower rates of correctness in the diagnoses. In this way it is considered necessary collecting more accurate data associated with the biomarkers that constitute this base.

Lista de ilustrações

Lista de abreviaturas e siglas

DA	Doença de Alzheimer
ND	Nenhuma demência
MCI	*Mild cognitive impairment*
MLP	*Multi-Layer Perceptron*
RC	*Reservoir Computing*
LSM	*Liquid State Machines*
ESN	*Echo state networks*
RNR	Redes Neurais Recorrentes
RNA	Redes Neurais Artificiais

Sumário

1 Introdução

1.1 Motivação e Caracterização do Problema

A Demência afeta certa de 50 milhões de pessoas no mundo, este número deverá aumentar para 131,5 milhões até 2050. A forma mais conhecida de demência é a doença de Alzheimer, que representa cerca 50 a 60% de todos os casos mundiais [1]. No Brasil, em 2010, 1 milhão de idosos tinham Alzheimer e estima-se que em 2020 serão 1,6 milhão [2]. Pondera-se que 58% das pessoas com demência vivem em países de baixa ou média renda, uma proporção que é antecipada a crescer para 68% até 2050.

O Deficit Cognitivo Leve, em inglês, *Mild Cognitive Impairment* (MCI) é definido como um declínio cognitivo maior do que o esperado para a idade e o nível de educação de um indivíduo que não interfere notavelmente com as atividades da vida diária [3]. Em muitos casos o MCI é considerado uma fase intermediária entre o desenvolvimento da DA, uma vez que a grande percentagem de indivíduos são identificados com DA após alguns anos do diagnóstico de MCI, mas ainda existem casos em que o quadro do paciente não se desenvolve para um declínio cognitivo maior ou DA. De qualquer forma, é por muitos considerado um indicador do aparecimento da Doença de Alzheimer

Apesar do crescimento significativo da população idosa e consequentemente dos casos de diagnóstico de DA, MCI e outras demências, as pesquisas ainda não são capazes de responder muitas questões relacionadas. Ainda não se sabe ao certo qual a causa desses males e quanto tempo antes dos primeiros sintomas se manifestarem a demência vem se desenvolvendo no paciente. Portanto o diagnóstico precoce ainda não é preciso, tomando vários ramos, dentre eles o de classificação com base em diagnósticos prévios abordado nesse trabalho.

As pesquisas mostram que em 2011 a maioria das pessoas que vivem com demência não recebeu um diagnóstico formal. Em países de alta renda, apenas cerca de 20 a 50% dos casos de demência foram reconhecidos e documentados [4].

A Federação Internacional de Associações de Alzheimer acredita que o principal objetivo do diagnóstico precoce de demências como DA e MCI é o acesso oportuno à informação por parte do paciente e da família, aconselhamento desde clínico até financeiro, apoio emocional e acesso à tratamentos e cuidados eficazes desde o momento do diagnóstico até o final da vida. Acredita também que possíveis terapias medicamentosas que possam ser descobertas pelas pesquisas são susceptíveis de funcionar melhor quando aplicadas antes do dano extenso e permanente ter ocorrido no cérebro do paciente, portanto, nos estágios iniciais da demência, ou mesmo antes da doença ser clinicamente evidente [4].

Com base nessa necessidade de um diagnóstico mais preciso e precoce do Mal de Alzheimer, o avanço da aplicação de Redes Neurais Artificiais no campo do diagnóstico de doenças tem criado alternativas para esse problema. Essas redes propõem utilizar os dados já existentes, coletados nos laboratórios, para criar um modelo matemático capaz de prever casos futuros. As redes neurais artificiais possuem a característica de aprender analogamente ao cérebro humano e para isso precisam de adquirir memória de experiência. A memória das redes neurais é formada a partir do treinamento dos casos de doença previamente diagnosticados que são documentados nas bases de dados. Dessa forma uma rede, assim como um humano, aprende com os casos já existentes. Dai pode-se afirmar a importância da coleta dos dados para execução das redes de classificação do Alzheimer. É importante que o universo dos exemplos coletados e dos marcadores biológicos selecionados sejam suficientes para que o diagnóstico não seja impreciso ou fuja da realidade afim de adquirir confiabilidade nas informações geradas.

Atualmente os estudos mais importantes no ramo da identificação dessas doenças utilizaram uma base de dados contendo 120 marcadores biológicos relativos ao diagnóstico de DA. Essa base de dados se constitui em cerca de 220 amostras sanguíneas contendo valores de concentrações de 120 proteínas do plasma sanguíneo de pacientes identificados com DA.

Ray et al utilizou um algoritmo de centroide encolhido chamado Análise Preditiva de Microarrays (PAM) para encontrar um subconjunto desses marcadores biológicos contendo 18 proteínas capazes de classificar as amostras com 95% de concordância positiva e 83% de concordância negativa com o diagnóstico clínico [5]. Um ano depois os pesquisados Gómez e Moscato foram capazes de, utilizando 24 diferentes classificadores disponíveis no pacote de software Weka, gerar uma assinatura de 5 proteínas para a mesma base de dados com em média 96% de precisão ao prever DA [6].

Em 2015 na Universidade de Pernambuco, Lara, utilizando os algoritmos de seleção de variável *Random Forest* e *Information Gain* identificou um subconjunto da base de dados composto por 3 proteínas que treinadas nas redes *multi-layer Perceptron*, *Reservoir Computing* e *Extreme Learning Machine* (ELM) apresentou taxas de acerto estatisticamente iguais às encontradas por Goméz e Moscato. Lara provou que é possível diminuir o custo financeiro para realizar a coleta de dados dos pacientes durante o diagnóstico de DA, uma vez que conseguiu encontrar para uma assinatura com quantidade menor de proteínas o mesmo desempenho de classificação dos trabalhos anteriores.

Este trabalho se propõe a estudar a viabilidade dos dados de entradas para a classificação de DA em pacientes com ou sem MCI diagnosticado previamente. Assim sendo, foram utilizadas as assinaturas de proteínas de melhor desempenho descobertas nos trabalhos anteriores para gerar uma variação na execução da classificação dos dados por redes neurais e obter estatísticas que fundamentaram essa análise estatística. As redes

neurais utilizadas foram *Multi-Layer Perceptron* (MLP) e *Reservoir Computing* (RC).

1.2 Objetivo Geral

Este trabalho tem como objetivo fazer uma análise de sensibilidade da base de dados utilizada pela literatura para classificação da doença de Alzheimer em pessoas sem demência anterior ou que foram diagnosticadas com Deficit Cógnitivo Leve anos antes.

Através de variações na divisão de dados durante o treinamento das Redes Neurais Artificais pretende-se avaliar a viabilidade da base de dados coletada atualmente para todos os estudos na área. Pretende-se coletar informações relacionadas à distribuição, variância e amplitude das taxas de acertos da classificação das redes neurais para cada uma das bases geradas.

1.3 Objetivos Específicos

- Implementar um algoritmos para pré-processamento dos dados e geração dos conjuntos aleatórios de treinamento, validação cruzada e teste que serão utilizados nas Redes Neurais;
- Implementar as 2 topologias escolhidas de redes neurais na linguagem Java e validar seu comportamento;
- Comparar todos os resultados obtidos neste trabalho estatisticamente em uma análise de sensibilidade;

1.4 Estrutura do Documento

Este trabalho foi dividido em 5 capítulos, incluído este que introduz a motivação e objetivos para desenvolvimento deste trabalho. Em seguinda o Capítulo 2 expõe os conceitos utilizados como fundamentação teórica para os testes e análises realizadas, bem como para entendimento do problema à ser resolvido pelo desenvolvimento do trabalho.

O capítulo seguinte, 3, descreve os métodos utilizados para obtenção dos resultados que são explanados no capítulo 4. Bem como parâmetros de entrada, base de dados utilizada e testes estatísticos.

Finalmente o capítulo 5 encerra este trabalho com uma discussão dos resultados encontrados

2 Fundamentação Teórica

Este capítulo apresenta alguns conceitos fundamentais a respeito de principais Demências que acometem os idosos atualmente.

Também é detalhada a teoria que fundamenta as redes neurais utilizadas para geração de diagnóstico de DA.

2.1 Demências

A Demência, nome coletivo para síndromes cerebrais progressivas que afetam memória, pensamento, comportamento e emoção é a principal causa de deficiência e dependência entre os idosos (ADI). Os tipos mais comuns de demência são Doença de Alzheimer, descrita no próximo tópico; demência vascular, causada por falhas de oxigenação no cérebro que ocorrem quando os vasos sanguíneos são danificados; demência com corpos de Lewy, causada pela morte de células cerebrais, neurônios, devido ao acumulo das proteínas anormais (alfa-sinucleína), chamadas de corpos de Lewy; demência frontotemporal, causada por morte das células cerebrais do lobo frontal do cérebro, sendo mais rara, acomete comumente pessoas entre 40 e 50 anos.

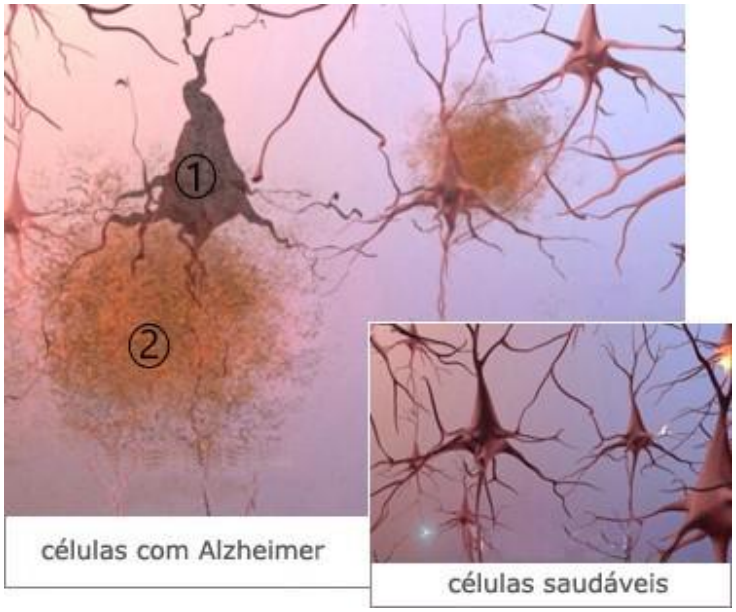
2.1.1 Doença de Alzheimer

O mal de Alzheimer é uma doença degenerativa das células do cérebro que atinge atualmente cerca de 1,6 milhões de pessoas no Brasil [2]. Em 1906 o médico Alemão Alois Alzheimer observou uma redução significativa e depósitos anormais dentro e ao redor das células nervosas durante autópsia da paciente Auguste D., que em seu estágio terminal apresentava sinais fortes de demência, tais como perda de memória. Cerca de 20 anos depois, com o avanço das pesquisas relacionadas ao mal que se intitulou de Alzheimer, foram identificadas a proteína amiloide cerebrovascular, conhecida como beta-amiloide e a proteína tau no cérebro de pacientes, seu acumulo foi associado à morte das células cerebrais que causam a doença.

Como pode ser visto na figura 1, a proteína beta-amiloide(1) se concentra em placas ao redor das células nervosas bloqueando as sinapses (comunicação entre os neurônios) e também podem ativar as células do sistema imunológico que causam inflamações e devoram células deficientes. A proteínas tau(2), por sua vez são responsáveis pela estabilização dos filamentos de transporte celular conhecidos como microtubulos, o aumento da concentração dessa proteína converte-os em filamentos torcidos chamados de emaranhados. Esses filamentos não conseguem mais se manter retos., se rompendo e desintegrando.

Dessa forma nutrientes e outros suprimentos essenciais não conseguem mais se movimentar através das células, que acabam morrendo. Ainda não foi constatado pelas pesquisas que as duas proteínas sejam a principal causa da morte das células cerebrais, porém elas são as principais suspeitas.

Figura 1 – Células nervosas Cerebrais de pacientes com e sem a DA



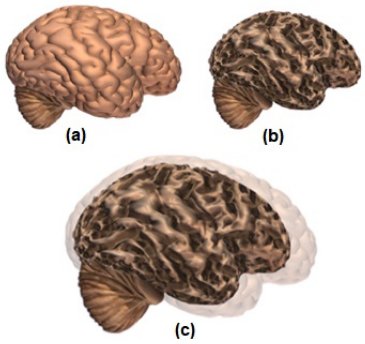
Fonte: extraído de [7].

Esse distúrbio cerebral é progressivo e danifica e eventualmente destrói as células cerebrais, causando com o passar do tempo um encolhimento cerebral que pode afetar quase todas as funções do corpo humano. A figura 2 ilustra respectivamente o cérebro de uma pessoa sem a doença, o cérebro de uma pessoa em estágio avançado da doença e o comparativo do tamanho desses dois cérebros.

Embora as pesquisas que se seguiram revelaram muito sobre a DA, ainda há muito a se descobrir sobre as mudanças biológicas que causam a doença, porque ela progride mais rapidamente em alguns pacientes e como a doença pode ser melhor diagnosticada, prevenida, retardada ou curada.

Atualmente sabe-se que o mal pode atingir de acordo com os grupos de classificação, homens e mulheres acima de 65 anos, quando é chamada de senil, e pessoas com menos de 65 anos, quando é titulada pré-senil.e pode ser dividida em três estágios: leve, moderada e grave [2] Na fase leve, são observados sintomas como perda de memória recente, dificuldade para encontrar palavras, desorientação no tempo e no espaço, dificuldade para tomar decisões, perda de iniciativa e de motivação, sinais de depressão, agressividade, diminuição do interesse por atividades e passatempos.

Figura 2 – Comparativo dos cérebro de pessoas com e sem DA. (a) Cérebro de pessoas saudáveis. (b) cérebro de pacientes com o Mal de Alzheimer. (c) comparativo entre o tamanho dos cérebros citados anteriormente.



Fonte: extraído de [7].

Na fase moderada são observados aumento de todos os sintomas iniciais e da perda de memória, dificuldade na execução de atividades sociais tais como fazer compras, cozinhar, limpas e dificuldades com a fala e orientação espaço-temporal.

Na fase mais grave o déficit cognitivo abrange grandes áreas cerebrais de forma que as pessoas precisam de ajuda na execução de atividades básicas da vida diária, bem como tomar banho, se vestir, comer e usar o banheiro, perdem suas habilidades de comunicação, não reconhecem parentes, amigos e objetos familiares, se tornam acamados e dependentes de cuidados integrais. O ritmo de aparecimento desses sintomas varia de acordo com a pessoa. A DA tem etiologia ainda desconhecida, seu aparecimento é comumente confundido com outros tipos de demência que acometem pessoas na fase idosa, tornando o diagnóstico mais difícil e tardio, uma vez que pacientes e familiares tendem a não procurar o diagnóstico da doença assumindo que os sintomas apresentados sejam apenas do processo de envelhecimento humano normal. As terapias medicamentosas atualmente disponíveis tratam sintomas, não alteram fundamentalmente o curso da doença. Os inibidores da colinesterase (donepezil, rivastigmina e galantamina) são licenciados para doença de Alzheimer leve a moderada [8].

Acredita-se que os danos cerebrais causados por esse distúrbio começam a surgir no cérebro humano anos antes dos sintomas iniciais serem percebidos. Quando as mudanças iniciais ocorrem, o cérebro compensa por elas, permitindo que os indivíduos continuem a funcionar normalmente. À medida que o dano neuronal aumenta, o cérebro já não pode compensar as mudanças e os indivíduos começam a mostrar declínio cognitivo sutil, iniciando-se assim a primeira fase da doença [7].

O diagnóstico precoce da DA nos dias de hoje se baseiam em diversos ramos, no

estudo de marcadores biológicos, ou seja, aspectos biológicos dos pacientes que já foram diagnosticados com a doença para ajudar a definir diagnósticos futuros; em Neuroimagem, ressonância magnética ou tomografia computadorizada; estudo dos níveis das proteínas do líquido cefalorraquidiano, beta-amiloide e tau no cérebro humano que estão associados à morte das células cerebrais; associações genéticas. Cientistas descobriram que a presenca gene APOE-e4 no DNA de uma pessoa é um indicador de risco de desenvolvimento da doença, não garantindo 100% que isso irá ocorrer, e além disso, é também um indicador de que os sintomas apareçam em uma idade mais jovem do que o habitual. Foi descoberto também um gene determinístico do apareceminento da doença, que está associado ao aumento da produção de beta-amilóide, o fragmento de proteína que é um dos principais associado à morte das células cerebrais. Esse gene é muito raro e está associado apenas a 5% dos casos de Alzheimer diagnosticados. Os pesquisadoras sinalizam também dois fatores de risco para o aparecimento do Mal, tais como a idade avançada, após 65 anos, o risco de aparecimento da doença duplica a cada cinco anos. Após 85 anos, o risco atinge quase 50%; e o histórico familiar, as pesquisas mostram que aqueles que têm um pai, irmão, irmã ou filho com doença de Alzheimer são mais propensos a desenvolver a doença.

2.1.2 Déficit Cognitivo leve

O Déficit cognitivo leve, em inglês, *Mild Cognitive Impairment*, causa mudanças cognitivas que são suficientemente graves para serem percebidas pelos indivíduos que as experimentam ou para outras pessoas, mas as mudanças não são suficientemente graves para interferir com a vida diária e independência do paciente. Estudos sugerem que 15 a 20 por cento daqueles com 65 anos ou mais podem ter MCI. (ALZ MILD) Acredita-se que indivíduos diagnosticados com MCI tem maior probabilidade de desenvolver a DA.

Assim como a DA, MCI ainda tem muitas questões deixadas em aberto pela ciência, as pesquisas ainda não foram capazes de identificar possíveis causas do seu aparecimento e nenhuma relação com o DA além da probabilidade estatística baseada em diagnósticos prévios. Sabe-se que as drogas aprovadas para tratar os sintomas da doença de Alzheimer não mostraram nenhum benefício duradouro em adiar ou prevenir a progressão da MCI para a demência.

2.2 Redes Neurais

Redes Neurais é um ramo da inteligência artificial que se baseia no funcionamento do cérebro humano. Observou-se que o cérebro humano é um sistema de processamento maciçamente paralelo que é capaz de realizar o processamento de informações mais rápido que o mais rápido dos computadores existentes [9]. As Redes Neurais Artificiais (RNAs) se propõem a construir um modelo matemático que simule o funcionamento do cérebro

usando como base o neurônio.

Essas células cerebrais humanas, neurônios, tem papel muito importante no processamento de informações humana, porém são formadas por uma estrutura simples que possui um corpo celular, conhecido como axônio e os dendritos. O primeiro é responsável por transferir informações pelo sistema nervoso, o segundo recebe as informações trazidas pelo axônio de outro neurônio através de sinapses. As Sinapses são regiões de contato entre os neurônios, nessas regiões que são transferidos os impulsos nervosos entre o dendrito do neurônio que recebe informação e o axônio do neurônio que envia informações em forma de impulsos elétricos.

O neurônio artificial representa o biológico utilizando uma regra de propagação e uma função de ativação. A regra de propagação é um somatório das informações passadas como entradas (Xi) multiplicada por um peso (Wi) definido afim de representar o efeito de uma sinapse, a resposta ao estímulo do neurônio, saída, é dada pela função de ativação definida.

2.2.1 Perceptron

O Perceptron é o modelo mais simples de rede neural, conhecido como rede neural de realimentação, pois é capaz de fazer com que neurônio artificial aprenda de acordo com a quantidade de vezes que é executado, tornando sua resposta ao estímulo mais precisa.

Nesta rede vários neurônios da camada entradas estão conectadas a um só neurônio na camada de saída, a resposta ao estimulo criada pelas entradas da rede é uma função linear que tem como entradas os valores dos neurônios de entrada e uma ponderação para cada um deles chamada de peso sináptico. Durante o treinamento da rede os pesos sinápticos são reajustados de acordo com a saída esperada pela função, fazendo com que a rede aprenda durante o treinamento.

O resultado do Perceptron é definido pela Lei de Tudo ou Nada que define quando o neurônio está ativo (resultado 1) ou inativo (resultado 0), definida utilizando a função degrau. Uma das limitações do Perceptron é a resolução de problemas matemáticos cujos conjuntos de dados pertençam obrigatoriamente a duas classes distintas, chamados de linearmente separáveis, devido à sua função de saída ser a função Degrau. Essa limitação foi posteriormente resolvida pela rede MLP [9].

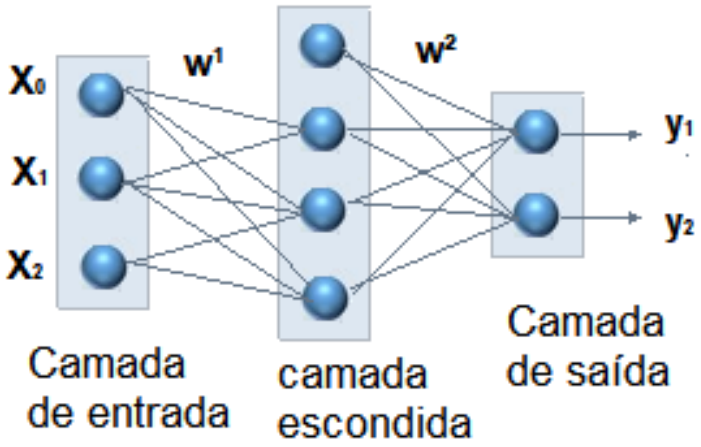
2.2.2 Multi-Layer Perceptron

Multi-layer Perceptron (MLP) é uma rede neural de realimentação que consiste na generalização da rede simples Perceptron em camadas de neurônios organizadas em uma ordem onde os neurônios de uma camada estimulam todos os da camada seguinte e após

cada execução do treinamento a rede propaga seu sinal entre os neurônios estimulando o aprendizado.

É composta por no mínimo três camadas de neurônios artificiais, a primeira, camada de entrada representando as entradas da rede, onde cada neurônio representa uma variável da base de dados utilizada para classificação; uma ou mais camadas intermediárias, também conhecidas como camadas escondidas, responsáveis pela não linearidade da rede, permitindo que o MLP possa resolver problemas reais, ou seja, funções não linearmente separáveis; e uma camada de saída representando a resposta da rede a um estímulo, isto é, o resultado da classificação da rede em relação às entradas fornecidas. A figura 3 ilustra um exemplo de arquitetura de uma MLP:

Figura 3 – Exemplo de arquitetura da *Multi-Layer Perceptron*



Fonte: extraído de [14].

O MLP, por ser uma adaptação do Perceptron, também é uma rede de realimentação, faz o reajuste dos pesos durante o treinamento da rede, aplicando o algoritmo de *Backpropagation*. O algoritmo é dividido em duas fases: *forward* e *backward*, sendo a primeira onde ocorre a propagação progressiva do sinal de entrada e a segunda onde ocorre a retropropagação do erro gerado pela saída para que a rede possa aprender. O aprendizado da rede está associado ao ajuste dos pesos na fase *backward* de acordo com o erro gerado pelas execuções anteriores. Durante a fase *forward* é feito o cálculo da entrada líquida para as camadas de acordo com a equação 2.1 e depois calculado o resultado da saída utilizando função de ativação do neurônio. As funções de ativação mais comumente utilizadas pela literatura são a linear, tangente hiperbólica e sigmoide logística, descritas

respectivamente pelas equações 2.2 2.3 e 2.4.

$$net_i^m = \sum_{j=0}^{N} W_{ij}^m . X_j \tag{2.1}$$

$$F(net_i^m) = net_i^m \tag{2.2}$$

$$F(net_i^m) = \frac{1}{1 + e^{-net}} \tag{2.3}$$

$$F(net_i^m) = \frac{e^{net} - e^{-net}}{e^{net} + e^{-net}} \tag{2.4}$$

Como a camada intermediária do MLP também possui uma função de ativação, o resultado de suas saídas por sua vez será utilizado em conjunto com os pesos sinápticos para calcular a entrada líquida para a função de ativação da camada de saída. Desta forma, a fase *forward* do algoritmo consiste em gerar a entrada líquida e a função de ativação para as duas camadas, intermediária e de saída.

Durante a fase *backward* para que os neurônios aprendam com a classificação dessa amostra é necessário calcular o erro relacionado à classificação das entradas. O erro é calculado com base no valor correto do sinal (desejável) definido na base de dados e no valor calculado pela rede neural de acordo com a equação 2.5:

$$e = d_i - y_i \tag{2.5}$$

Onde d é o valor desejável que já foi previamente definido na base de dados e y é o valor calculado pela rede neural. Em posse do valor desse erro podemos calcular o reajuste dos pesos nas camadas de acordo com as equações,2.6, 2.7 e 2.8.

$$W_{i,j}^m(t+1) = W_{i,j}^m(t) + \alpha . \delta_i^m . f^{m-1}(net_j^{m-1}) + \beta . \Delta W_{i,j}^m(t-1) \tag{2.6}$$

$$\delta_i^m = f'^m(net_i^m) . e_i \tag{2.7}$$

$$\delta_i^{m-1} = f'^{m-1}(net_j^{m-1}) . \sum_{i=1}^{N} W_{i,j}^m . \delta_j^m \tag{2.8}$$

2.2.3 Reservoir Computing

Diferente da MLP, RNA de alimentação adiante, *Reservoir Computing* (RC) possui uma arquitetura recorrente, ou seja, a presença de conexões recorrentes ou realimentação

de informação que pode conduzir a comportamentos complexos, mesmo com um número reduzido de parâmetros.

As redes neurais recorrentes (RNR) oferecem um método atraente para resolver tarefas de engenharia complicadas. Elas têm as vantagens das redes de feedforward, que incluem robustez, aprender por exemplos e a capacidade de modelar sistemas altamente não-lineares e adicionam a isso uma capacidade de processamento temporal inerente [10].

Uma RC se concentra em dois princípios de Redes Neurais Recorrentes (RNR) Liquid State Machines (LSM) e Echo State Network (ESN) que foram desenvolvidos independente e simultaneamente por Jaeger [11] e Mass [12].

A topologia da rede é definida por duas camadas lineares uma de entrada e outra de saída e uma cama intermediária chamada Reservoir, como ilustrado na figura 4. O Reservoir é uma camada recorrente de neurônios que são ligados à camada de saída e de entrada e podem ser ligados entre si mesmos em conexões recorrentes. A presença dessas conexões internas do Reservoir introduz uma forma de memória na rede devido ao fato de que a informação não flui em uma direção pela rede, mas continua circulando por dentro e está integrada com informações de execuções anteriores. O fato de que a memória existe dentro da rede deve-se ao fato de que a rede é um sistema dinâmico. De fato: os valores de ativação (os estados) dos neurônios não são apenas determinados pela entrada atual, mas também pelo estado anterior da rede (e, portanto, recursivamente, por todas as entradas anteriores). Apenas a camada de saída é treinada e é por este motivo que ela possui uma função de saída. Essa função pode ser, por exemplo, um classificador linear ou um algoritmo de regressão [13]

A quantidade de neurônios na camada de saída e de entrada é definida de acordo com os dados propostos no problema, ou seja, a quantidade de neurônios de entrada deve ser equivalente à quantidade de parâmetros de entrada do problema e a quantidade de neurônios na camada de saída deve ser equivalente à quantidade de variáveis de saída definidas na base de dados. Na camada de recorrência, a quantidade de neurônios é definida empiricamente, pois ainda não se tem nenhuma definição precisa de como calcular a melhor quantidade de neurônios afim de obter o melhor desempenho da rede.

Inicialmente, para treinamento dessa rede, é definida uma matriz de pesos da ligação entre a camada de entrada e o Reservoir, W_{ent} com tamanho ExN onde E é a quantidade de entradas e N a quantidade de nós da camada recorrente; e uma matriz de pesos internos do Reservoir Wres de tamanho NxN onde N é a quantidade de nós da camada recorrente. A taxa de conectividade da rede define a quantidade de nós ligados entre si dentro da camada intermediária, de acordo com a porcentagem definida para essa taxa aumenta-se a quantidade de pesos não nulos na matriz Wres.

As N execuções iniciais da rede Reservoir Computing podem ser definidas como

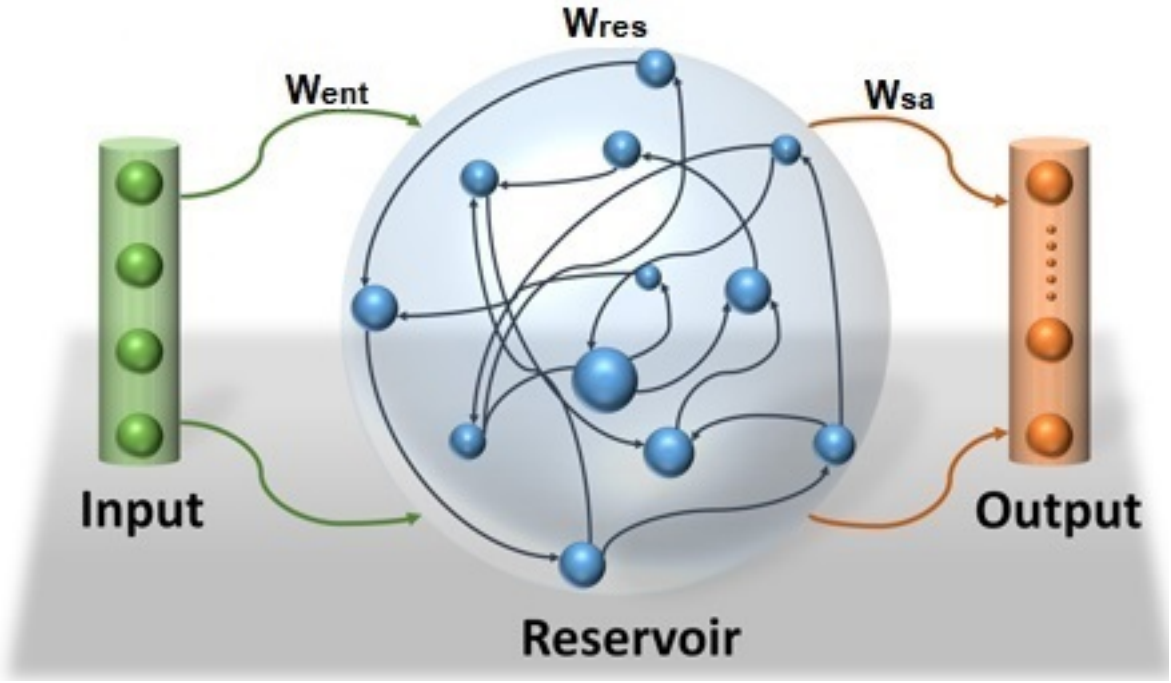


Figura 4 – Exemplo de arquitetura de *Reservoir Computing*

Fonte: extraído de [14].

"warm up". "Warm up" são ciclos que são utilizados apenas para atualização do estado do Reservoir afim de que os pesos iniciais sejam apagados da memória da rede, e assim, a mesma perca a influência do valor zero atribuídos a esses estados iniciais. Portando durante essa fase não é necessário a geração da matriz de pesos da camada de saída nem calcular o valor das saídas da rede, não é considerada o treinamento da rede em si. Após os ciclos do warm up o treinamento é iniciado, durante o treinamento os estados do Reservoir são atualizados durante os ciclos segundo a seguite equação:

$$x[t+1] = f(W_{res}.X[t] + W_{ent}.U[t]) \tag{2.9}$$

$W_{res}X[t]$ indica matriz que guarda o produto valores das variáveis de entrada pelos valores dos pesos que conectam a camada de entrada ao reservatório no instante t

$W_{ent}U[t]$ indica a matriz de pesos do reservatório no mesmo instante e X[t+1] representa o estado seguinte à t de um neurônio do Reservoir.

Ao fim do treinamento os únicos pesos ajustados na rede são os pesos da camada de saída, utilizando a seguinte equação:

$$W_{sa} = W_{res}^{+}.Y \tag{2.10}$$

Onde,

W_{res} representa a pseudo-inversa da matriz com os estados dos neurônios do reservoir, para isso é comumente utilizada a técnica *one-shot* da pseudo-inversa de Moore-Penrose.

W_{sa} representa a matriz de pesos que ligam os neurônios do reservoir à camada de saída

Y representa os valores desejados para os neurônios da camada de saída.

A pseudo-inversa possui um papel estratégico na resolução de sistemas lineares inconsistentes, uma vez que generaliza o conceito de inversa de modo que toda matriz real (inclusive as retangulares) possuam essa inversa generalizada, tornando possível calcular a inversa de matrizes como W_{res} que é muitas vezes definida não inversível uma vez que pode não ser quadrada e e seu determinante se aproxima de 0 .

2.2.4 Validação Cruzada

Além da quantidade de ciclos de treinamento definida para uma topologia de uma RNA, é importante definir critérios de parada do treinamento afim de obter uma boa generalização dos dados e evitar o fenômeno conhecido como *overfitting*.

Overfitting é o ponto do treinamento no qual a rede já memorizou os exemplos de treinamento, mas não aprendeu a generalizar para novas situações, o gráfico de erro do conjunto de treinamento continua a diminuir, porém o erro de classificação de novos valores da rede aumenta de valor, como pode ser visto no gráfico 5. A partir do ponto de parada definido temos o *overfitting*, após esse ponto o que a rede aprende é apenas ruído contido nos dados de treinamento.

Para aplicação da validação cruzada é necessário determinar o subconjunto de exemplos da base de dados que será reservado para buscar o melhor ponto de parada, aplicar o cálculo da entrada líquida e a função de ativação para definir o valor da saída e então calcular o erro médio quadrático (EMQ) através da equação:

$$EMQ = \frac{1}{NM}\sum_{p=1}^{N}(y_i - d_i)^2 \tag{2.11}$$

N representa o número de exemplos que foi utilizado.

M representa o número de neurônios na camada de saída.

d_i representa o valor desejado pela rede neural.

y_i representa o valor calculado pela rede neural.

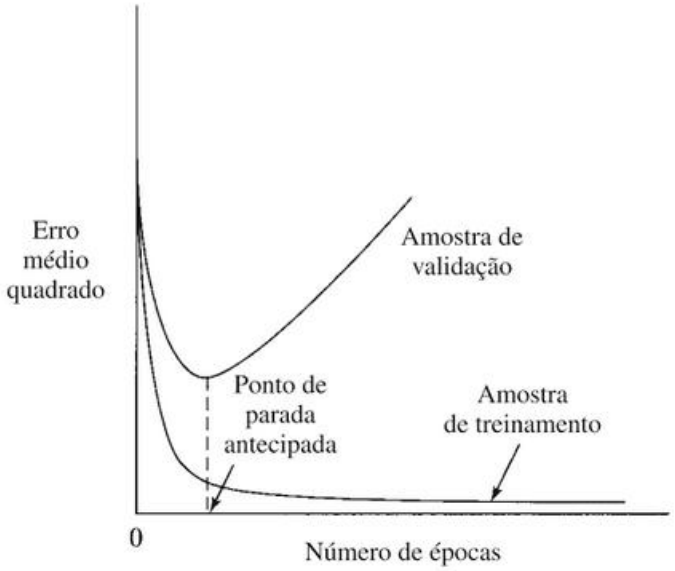


Figura 5 – Gráfico Validação Cruzada

Fonte: extraído de [14].

2.2.5 Testes das Redes Neurais

Após o treinamento de cada uma das redes é necessário verificar se já são atendidos os critérios de parada do treinamento, que são detalhados no item deste capítulo. Uma vez que os critérios de parada são atendidos, a rede já está treinada e pronta para classificar outras amostras fora do conjunto de treinamento, neste passo iniciam-se os testes. Nessa fase vamos testar se a rede aprendeu durante o treinamento, ou seja, o MLP ou RC é utilizado para classificar a amostra e então é possível definir a taxa de acerto entre a saída calculada pela rede e saída informada pela base de dados.

A saída calculada pela rede é o resultado de uma função (geralmente linear), um valor real, porém a base de dados utilizada possui duas saídas representadas em valores binários (0 ou 1), então se faz necessário definição de um parâmetro para comparação com os valores definidos na base de dados.

Neste trabalho foi escolhida a regra de aprendizagem competitiva chamada de "Vencedor Leva Tudo" para igualar os valores reais da saída. Na aprendizagem competitiva os neurônios de saída de uma rede neural competem entre si para se tornarem ativos (disparar), somente um neurônio está ativo em um determinado instante. É essa característica que torna a aprendizagem competitiva muito adequada para descobrir caraterísticas estatisticamente salientes que podem ser utilizadas para classificar um conjunto de padrões de entrada [14].

Essa regra define que para um neurônio K ser um vencedor (representar o impulso),

sua saída calculada $f(net_k)$ para um padrão x de entradas deve ser a maior entre todas as outras saídas de neurônios na camada de saída, podemos descrever:

$$\begin{cases} 1, se f(net_k) > f(net_j) \;\; para j, \;\; j \neq k \\ 0, caso \;\; contrario \end{cases}$$

Finalmente, o cálculo da taxa de acerto desse teste é definido então, dado um número N de exemplos de teste, pela quantidade de exemplos que foi corretamente medido pela RNA.

3 Metodologia

3.1 Base de dados

Neste trabalho foi utilizada a base de dados da literatura, já utilizada por Ray et al [5], Gómez e Moscato [6] e Lara [15], para definição de assinaturas de proteínas para classificação de DA em pacientes com e sem MCI identificada anteriormente.

A fim de constituir essa base foram coletadas 222 amostras sanguíneas de pessoas nas seguintes condições:

- Pacientes diagnosticados com DA;
- Pacientes sem nenhuma demência diagnosticada;
- Pacientes diagnosticados inicialmente com MCI que desenvolveram DA dentro de um período posterior de 2 a 6 anos;
- Pacientes diagnosticados inicialmente com MCI que não desenvolveram DA (Permanecendo com diagnóstico de MCI ou outras demências);

Essas amostras são formadas pelo valor de concentrações de 120 proteínas no plasma sanguíneo que são considerados marcadores biológicos da identificação de DA. Marcadores biológicos são entidades que podem ser medidas para indicar de forma precisa e confiável a presença de uma doença, nesse caso, a DA pode ser identificada de acordo com a concentração desse conjunto de 120 proteínas do plasma sanguíneo humano.

Desta forma, 47 amostras dessa base são de pacientes que receberam diagnóstico de MCI anos antes do diagnóstico (positivo ou negativo) de DA, essa quantidade foi utilizada para gerar testes relacionados ao aparecimento de DA em pacientes com diagnóstico prévio de MCI. O restante dos 175 pacientes participou de testes para diagnóstico apenas da DA, essa quantidade foi utilizada para treinamento e validação cruzada das redes MLP e RC e testes em pacientes sem diagnóstico prévio de MCI. A base de dados foi dividida em diversas combinações de conjuntos de treinamento, validação cruzada e testes afim de gerar estatísticos que comprovem sua eficiência na classificação do Mal de Alzheimer.

3.1.1 Pré-processamento

Redes Neurais são incapazes de generalização dos dados, o que significa que os dados nos conjuntos de validação cruzada e testes devem estar no mesmo intervalo adotado pelo conjunto de treinamento. Para garantir que todos os dados estejam no mesmo intervalo

de variação devemos normalizar os dados de entrada antes da divisão dos conjuntos. A normalização é o processo de dimensionamento dos números em um conjunto de dados em um intervalo especificado. Além disso, as variáveis têm que ser normalizadas de tal maneira que seus valores sejam proporcionais aos limites das funções de ativação usada na camada de saída. (Valença) Neste trabalho foi aplicada a técnica de normalização Linear de acordo com a função abaixo, limitando os valores entre o intervalo [-1,1]:

$$Y_n = (b - a).\frac{(X_n - X_{min})}{(X_{max} - X_{min})} + a \tag{3.1}$$

Onde,

Y_n é o valor normalizado,

b é o limite máximo de normalização,

a é o limite mínimo de normalização

X_{max} é o valor máximo que a variável assume no conjunto

X_{min} é o valor mínimo que a variável assume no conjunto

3.1.2 Divisão dos dados

Para aplicação da análise de sensibilidade da base utilizada pela literatura é necessária variação na disposição dos dados apresentados durante as execuções das redes neurais artificiais. Assim sendo, a base de dados foi dividida em diversos conjuntos de porcentagens diferentes e os resultados de suas execuções foram utilizados como base para estatísticas durante a análise de sensibilidade.

Conforme citado anteriormente no item 3.1 as 47 amostras relativas à pacientes diagnosticados previamente com MCI foram incluídos no conjunto de testes de DA/MCI. O teste aplicado nesse conjunto identifica pessoas que adquiriram DA de 2 a 6 anos após o diagnóstico de MCI. Nenhuma das amostras de pacientes diagnosticados com MCI anos antes foi utilizada nas fazes de treinamento e validação cruzada da execução das redes.

Os 175 pacientes sem diagnóstico prévio de MCI foram divididos aleatoriamente nas porcentagens descritas na figura 6 excluindo-se o conjunto de Testes de MCI que permaneceu o mesmo durante as execuções. Por fim foram geradas as quantidades dos conjuntos de acordo com as descritas na figura 7. A figura 8 explica a sintaxe aplicada aos resultados (Taxas de acerto) relacionado à cada uma das divisões de dados, de acordo com a porcentagem da divisão dos conjuntos (treinamento, validação cruzada e testes de DA), o teste que foi aplicado(teste Da ou teste DA/MCI) e a base de dados utilizada (Lara-3 ou Lara-5) Foi desenvolvido o código Java para efetuar o pré-processamento dos dados, que será descrito no próximo item e gerar aleatoriamente todos os conjuntos descritos nas

figuras acima. O pseudocódigo do algoritmo está descrito no item 16.

	Proporções de acordo com cada sintaxe				
Conjuntos de entrada	**(50-25-25)**	**(60-30-10)**	**(60-20-20)**	**(80-10-10)**	**(40-20-40)**
Treinamento	50%	60%	60%	80%	40%
Validação Cruzada	25%	30%	20%	10%	20%
Testes DA	25%	10%	20%	10%	40%
Testes DA/MCI	O conjunto de DA/MCI não sofreu nenhuma variação				

Figura 6 – Porcentagens para divisão das assinaturas escolhidas para execução das redes neurais

	Proporções de acordo com cada sintaxe				
Conjuntos de entrada	**(50-25-25)**	**(60-30-10)**	**(60-20-20)**	**(80-10-10)**	**(40-20-40)**
Treinamento	87	105	105	140	71
Validação Cruzada	44	53	35	17	35
Testes DA	44	18	35	18	70
Testes DA/MCI	47	47	47	47	47

Figura 7 – Quantidades de exemplos em cada um dos conjuntos

Base	Proporção	Cenário de Teste	Rótulo
Lara-5	50-25-25	DA	Lara-5 (50-25-25) -DA
	60-30-10		Lara-5 (60-30-10) -DA
	60-20-20		Lara-5 (60-20-20) -DA
	80-10-10		Lara-5 (80-10-10) -DA
	40-20-40		Lara-5 (40-20-40) -DA
	50-25-25	DA/MCI	Lara-5 (50-25-25) -DA/MCID
	60-30-10		Lara-5 (60-30-10) -DA/MCI
	60-20-20		Lara-5 (60-20-20) -DA/MCI
	80-10-10		Lara-5 (80-10-10) -DA/MCI
	40-20-40		Lara-5 (40-20-40) -DA/MCI
Lara-3	50-25-25	DA	Lara-3 (50-25-25) -DA
	60-30-10		Lara-3 (60-30-10) -DA
	60-20-20		Lara-3 (60-20-20) -DA
	80-10-10		Lara-3 (80-10-10) -DA
	40-20-40		Lara-3 (40-20-40) -DA
	50-25-25	DA/MCI	Lara-3 (50-25-25) -DA/MCI
	60-30-10		Lara-3 (60-30-10) -DA/MCI
	60-20-20		Lara-3 (60-20-20) -DA/MCI
	80-10-10		Lara-3 (80-10-10) -DA/MCI
	40-20-40		Lara-3 (40-20-40) -DA/MCI

Figura 8 – Sintaxe aplicada para cada uma das divisões de conjuntos aplicados nas redes neurais. A primeira coluna indica de qual base de dados está sendo feita a divisão, a segunda coluna indica a proporção da divisão dos conjuntos de treinamento, validação cruzada e teste DA, a terceira coluna indica qual conjunto de teste foi aplicado e por fim a sintaxe aplicada à cada um dos resultados.

Algoritmo 1: Pseudocódigo de Pré-processamento e Divisão de dados

Entrada: arquivo csv da base de dados

Saída: arquivos csv referentes aos conjuntos (entrada e saída) da assinaturas selecionada

```
1  início
2      Lê o arquivo CSV da base de dados;
3      Normaliza os dados;
4      Define o conjunto de testes DA/MCI;
5      Define o conjunto restante sem os dados de teste DA/MCI;
6      Escreve em CSV o conjunto de testes DA/MCI;
7      Gera aleatoriedade no conjunto restante dos dados;
8      Divide o conjunto restante em dois conjuntos: entrada e saída
9      para cada conjunto de porcentagens da tabela faça
10         Carregar os tamanhos dos conjuntos de treinamento, validação e teste de
           acordo com a porcentagem definida;
11         para cada conjunto de dados (entrada e saída) faça
12             Define os conjuntos de Treinamento, Validação Cruzada e Testes de
               acordo com os tamanhos definidos
13             Escreve em CSVs os conjuntos;
14         fim
15     fim
16 fim
```

3.1.3 Assinaturas de proteínas

Conforme citado na seção 3.1 a base de dados utilizada nessa monografia foi comumente aplicada na literatura por Ray et Al [5], posteriormente pelos pesquisadores Gómez e Moscato [6] e por fim por Lara [15]. Cada um dos trabalhos procurou obter o conjunto de proteínas derivadas da base de 120 proteínas afim de alcançar uma melhor taxa de classificação para a doença de Alzheimer. Os trabalhos desenvolvidos por Ray et al resultaram em uma base de dados com 18 proteínas ao passo que Gómez e Moscato encontraram uma assinatura de 5 proteínas, um subconjunto da assinatura de 18 proteínas de Ray et al, que tem o mesmo desempenho geral. Em estudos mais recentes na Universidade de Pernambuco Lara encontrou uma assinatura com 3 proteínas que de acordo com seus testes estatísticos possui uma taxa de acerto aproximada ao obtido por Gómez e Moscato, porém resulta-se melhor devido à diminuição da quantidade de proteínas que causa diminuição nos custos de diagnóstico.

Para execução das variações das redes neurais foram utilizadas a assinatura de melhor desempenho obtida por Lara [15] com 3 proteínas do plasma e a segunda melhor

assinatura de melhor desempenho por Lara com 5 proteínas, visto que essas duas assinaturas superaram em termos de desempenho e quantidade de proteínas as demais assinaturas identificadas pela literatura. A figura 9 mostra a disposição das proteínas em cada uma das sub-bases.

Nome da Base	Proteínas
Lara-5	IL-1a, TNF-a, G-CSF , PDGFBB, M-CSF
Lara-3	IL-1a, TNF-a, G-CSF

Figura 9 – Assinaturas da base de dados utilizadas para execução do MLP e RC

3.2 Parametrização das Redes Neurais

O algoritmo de MLP e RC utilizados nesse trabalho foram implementados na linguagem JAVA no ambiente eclipse. Cada combinação entre as assinaturas selecionadas e as variações de conjuntos de entrada geradas foi executado 30 fazer para garantir a veracidade das informações. Afim de obter dados estatísticos de classificação para uma análise de sensibilidade foram colhidos a cada execução a taxa de acerto da rede, que representa a porcentagem de exemplos do conjunto de testes nos quais as redes fizeram uma classificação correta. O conceito de classificação correta das redes neurais foi abordado no item 2.2 desse trabalho.

A execuções da rede MLP foram iniciadas com as seguintes configurações:

- Quantidade de neurônios na camada de entrada: 5 ou 3 (a depender da assinatura de proteínas que estiver sendo utilizada);
- Quantidade de neurônios na camada escondida: 8;
- Quantidade de neurônios na camada de saída: 2;
- Função de ativação dos neurônios da camada escondida: Tangente Hiperbólica;
- Função de ativação dos neurônios da camada de saída: Linear;
- Taxa de aprendizagem 0,7;
- Taxa de momento 0,4;

A execuções da rede RC foram iniciadas com as seguintes configurações:

- Quantidade de neurônios na camada de entrada: 5 ou 3 (a depender da assinatura de proteínas que estiver sendo utilizada);
- Quantidade de neurônios no reservatório: 4;
- Quantidade de neurônios na camada de saída: 2;
- Taxa de conectividade: 20%;
- Quantidades de ciclos no Warm up: 100;
- Função de ativação no Reservatório: Tangente Hiperbólica;
- Função de ativação na Camada de Saída: Linear;

3.3 Testes estatísticos

3.3.1 Teste de Shapiro-Wilk

Toda variável aleatória assume uma determinada distribuição de frequências na população, que podem ter formas variadas. Na literatura estatística encontra-se muitas distribuições teóricas. Essas são modelos que procuram representar o comportamento de determinado evento em função da frequência de sua ocorrência. No caso das variáveis contínuas, esse evento será um intervalo de valores. As distribuições de frequências são, na verdade, distribuições de probabilidade, onde para um evento teremos uma probabilidade de ocorrência associada. Em outras palavras, a partir de uma distribuição de probabilidade completamente especificada, pode-se calcular a probabilidade de uma variável aleatória assumir determinado intervalo de valores. Para aplicação dos testes paramétricos t nessas variáveis é necessária a suposição de normalidade da(s) variável(is) aleatória(s) [16].

Para suposição da normalidade da taxa de acerto das redes neurais vamos utilizar o teste desenvolvido em 1965 por Shapiro Wilk [17] responsável por gerar um *p-value* onde é possível assumir a hipótese de que a o conjunto de dados é normalmente distribuído.

Para execução do teste Shapior-Wilk vamos definir a hipóteses:

- Hipótese nula H_0, de que a variável aleatória adere à distribuição normal
- Hipótese alternativa H_1, de que a variável aleatória não adere à distribuição normal.

Foi adotado então que, se $p - valor < 0.05$, então reijeitamos H_0, ou seja podemos afirmar com nível de significância de 5% que a amostra não provém de uma população normal.

3.3.2 Teste F

O teste F é um teste não-paramétrico utilizado para identificar se há diferenças estatisticamente significativas entre os desvios-padrão de duas amostras de populações Normais independentes, ou seja, se elas são ou não homocedásticas. Sendo assim é possível identificar se as variâncias entre dois conjuntos de amostras eram estatisticamente iguais.

Para execução do teste Shapior-Wilk vamos definir a hipóteses:

- Hipótese nula H_0, de que as amostras possuem variâncias estatisticamente iguais.
- Hipótese alternativa H_1, de que as amostras não possuim variâncias estatisticamente iguais

Da mesma forma que no teste anterior, podemos inferir que se $p\text{-}value < 0.05$, então reijeitamos H_0, ou seja podemos afirmar com nível de significância de 5% que as amostras não possuem variâncias estatisticamente iguais.

3.3.3 Teste t-Student

O teste t-Student foi desenvolvido pelo pesquisador Irlandês William Sealy Gosset, e é utilizado para avaliar a diferença entre as médias de duas amostras. Para aplicação desse teste as duas amostras devem prover de uma distribuição normal e possuir variâncias estatisticamente iguais. Estes dois critérios foram verificados nesse trabalho utilizando o teste de Shapiro-Wilk e o Teste F de Snedecor.

Semelhante aos testes descritos anteriormente, para interpretação do teste de t-Student também se faz necessário definição das hipóteses para o problema. São elas:

- Hipótese nula H_0, de que as amostras possuem médias estatisticamente iguais.
- Hipótese alternativa H_1, de que as amostras não possuem médias estatisticamente iguais

3.3.4 Teste de Wilcoxon

O teste Wilcoxon é um teste não paramétrico, ou seja, não faz nenhuma hipótese em relação a distribuição das probabilidades dos dados. Este teste tem o objetivo de verificar que duas amostras mesmo sendo de amostras independentes, elas fazem parte de uma população onde as medianas são iguais [18].

Também se faz necessário definição das hipóteses seguintes:

- Hipótese nula H_0, de que as amostras possuem medianas estatisticamente iguais.

- Hipótese alternativa H_1, de que as amostras não possuim medianas estatisticamente iguais

3.3.5 Gráficos de Box Plot

O Box Plot, também conhecido como Gráfico de caixa é utilizado para avaliar a distribuição empírica dos dados. Fornece informação sobre as seguintes características do conjunto de dados: localização, dispersão, assimetria, comprimento da cauda e medidas discrepantes. A dispersão é representada pela amplitude do gráfico, que pode ser calculada como máximo valor – mínimo valor. Quanto maior for a amplitude, maior a variação nos dados. O Gráfico é formado pelos seguintes parâmetros:

- Mediana: A linha preta dentro da caixa indica o valor da mediana dos dados
- Caixa da amplitude interquartílica: A caixa de amplitude interquartílica representa a metade, 50% dos dados. Ela mostra a distância entre o primeiro e o terceiro quartis (Q3-Q1)
- limites: São representados pelos traços se estendem de ambos os lados da caixa. Os traços representam a amplitude dos dados, o limite máximo e mínimo
- outliers: São representados por pontos fora dos limites que são considerados valores discrepantes da amostra.

Quando os dados são assimétricos, a maior parte dos dados está localizada no lado superior ou inferior do gráfico. A assimetria indica que os dados podem não ser normalmente distribuídos [19].

A Figura 10 a seguir apresenta um exemplo do formato de um boxplot:

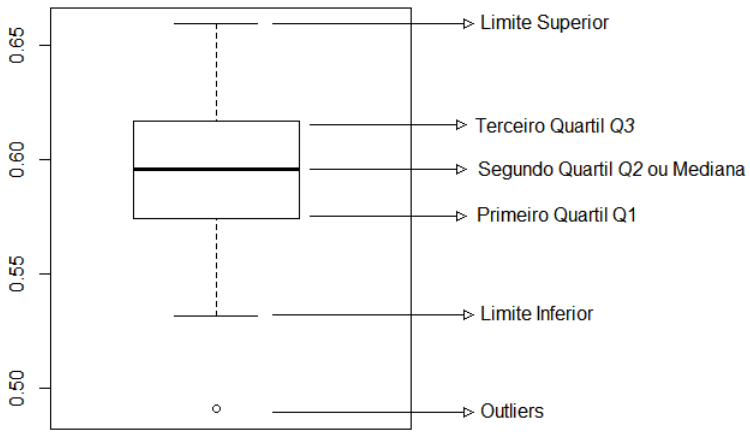


Figura 10 – Exemplo de Gráfico Box Plot

4 Resultados

As redes MLP e RC foram executadas 30 vezes para cada uma das bases geradas e foi colhida a informação referente à taxa de acerto das redes após os dois testes aplicados à cada uma das execuções, Testes de DA, para diagnóstico de DA em pessoas que não possuíam nenhum diagnóstico anterior e Testes de DA/MCI para diagnóstico de DA em pessoas anteriormente identificadas com MCI.

Nesses capítulo são descritos os testes estatísticos aplicados depois da execução das redes Neurais para análise de sensibilidade da base de dados. Os testes de t-Student e Wilcoxon foram adotados para avaliar o efeito causado pela variação dos tamanhos conjuntos sobre a taxa de acerto das redes neurais. Pretendeu-se analisar a diferença de performance alcançada pelos testes das redes de acordo com a quantidade de amostras definida nos conjuntos de entrada, ou seja, amostras que resultaram em médias ou medianas estatisticamente iguais não sofreram efeito negativo à variabilidade do tamanho dos conjuntos de entrada. Continuando a análise foram gerados Gráficos Box Plot afim de avaliar a assimetria, dispersão, amplitude e variabilidade dos dados. Os testes de Shapiro-Wilk e F de Snedecor foram aplicados com objetivo de definir qual teste seria utilizado para comparação de performance das amostras, paramétrico ou não-paramétrico.

4.1 Aplicação dos testes de Shapiro-Wilk

4.1.1 Em amostras provenientes dos testes de DA na base de dados Lara-5

As redes neurais MLP e RC foram executadas 30 vezes para as cinco bases geradas a partir da assinatura de Lara-5 divididas de acordo com as porcentagens definidas na figura 6. Os resultados do *p-value* do teste Shapiro-Wilk estão descritos nas figuras 11 e 12.

Base	Shapiro-Wilk	Desvio Padrão
Lara-5 (50-25-25) - DA	0,001502	0,147913
Lara-5 (60-30-10) - DA	3,06E-03	0,1471189
Lara-5 (60-20-20) - DA	1,19E-05	0,1100572
Lara-5 (80-10-10) - DA	1,48E-05	0,08956084
Lara-5 (40-20-40) - DA	6,94E-05	0,1497797

Figura 11 – Resultado do *p-value* encontrado após a aplicação do teste de Shapiro-Wilk para as amostras da MLP na base Lara-5

Todos os resultados apresentados possuem $p\text{-}value < 0.05$, então podemos determinar que nenhuma das execuções das redes MLP ou RC pode ser considerada normalmente

Base	Shapiro-Wilk	Desvio Padrão
Lara-5 (50-25-25) - DA	0,0006844	0,08529118
Lara-5 (60-30-10) - DA	2,65E-05	0,06015448
Lara-5 (60-20-20) - DA	0,002905	0,0819708
Lara-5 (80-10-10) - DA	0,02412	0,1017967
Lara-5 (40-20-40) - DA	0,009561	0,06210023

Figura 12 – Resultado do *p-value* encontrado após a aplicação do teste de Shapiro-Wilk para as amostras da RC na base Lara-5

distribuída. Dessa forma todos os testes comparativos foram gerados com o teste de Wilcoxon para comparação das medianas das amostras analisadas. Como o teste é não-paramétrico não foi necessário aplicação do Teste F de vericação da possibilidade de aplicação de um teste paramétrico.

4.1.2 Em amostras provenientes dos testes de DA na base de dados Lara-3

Dando continuidade aos testes estatísticos aplicados para o conjunto de Testes de DA, as redes neurais MLP e RC foram executadas 30 vezes para as cinco bases geradas a partir da base Lara-3 divididas de acordo com as porcentagens definidas na figura 6. Os resultados estão ilustrados nas figuras 13 e 14.

Base	Shapiro-Wilk	Desvio Padrão
Lara-3 (50-25-25) - DA	0,007369	0,1126307
Lara-3 (60-30-10) - DA	0,005001	0,1148169
Lara-3 (60-20-20) - DA	0,003216	0,1491866
Lara-3 (80-10-10) - DA	0,002735	0,096059
Lara-3 (40-20-40) - DA	0,08917	0,1271505

Figura 13 – Resultado do *p-value* encontrado após a aplicação do teste de Shapiro-Wilk para as amostras da MLP na base Lara-3

Base	Shapiro-Wilk	Desvio Padrão
Lara-3 (50-25-25) - DA	0,009687	0,06006477
Lara-3 (60-30-10) - DA	9,63E-06	0,06645881
Lara-3 (60-20-20) - DA	3,96E-06	0,06699894
Lara-3 (80-10-10) - DA	0,2424	0,07768193
Lara-3 (40-20-40) - DA	0,001343	0,03689827

Figura 14 – Resultado do *p-value* encontrado após a aplicação do teste de Shapiro-Wilk para as amostras da RC na base Lara-3

Dentre as execuções do MLP apenas a Lara-3(40-20-20)-DA provém de uma população normal, ao passo que para as execuções do RC apenas Lara-3(80-10-10)-DA é

considerada normalmente distribuída, não podendo ser comparadas com nenhuma outra amostra da base. Sendo assim, foi aplicado o teste de Wilcoxon para comparação das taxas de acertos entre as diferentes bases de uma mesma assinatura. Não foi necessário aplicação do teste F para análise estatística de variâncias.

4.1.3 Em amostras provenientes dos testes de DA/MCI na base de dados Lara-5

Em sequência foram aplicados os testes estatísticos para os testes das redes neurais na base de dados de DA/MCI. Esse conjunto pretende identificar DA em pacientes que foram diagnosticados com MCI anos antes do diagnóstico atual. Para os testes do conjunto de Testes de DA/MCI nas diferentes porcentagens foram utilizados os mesmos conjuntos de treinamento e validação aplicados nos testes de DA como descritos na sessão anterior. Ou seja, durante a execução do mesmo conjunto de treinamento e validação foram efetuados dois testes, um para DA e outro para MCI.

As figuras 15 e 16 mostram os resultados dos testes de Shapiro-Wilk para execução desses testes.

Base	Shapiro-Wilk	Desvio Padrão
Lara-5 (50-25-25) - DA/MCI	0,01878	0,05745451
Lara-5 (60-30-10) - DA/MCI	0,05109	0,067994
Lara-5 (60-20-20) - DA/MCI	0,000001019	0,07857024
Lara-5 (80-10-10) - DA/MCI	4,107E-07	0,08013416
Lara-5 (40-20-40) - DA/MCI	0,00001691	0,06139482

Figura 15 – Resultado do *p-value* encontrado após a aplicação do teste de Shapiro-Wilk para as amostras do conjunto de testes de DA/MCI da execução da MLP na base Lara-5

Base	Shapiro-Wilk	Desvio Padrão
Lara-5 (50-25-25) - DA/MCI	0,01778	0,04894521
Lara-5 (60-30-10) - DA/MCI	0,3944	0,05155971
Lara-5 (60-20-20) - DA/MCI	0,04818	0,04016922
Lara-5 (80-10-10) - DA/MCI	0,03456	0,04158848
Lara-5 (40-20-40) - DA/MCI	0,09894	0,0494739

Figura 16 – Resultado do *p-value* encontrado após a aplicação do teste de Shapiro-Wilk para as amostras do conjunto de testes de DA/MCI da execução da RC na base Lara-5

Os resultados mostram que para os testes nas execuções de MLP apenas os testes realizados na base dividida em Lara-5(60-20-20)-DA/MCI provém de uma população

normal. Não podendo ser comparada com nenhum outro conjunto de porcentagens da base. Dessa maneira, todas as comparações entre as possíveis divisões de conjuntos da base de dados foram feitas utilizando o teste de Wilcoxon, não sendo necessária aplicação do teste F.

Ao analisarmos, por sua vez, o resultado dos testes para a rede RC, observamos que, os resultados dos testes Lara-5(60-30-10)-DA/MCI e Lara-5(40-20-20)-DA/MCI obtiveram *p-value* > 0.05, o que indica que as amostras podem ser consideradas normalmente distribuídas.

Para dar continuidade ao comparativo entre essas duas amostras foi aplicado o Teste F afim de verificar se ambas possuem variância estatisticamente igual. O *p-value* resultante do teste F para essas duas amostras foi de 0,8255, indicando com uma taxa de significância de 5% que as amostras possuem variâncias estatisticamente iguais. A vista destes dois testes, foi possível aplicação do teste paramétrico de t-Student para comparação entre as amostras Lara-5(60-30-10)-DA/MCI e Lara-5(40-20-20)-DA/MCI. Para as demais comparações foi aplicado o teste não paramétrico de Wilcoxon.

4.1.4 Em amostras provenientes dos testes de DA/MCI na base de dados Lara-3

Por fim, foram aplicados os testes estatísticos nas taxas de acerto geradas pelo conjunto de teste DA/MCI para as possíveis divisões da base Lara-3. Os resultados para cada uma das redes estão descritos nas figuras 17 e 18:

Base	Shapiro-Wilk	Desvio Padrão
Lara-3 (50-25-25) - DA/MCI	0,05481	0,05899607
Lara-3 (60-30-10) - DA/MCI	0,004638	0,03830602
Lara-3 (60-20-20) - DA/MCI	0,05707287	0,06278
Lara-3 (80-10-10) - DA/MCI	0,0009233	0,04883878
Lara-3 (40-20-40) - DA/MCI	0,03101	0,05613522

Figura 17 – Resultado do *p-value* encontrado após a aplicação do teste de Shapiro-Wilk para as amostras do conjunto de testes de DA/MCI da execução da MLP na base Lara-3

As figuras mostram que dos testes de Shapiro-Wilk aplicados nas amostras de execução do MLP apenas os Lara-3(50-25-25)-DA/MCI e Lara-3(60-20-20)-DA/MCI podem ser considerados como pertencentes à uma distribuição normal.

Para a comparação entre essas duas amostras foi obtido *p-value* de 0,8596 durante a aplicação do Teste F, indicando que as amostras possuem variâncias estatisticamente iguais, sendo possível a aplicação do teste paramétrico de t-student. As demais comparações entre as bases foram testadas com a aplicação do teste de Wilcoxon.

Base	Shapiro-Wilk	Desvio Padrão
Lara-3 (50-25-25) - DA/MCI	0,05619	0,03703524
Lara-3 (60-30-10) - DA/MCI	0,1035	0,03703524
Lara-3 (60-20-20) - DA/MCI	0,0514	0,03475202
Lara-3 (80-10-10) - DA/MCI	0,02194	0,03516137
Lara-3 (40-20-40) - DA/MCI	0,003178	0,03898597

Figura 18 – Resultado do *p-value* encontrado após a aplicação do teste de Shapiro-Wilk para as amostras do conjunto de testes de DA/MCI da execução da RC na base Lara-3

4.2 Testes comparativos t-Student e Wilcoxon

Os testes de t-Student e Wilcocoxon foram aplicados para todas as combinações de bases de dados geradas para uma mesma assinatura. A escolha do teste à ser adotada foi definida de acordo com os resultados dos testes do item 4.1.

4.2.1 Em amostras provenientes dos testes de DA na base de dados Lara-5

Nº	Amostra Base 1	Amostra Base 2	t-Student	Wilcoxon
1	Lara-5 (50-25-25) - DA	Lara-5 (60-30-10) - DA		0,1674
2	Lara-5 (50-25-25) - DA	Lara-5 (60-20-20) - DA		9,53E-01
3	Lara-5 (50-25-25) - DA	Lara-5 (80-10-10) - DA		0,3027
4	Lara-5 (50-25-25) - DA	Lara-5 (40-20-40) - DA		0,04182
5	Lara-5 (60-30-10) - DA	Lara-5 (60-20-20) - DA		2,60E-01
6	Lara-5 (60-30-10) - DA	Lara-5 (80-10-10) - DA		0,05627
7	Lara-5 (60-30-10) - DA	Lara-5 (40-20-40) - DA		1,97E-03
8	Lara-5 (60-20-20) - DA	Lara-5 (80-10-10) - DA		9,13E-02
9	Lara-5 (60-20-20) - DA	Lara-5 (40-20-40) - DA		0,001471
10	Lara-5 (80-10-10) - DA	Lara-5 (40-20-40) - DA		0,002389

Figura 19 – Resultado do *p-value* encontrado após a aplicação dos testes de t-Student e Wilcoxon para as amostras de testes DA da execução da MLP na base Lara-5

Como pode ser visto nas figuras 19 e 20, os resultados dos testes de Wilcoxon para as amostras de Teste DA da base Lara-5 indicaram uma diferença estatística entre 4 das 10 comparações de amostras da MLP e 6 das 10 comparações de amostras da MLP. Podemos inferir que 40% e 60% por cento das comparações feitas nas amostras provindas da mesma base de dados e dos mesmos parâmetros de execução não são estatísticamente iguais.

Nº	Amostra Base 1	Amostra Base 2	t-Student	Wilcoxon
1	Lara-5 (50-25-25) - DA	Lara-5 (60-30-10) - DA		0,00024
2	Lara-5 (50-25-25) - DA	Lara-5 (60-20-20) - DA		0,3224
3	Lara-5 (50-25-25) - DA	Lara-5 (80-10-10) - DA		0,8124
4	Lara-5 (50-25-25) - DA	Lara-5 (40-20-40) - DA		3,15E-06
5	Lara-5 (60-30-10) - DA	Lara-5 (60-20-20) - DA		0,01153
6	Lara-5 (60-30-10) - DA	Lara-5 (80-10-10) - DA		0,002055
7	Lara-5 (60-30-10) - DA	Lara-5 (40-20-40) - DA		0,06203
8	Lara-5 (60-20-20) - DA	Lara-5 (80-10-10) - DA		0,1654
9	Lara-5 (60-20-20) - DA	Lara-5 (40-20-40) - DA		0,005345
10	Lara-5 (80-10-10) - DA	Lara-5 (40-20-40) - DA		0,000105

Figura 20 – Resultado do *p-value* encontrado após a aplicação dos testes de t-Student e Wilcoxon para as amostras de testes DA da execução do RC na base Lara-5

Nº	Amostra Base 1	Amostra Base 2	t-Student	Wilcoxon
1	Lara-3 (50-25-25) - DA	Lara-3 (60-30-10) - DA		0,004784
2	Lara-3 (50-25-25) - DA	Lara-3 (60-20-20) - DA		6,52E-01
3	Lara-3 (50-25-25) - DA	Lara-3 (80-10-10) - DA		0,1074
4	Lara-3 (50-25-25) - DA	Lara-3 (40-20-40) - DA		0,07452
5	Lara-3 (60-30-10) - DA	Lara-3 (60-20-20) - DA		1,88E-02
6	Lara-3 (60-30-10) - DA	Lara-3 (80-10-10) - DA		0,2114
7	Lara-3 (60-30-10) - DA	Lara-3 (40-20-40) - DA		4,45E-01
8	Lara-3 (60-20-20) - DA	Lara-3 (80-10-10) - DA		2,25E-01
9	Lara-3 (60-20-20) - DA	Lara-3 (40-20-40) - DA		0,06485
10	Lara-3 (80-10-10) - DA	Lara-3 (40-20-40) - DA		0,6179

Figura 21 – Resultado do *p-value* encontrado após a aplicação dos testes de t-Student e Wilcoxon para as amostras de testes DA da execução da MLP na base Lara-3

4.2.2 Em amostras provenientes dos testes de DA na base de dados Lara-3

As figuras 21 e 22 ilustram os resultados dos testes de Wilcoxon para as amostras de Teste DA da base Lara-3. Os *p-values* resultantes dos testes indicaram uma diferença estatística entre apenas 2 das 10 comparações de amostras da MLP e 8 das 10 comparações de amostras da RC. No caso da sub base de dados em questão, para execução do MLP a variância dos tamanhos de conjuntos de entrada da rede não interfere tanto quanto para RC que possui cerca de 80% dos seus resultados estatísticos divergentes.

Nº	Amostra Base 1	Amostra Base 2	t-Student	Wilcoxon
1	Lara-3 (50-25-25) - DA	Lara-3 (60-30-10) - DA		0,00024
2	Lara-3 (50-25-25) - DA	Lara-3 (60-20-20) - DA		0,3224
3	Lara-3 (50-25-25) - DA	Lara-3 (80-10-10) - DA		0,8124
4	Lara-3 (50-25-25) - DA	Lara-3 (40-20-40) - DA		3,15E-06
5	Lara-3 (60-30-10) - DA	Lara-3 (60-20-20) - DA		0,01153
6	Lara-3 (60-30-10) - DA	Lara-3 (80-10-10) - DA		0,002055
7	Lara-3 (60-30-10) - DA	Lara-3 (40-20-40) - DA		0,06203
8	Lara-3 (60-20-20) - DA	Lara-3 (80-10-10) - DA		0,1654
9	Lara-3 (60-20-20) - DA	Lara-3 (40-20-40) - DA		0,005345
10	Lara-3 (80-10-10) - DA	Lara-3 (40-20-40) - DA		0,000105

Figura 22 – Resultado do *p-value* encontrado após a aplicação dos testes de t-Student e Wilcoxon para as amostras amostras de testes DA da execução do RC na base Lara-3

Nº	Amostra Base 1	Amostra Base 2	t-Student	Wilcoxon
1	Lara-5 (50-25-25) - DA/MCI	Lara-5 (60-30-10) - DA/MCI		0,7483
2	Lara-5 (50-25-25) - DA/MCI	Lara-5 (60-20-20) - DA/MCI		4,06E-03
3	Lara-5 (50-25-25) - DA/MCI	Lara-5 (80-10-10) - DA/MCI		0,00604
4	Lara-5 (50-25-25) - DA/MCI	Lara-5 (40-20-40) - DA/MCI		8,96E-05
5	Lara-5 (60-30-10) - DA/MCI	Lara-5 (60-20-20) - DA/MCI		4,83E-03
6	Lara-5 (60-30-10) - DA/MCI	Lara-5 (80-10-10) - DA/MCI		0,005584
7	Lara-5 (60-30-10) - DA/MCI	Lara-5 (40-20-40) - DA/MCI		3,23E-04
8	Lara-5 (60-20-20) - DA/MCI	Lara-5 (80-10-10) - DA/MCI		7,45E-01
9	Lara-5 (60-20-20) - DA/MCI	Lara-5 (40-20-40) - DA/MCI		0,9869
10	Lara-5 (80-10-10) - DA/MCI	Lara-5 (40-20-40) - DA/MCI		0,9069

Figura 23 – Resultado do *p-value* encontrado após a aplicação dos testes de t-Student e Wilcoxon para as amostras de testes DA/MCI da execução da MLP na base Lara-5

4.2.3 Em amostras provenientes dos testes de DA/MCI na base de dados Lara-5

Os resultados dos testes de Wilcoxon para as amostras de Teste DA/MCI da base Lara-5 conforme descritos nas figuras 23 e 24 indicaram uma diferença estatística entre apenas 6 das 10 comparações de amostras da MLP e 3 das 10 comparações de amostras da MLP. Apesar da variação nos tamanhos do conjunto de treinamento e validação, o conjunto de DA/MCI manteve-se o mesmo durante toda os testes, sendo assim esperava-se uma aproximação maior entre os resultados das variâncias do que o que foi encontrado para o conjunto de testes DA.

Nº	Amostra Base 1	Amostra Base 2	t-Student	Wilcoxon
1	Lara-5 (50-25-25) - DA/MCI	Lara-5 (60-30-10) - DA/MCI		0,05962
2	Lara-5 (50-25-25) - DA/MCI	Lara-5 (60-20-20) - DA/MCI		0,7196
3	Lara-5 (50-25-25) - DA/MCI	Lara-5 (80-10-10) - DA/MCI		0,3115
4	Lara-5 (50-25-25) - DA/MCI	Lara-5 (40-20-40) - DA/MCI		0,03623
5	Lara-5 (60-30-10) - DA/MCI	Lara-5 (60-20-20) - DA/MCI		0,01347
6	Lara-5 (60-30-10) - DA/MCI	Lara-5 (80-10-10) - DA/MCI		0,3345
7	Lara-5 (60-30-10) - DA/MCI	Lara-5 (40-20-40) - DA/MCI	0.871	
8	Lara-5 (60-20-20) - DA/MCI	Lara-5 (80-10-10) - DA/MCI		0,07356
9	Lara-5 (60-20-20) - DA/MCI	Lara-5 (40-20-40) - DA/MCI		0,002616
10	Lara-5 (80-10-10) - DA/MCI	Lara-5 (40-20-40) - DA/MCI		0,1565

Figura 24 – Resultado do *p-value* encontrado após a aplicação dos testes de t-Student e Wilcoxon para as amostras amostras de testes DA/MCI da execução do RC na base Lara-5

4.2.4 Em amostras provenientes dos testes de DA/MCI na base de dados Lara-3

As figuras 25 e 26 contém os valores referentes ao *p-value* gerados pelos testes estatísticos nas duas redes para a base de dados Lara-3. Este foi o único caso em que todas as comparações apresentaram $p\text{-}value > 0.05$, sendo assim consideradas sua médias, no caso do T-student, e medianas, no caso do teste de Wilcoxon, estatisticamente iguais. Podemos inferir por esse resultado que a variabilidade da disposição dos dados nos conjuntos de entrada das redes MLP e RC não afeta o desempenho desta assinatura de dados na classificação da doença em pacientes que já foram diagnosticados com MCI.

Nº	Amostra Base 1	Amostra Base 2	t-Student	Wilcoxon
1	Lara-3 (50-25-25) - DA/MCI	Lara-3 (60-30-10) - DA/MCI		0,8686
2	Lara-3 (50-25-25) - DA/MCI	Lara-3 (60-20-20) - DA/MCI	0,3723	
3	Lara-3 (50-25-25) - DA/MCI	Lara-3 (80-10-10) - DA/MCI		0,8918
4	Lara-3 (50-25-25) - DA/MCI	Lara-3 (40-20-40) - DA/MCI		0,8116
5	Lara-3 (60-30-10) - DA/MCI	Lara-3 (60-20-20) - DA/MCI		1,22E-01
6	Lara-3 (60-30-10) - DA/MCI	Lara-3 (80-10-10) - DA/MCI		0,9637
7	Lara-3 (60-30-10) - DA/MCI	Lara-3 (40-20-40) - DA/MCI		7,09E-01
8	Lara-3 (60-20-20) - DA/MCI	Lara-3 (80-10-10) - DA/MCI		2,43E-01
9	Lara-3 (60-20-20) - DA/MCI	Lara-3 (40-20-40) - DA/MCI		0,2797
10	Lara-3 (80-10-10) - DA/MCI	Lara-3 (40-20-40) - DA/MCI		0,9404

Figura 25 – Resultado do *p-value* encontrado após a aplicação dos testes de t-Student e Wilcoxon para as amostras da MLP na base Lara-3

Nº	Amostra Base 1	Amostra Base 2	t-Student	Wilcoxon
1	Lara-3 (50-25-25) - DA/MCI	Lara-3 (60-30-10) - DA/MCI	1	
2	Lara-3 (50-25-25) - DA/MCI	Lara-3 (60-20-20) - DA/MCI	0,7341	
3	Lara-3 (50-25-25) - DA/MCI	Lara-3 (80-10-10) - DA/MCI	0,7341	
4	Lara-3 (50-25-25) - DA/MCI	Lara-3 (40-20-40) - DA/MCI		0,455
5	Lara-3 (60-30-10) - DA/MCI	Lara-3 (60-20-20) - DA/MCI		0,6842
6	Lara-3 (60-30-10) - DA/MCI	Lara-3 (80-10-10) - DA/MCI		0,5934
7	Lara-3 (60-30-10) - DA/MCI	Lara-3 (40-20-40) - DA/MCI		0,489
8	Lara-3 (60-20-20) - DA/MCI	Lara-3 (80-10-10) - DA/MCI		0,3389
9	Lara-3 (60-20-20) - DA/MCI	Lara-3 (40-20-40) - DA/MCI		0,2525
10	Lara-3 (80-10-10) - DA/MCI	Lara-3 (40-20-40) - DA/MCI		0,8858

Figura 26 – Resultado do *p-value* encontrado após a aplicação dos testes de t-Student e Wilcoxon para as amostras da RC na base Lara-3

4.3 Gráficos Box Plot

4.3.1 Para as amostras provenientes dos testes de DA na base de dados Lara-5

O gráfico 27 indica que a maioria dos dados das amostras Lara-5(50-25-25)-DA e Lara-5(60-30-10)-DA está abaixo de 80% de desempenho, em contrapartida a maioria dos dados das amostras Lara-5(60-20-20)-DA e Lara-5(80-10-10)-DA está acima de 65%.

De forma mais atenuante,de acordo com o gráfico 28, para as execuções na rede neural RC a divergência na mediana das amostras, na amplitude e também na distribuição dos dados foram consideradas ainda mais significativas. Em casos como a amostra Lara-5(40-20-40)-DA temos uma amplitude de 80% a 90% ao passo que na amostra Lara-5(80-10-10)-DA temos uma amplitude de 55% a 90% aproximadamente. Podemos ainda observar a presença de valores discrepantes em todos os as amostras do Box Plot, que podem ser resultado de erros de entrada de dados ou de medição. O gráfico infere que para a mesma base de dados Lara-5, variando-se a quantidade de exemplos nos conjuntos de entrada, obtém-se uma grande variação no desempenho da rede RC para testes de DA.

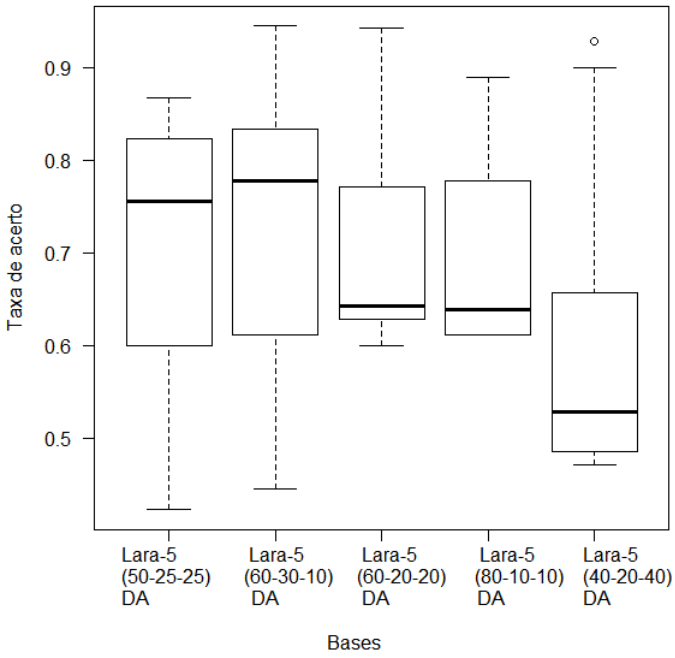


Figura 27 – Gráfico de Box Plot das amostras de testes DA da execução da MLP na base Lara-5

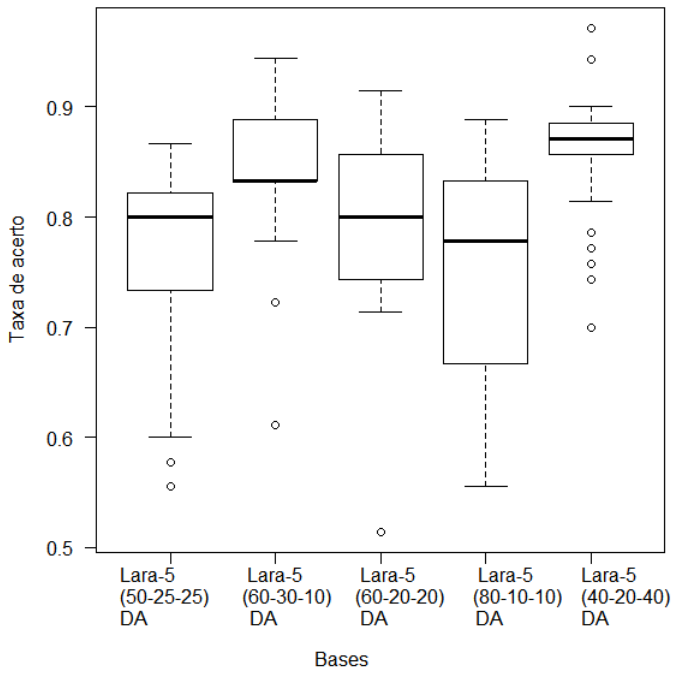


Figura 28 – Gráfico de Box Plot das amostras de testes DA da execução da RC na base Lara-5

4.3.2 Para as amostras provenientes dos testes de DA na base de dados Lara-3

No gráfico 29 observa-se uma diferença mais sutil entre a amplitude dos resultados do MLP, porém a mesma foi considerada muito grande. Os limites do box plot variam de

40% a 95% e suas medianas assumem valores diferentes a cada variação do conjunto.

Em discordância do que foi observado para a rede MLP, as taxas de acerto para as execuções do RC obtiveram em sua maioria baixa amplitude, o que é caracteriza o cenário ideal. Porém ainda é possível observar uma grande quantidade de taxas de acerto identificadas entre 80% e 70% nas amostras Lara-3(50-25-25)-DA, Lara-3(60-30-30)-DA e Lara-3(80-10-10)-DA e uma concentração de taxa de acerto entre 80 e 90% para as amostras Lara-3(60-20-20)-DA e Lara-3(40-20-20)-DA, o que evidencia mais uma vez a variação no desempenho para das execuções para uma mesma base de dados. As médias das taxas de acerto também apresentam grande divergência para esse caso. Para Lara-3(60-30-10)-DA a média foi de 75% e para Lara-3(60-20-20)-DA a média atingida foi de 87,2%, alcançando uma diferença de mais de 10

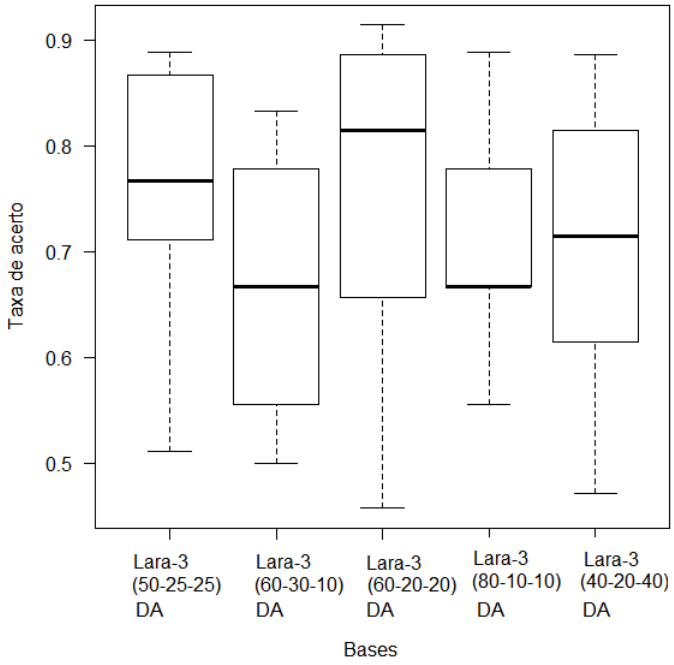


Figura 29 – Gráfico de Box Plot das amostras de testes DA da execução da MLP na base Lara-3

4.3.3 Para as amostras provenientes dos testes de DA/MCI na base de dados Lara-5

Analogamente à análise dos gráficos anteriores, é possível identificar também no Box Plot 31 e 32 para execução de MLP e RC dos testes de DA/MCI para base de 5 proteínas uma grande variação na amplitude e assimetria dos dados. A base de dados é

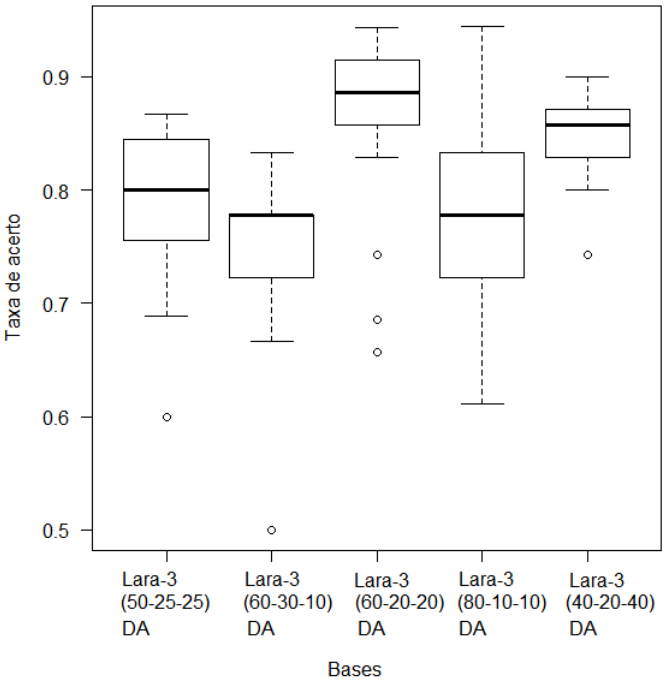


Figura 30 – Gráfico de Box Plot das amostras de testes DA da execução da RC na base Lara-3

então considerada sensível a variação dos conjuntos de dados de entrada, afetando o testes de diagnóstico de DA em pacientes previamente diagnosticados com MCI.

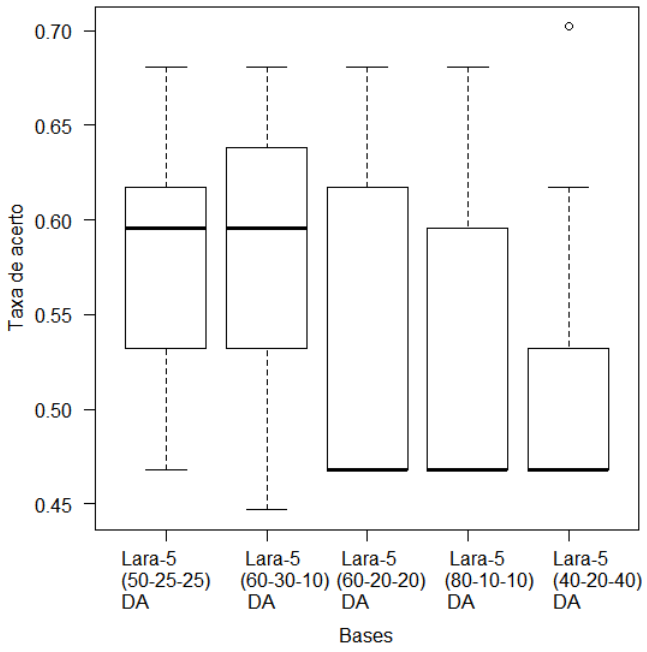


Figura 31 – Gráfico de Box Plot das amostras de testes DA/MCI da execução da MLP na base Lara-5

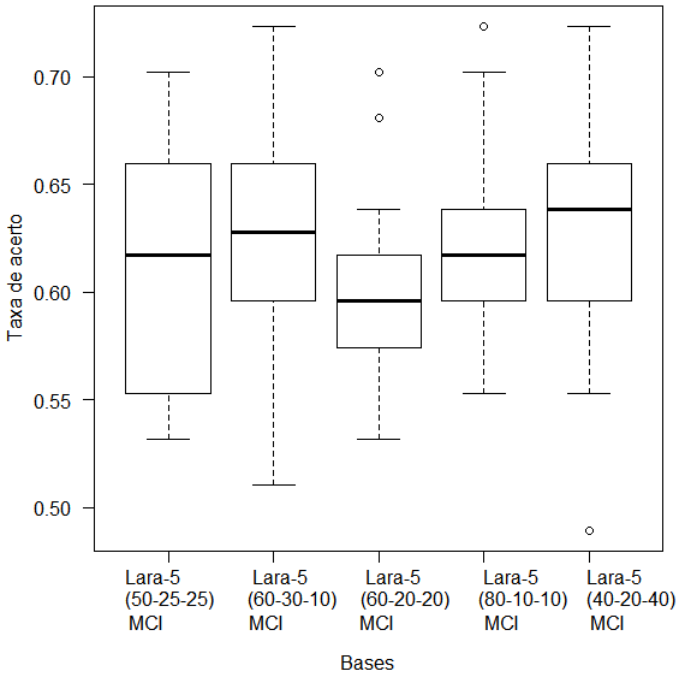


Figura 32 – Gráfico de Box Plot das amostras de testes DA/MCI da execução da RC na base Lara-5

4.3.4 Para as amostras provenientes dos testes de DA/MCI na base de dados Lara-3

Os dois casos ilustrados nos gráficos 33 e 34,os testes de DA/MCI para execução de MLP e RC para base Lara-3, apresentaram medianas estatisticamente iguais, bem como amplitude e dispersão dos dados similares. Podemos inferir que para os testes de DA/MCI a base de dados em questão não sofreu efeito negativo à variação dos conjuntos de entrada. Sendo esse o único resultado positivo apresentado nesse trabalho.

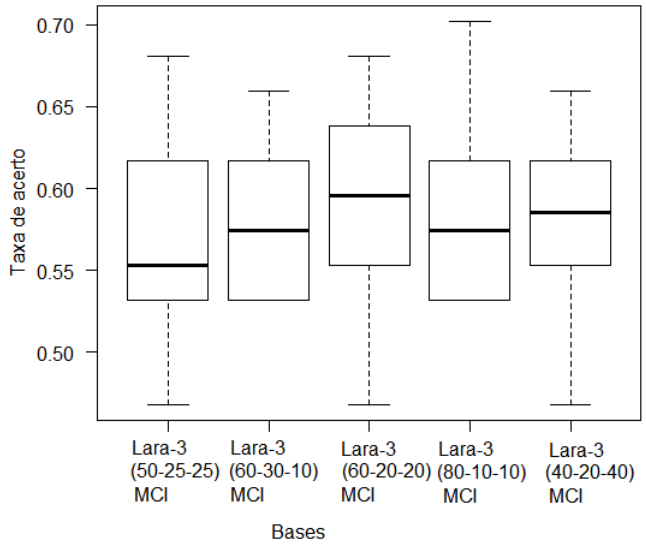


Figura 33 – Gráfico de Box Plot das amostras de testes DA/MCI da execução da MLP na base Lara-3

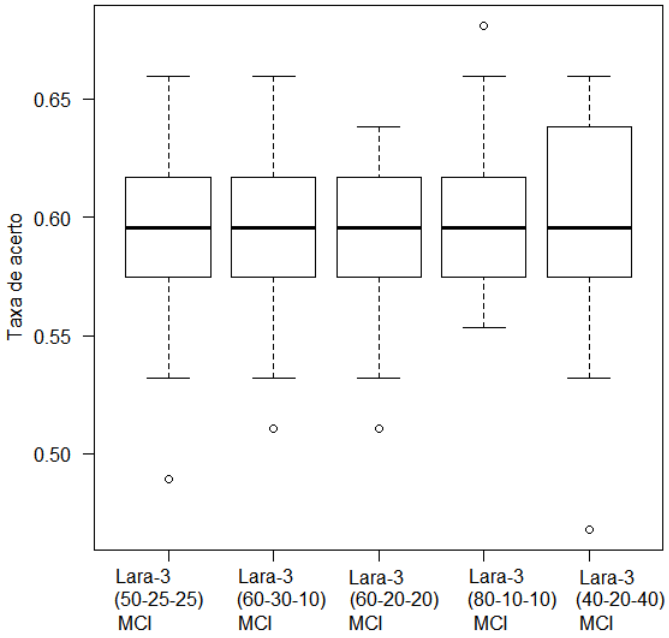


Figura 34 – Gráfico de Box Plot das amostras de testes DA/MCI da execução da RC na base Lara-3

5 Conclusão e Trabalhos Futuros

A doença de Alzheimer representa atualmente 60% dos casos de demência no mundo. Com o crescimento previsto da população idosa espera-se que o número de casos de idosos no mundo diagnosticados com o Mal aumente significativamente, podendo triplicar até 2050. Neste cenário, deve-se investir cada vez mais em pesquisas para melhorar a condição de vida desses pacientes.

Foi discutido nesse trabalho que o diagnóstico precoce da doença pode auxiliar não só o campo de pesquisas como também a busca pela melhor condição de vida dos pacientes e familiares. Possibilita maiores ações de prevenção dos sintomas da doença como a morte das células cerebrais, visto que hoje em dia muitas vezes a doença é diagnosticada quando o paciente apresenta um quadro muito avançado de morte das células nervosas impossibilitando ações preventivas. E também proporciona ao paciente um maior acesso à informação que possa garantir sua melhor qualidade de vida durante os estágios da doença.

As redes neurais artificiais se tornaram aliadas das pesquisas em busca de diagnósticos precoce para esse mal. A literatura fez avanços significativos quanto à exploração da base de dados disponibilizada contendo 220 amostras do plasma sanguíneo dos pacientes. Ray et Al, Gómez e Moscato e Lara foram capazes de identificar subconjuntos de proteínas dentre as 120 proteínas listadas na base de dados capazes de diagnosticar a DA com grande desempenho. Contudo avaliou-se a necessidade de um estudo sobre o universo dos dados trabalhados, uma vez que todos os trabalhos da literatura mantiveram a mesma divisão e aplicação dos dados.

Este trabalho foi responsável por criar um estudo de sensibilidade na base de dados utilizada na literatura para diagnóstico da doença. Procurou-se entender e medir através de testes estatísticos o efeito da variação nos conjuntos de dados de entrada sobre o desempenho das redes neurais *Multi-Layer Perceptron* e *Reservoir Computing*. Como resultado obteve-se um efeito negativo, na maioria dos casos, na variação do percentual de acertos das redes neurais ao decorrer das variações nos conjuntos utilizados na execução das redes.

Foi identificado que apenas para os testes da base Lara-3 do conjunto de testes de pacientes com MCI previamente diagnosticado obteve-se resultados estatisticamente positivos em relação à variação dos dados. Neste caso, em ambas as redes neurais, as medianas possuem valores aproximados, como podem ser vistos nos gráficos Box Plot, e todos os testes estatísticos de comparação obtiveram $p\text{-}value > 0.05$ indicando que, com um nível de significância de 5% podemos afirmar que os dados são estatisticamente iguais.

É esperado uma diminuição do desempenho nos testes de DA/MCI pois sabe-se

que esse teste foi aplicado em uma rede onde o treinamento foi efetuado sem exemplos característicos do conjunto de testes, ou seja, nos conjuntos de treinamento e validação não foram incluídos nenhum exemplo em que o paciente possui MCI diagnosticado previamente. Contudo os resultados exibidos possuem desempenho médio ainda inferior aos encontrados na literatura. Goméz e Moscato atingiram taxas de 65% a 64%, a média do desempenho do MLP para os conjuntos criados a partis das porcentagens Lara-3(50-25-25)-DA/MCI ,Lara-3(60-30-10)-DA/MCI , Lara-3(60-20-20)-DA/MCI, Lara-3(80-10-10)-DA/MCI e Lara-3(40-20-40)-DA/MCI, foi calculada em respectivamente em 57,7%,57,4%, 58,8%, 57,8% e 57,3%, ao passo que as médias de desempenho do RC foram de 59,4%, 59,4 %, 59,0%, 60,4% e 60,0%.

De acordo com os demais gráficos e resultados estatísticos apresentados nesse trabalho, podemos inferir que a bases de dados geradas a partir das assinaturas de 3 e 5 proteínas de Lara, salvo o caso citado anteriormente, são sensíveis à variação de tamanho dos seus conjuntos de entrada. As análises feitas resultaram em amostras de desempenho estatisticamente diferentes, com distribuição, média, mediana e amplitude desigual, mostrando que para uma mesma assinatura da base de dados a variação dos conjuntos de entradas das redes afeta diretamente no seu desempenho podendo causar alterações de mais de 10% na taxa de acerto da classificação da rede.

Podemos concluir que a base de dados apresentada não obtém um grau de confiabilidade na classificação da doença, uma vez que se mostra sensível as alterações nas parametrizações definidas para as redes neurais responsáveis pela execução da classificação.

Podemos listar como trabalhos futuros, à execução de mais testes estatísticos em amostras relacionadas à execução das redes para outras assinaturas encontradas, com o propósito de atestar os testes realizados nesse trabalho. E também a coleta de mais exemplos de plasma sanguíneo para melhorar o conhecimento associado à base de dados, como por exemplo a identificação de mais casos em que os pacientes foram diagnosticados anteriormente com MCI, possibilitando inclusão desses exemplos nas fases de treinamento e validação da execução da rede neural.

Referências

[1] PRINCE, M. et al. World alzheimer report 2016: improving healthcare for people living with dementia: coverage, quality and costs now and in the future. Alzheimer's Disease International (ADI), 2016. Citado na página 15.

[2] ASSOCIAÇ aO Brasileira de Alzheimer. ABRAz. 2017. `http://www.abraz.org.br/`. Acessado em Outubro de 2017. Citado 3 vezes nas páginas 15, 18 e 19.

[3] GAUTHIER, S. et al. Mild cognitive impairment. *The Lancet*, Elsevier, v. 367, n. 9518, p. 1262–1270, 2006. Citado na página 15.

[4] PRINCE, M.; BRYCE, R.; FERRI, C. *World Alzheimer Report 2011: The benefits of early diagnosis and intervention*. [S.l.]: Alzheimer's Disease International, 2011. Citado na página 15.

[5] RAY, S. et al. Classification and prediction of clinical alzheimer's diagnosis based on plasma signaling proteins. *Nature medicine*, v. 13, n. 11, 2007. Citado 3 vezes nas páginas 16, 30 e 34.

[6] RAVETTI, M. G.; MOSCATO, P. Identification of a 5-protein biomarker molecular signature for predicting alzheimer's disease. *PloS one*, Public Library of Science, v. 3, n. 9, p. e3111, 2008. Citado 3 vezes nas páginas 16, 30 e 34.

[7] ALZHEIMER'S Association. ALZ. 2017. `https://www.alz.org/`. Acessado em Outubro de 2017. Citado 2 vezes nas páginas 19 e 20.

[8] ALZHEIMER'S Disease International. 2017. `https://www.alz.co.uk/`. Acessado em Outubro de 2017. Citado na página 20.

[9] VALENÇA, M. J. Fundamentos das redes neurais: exemplos em java. *Recife: Livro Rápido*, JSTOR, v. 52, p. 591–611, 2007. Citado 2 vezes nas páginas 21 e 22.

[10] VERSTRAETEN, D. et al. An experimental unification of reservoir computing methods. *Neural networks*, Elsevier, v. 20, n. 3, p. 391–403, 2007. Citado na página 25.

[11] JAEGER, H. The "echo state"approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, v. 148, n. 34, p. 13, 2001. Citado na página 25.

[12] MAASS, W.; NATSCHLÄGER, T.; MARKRAM, H. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, MIT Press, v. 14, n. 11, p. 2531–2560, 2002. Citado na página 25.

[13] VERSTRAETEN, D. *Reservoir computing: computation with dynamical systems*. Tese (Doutorado) — Ghent University, 2009. Citado na página 25.

[14] HAYKIN, S. Neural networks: principles and practice. *Bookman*, 2001. Citado 3 vezes nas páginas 23, 26 e 28.

[15] COUTINHO, L. D. *Utilizando Redes Neurais Artificiais e Algoritmos de Seleção de Variáveis para Realizar Diagnóstico Precoce da Doença de Alzheimer e do Déficit Cognitivo Leve.* 2015. Dissertação (Mestrado), UPE (Universidade de Pernambuco), Recife, Brasil. Citado 2 vezes nas páginas 30 e 34.

[16] TORMAN, V. B. L.; COSTER, R.; RIBOLDI, J. Normalidade de variáveis: métodos de verificação e comparação de alguns testes não-paramétricos por simulação. *Clinical & Biomedical Research*, v. 32, n. 2, 2012. Citado na página 36.

[17] SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, JSTOR, v. 52, n. 3/4, p. 591–611, 1965. Citado na página 36.

[18] SOARES, E. *Multi-layer Perceptron e Reservoir Computing aplicadas em um processo chuva x vazão.* 2016. Dissertação de Mestrado (Bacharel em Informática), UPE (Universidade de Pernambuco), Recife, Brasil. Citado na página 37.

[19] INTERPRETAR os principais resultados para Boxplot. 2017. `https://support.minitab.com/pt-br/minitab/18/help-and-how-to/graphs/how-to/boxplot/interpret-the-results/key-results/`. Acessado em Novembro de 2017. Citado na página 38.