Universidade de Brasília – UnB
Campus Gama – FGA
Programa de Pós-Graduação

Detecção de alterações cerebrais anatômicas associadas à esquizofrenia com base em redes convolucionais aplicadas a imagens de ressonância magnética

Rodrigo Fay Vergara

Orientador: Dr. Cristiano Jacques Miosso

Rodrigo Fay Vergara

Detecção de alterações cerebrais anatômicas associadas à esquizofrenia com base em redes convolucionais aplicadas a imagens de ressonância magnética

Dissertação de Mestrado submetida ao programa de Pós-Graduação da Faculdade Gama da Universidade de Brasília, como parte dos requisitos necessários para a obtenção do grau de mestre em engenharia biomédica

Orientador: Dr. Cristiano Jacques Miosso

Brasília, DF
2018

Brasília/DF, Agosto de 2018

Ficha Catalográfica

Rodrigo Fay Vergara

Detecção de alterações cerebrais anatômicas associadas à esquizofrenia com base em redes convolucionais aplicadas a imagens de ressonância magnética

89p., 210 × 297 mm (FGA/UnB Gama, Mestrado em Engenharia Biomédica, 2018)

Dissertação de Mestrado em Engenharia Biomédica
Universidade de Brasília, Campus Gama
Programa de Pós-Graduação em Engenharia Biomédica

1. Detecção de Esquizofrenia
2. Redes Neurais Convolucionais
3. Imagens de Ressonância Magnética
4. Engenharia Biomédica

I. FGA UnB/UnB.
II. 096A/2018

Referência

Vergara, Rodrigo Fay (2018). Detecção de alterações cerebrais anatômicas associadas à esquizofrenia com base em redes convolucionais aplicadas a imagens de ressonância magnética Dissertação de mestrado em engenharia biomédica, Publicação 096A/2018 Programa de Pós-Graduação, Faculdade UnB Gama, Universidade de Brasília, Brasília, DF, 89p.

This paper is dedicated to those who suffer from a mental illness and the families that support them through difficult times.

Agradecimentos

Gostaria de agradecer primeiramente aos meus pais, Dirson Vergara e Rosane Fay, que me apoiaram e sempre estiveram ao meu lado em momentos difíceis. À minha namorada Laura Marques, por me ajudar a não desistir e estar comigo em todos momentos. Gostaria de agradecer também ao meu orientador Cristiano Miosso pela dedicação e contribuição na elaboração deste trabalho. Aos amigos e familiares pelo apoio e compreensão em momentos de ausência. A todos aqueles que direta ou indiretamente contribuíram para este trabalho.

UNIVERSIDADE DE BRASÍLIA
FACULDADE DO GAMA
ENGENHARIA BIOMÉDICA

"DETECÇÃO DE ALTERAÇÕES CEREBRAIS ANATÔMICAS ASSOCIADAS À ESQUIZOFRENIA COM BASE EM REDES CONVOLUCIONAIS APLICADAS A IMAGENS DE RESSONÂNCIA MAGNÉTICA"

RODRIGO FAY VERGARA

DISSERTAÇÃO DE MESTRADO SUBMETIDA À FACULDADE UNB GAMA DA UNIVERSIDADE DE BRASÍLIA, COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO TÍTULO DE MESTRE EM ENGENHARIA BIOMÉDICA.

APROVADA POR:

Cristiano Jacques Miosso

PROF. DR. CRISTIANO JACQUES MIOSSO RODRIGUES MENDES; FGA / UNB
ORIENTADOR)

PROF. DR. FABIANO ARAÚJO SOARES; FGA / UNB
(EXAMINADOR EXTERNO)

Renan Utida Ferreira

PROF. DR. RENAN UTIDA FERREIRA; FGA / UNB
(EXAMINADOR EXTERNO)

BRASÍLIA, 13 DE AGOSTO DE 2018

"We can only see a short distance ahead, but we can see plenty there that needs to be done"
Alan Turing

Resumo

A esquizofrenia é uma transtorno psíquico grave que afeta cerca de 1% da população mundial, e seu diagnóstico é realizado por um médico especializado baseando-se no Manual do Diagnóstico e Estatístico de Transtornos Mentais DSM-5. Contudo, este tipo de diagnóstico geralmente acontece de forma tardia e diminuindo as chances de tratamento.

Diante da complexidade do diagnóstico clássico da esquizofrenia apresentado pelo manual e da descoberta de mudanças anatômicas em áreas do cérebro existentes em pacientes com a doença, estudos recentes que utilizaram as características anatômicas para classificação obtiveram resultados promissores. Apesar de mostrarem-se promissores, apenas algumas regiões do cérebro foram utilizadas para classificação, porém, a esquizofrenia apresenta alterações anatômicas em diversas áreas, não havendo um padrão de escolha definitivo para o problema.

Por outro lado, houve avanços em técnicas de aprendizado de máquina como o Aprendizado Profundo (do inglês, *deep learning*). Nestas técnicas não há a necessidade da escolha de características para a classificação do estudo, em outras palavras, sendo uma técnica em que as estruturas aprendem as melhores características que descrevem o problema de forma automática, diferentemente de técnicas clássicas de classificação como a SVM (do inglês, *Support Vector Machine*), em que existe a necessidade da escolha destas características como forma de entrada.

Neste contexto, a pesquisa propõe a aplicação de uma técnica de *deep learning* chamada Rede Neural Convolucional (CNN, do inglês *Convolutional Neural Network*) para classificação automática de imagens de ressonância magnética estrutural do cérebro e diagnóstico da esquizofrenia, além de realizar a extração das características aprendidas no treinamento para utilização em outros classificadores clássicos para comparação.

O método proposto consiste no desenvolvimento de uma estrutura convolucional baseada em CNN, produzindo métricas de desempenho como precisão, acurácia e sensibilidade relativos ao diagnóstico. Foi utilizado um banco de dados de MRI do encéfalo humano ponderadas em T_2 de 87 indivíduos diagnosticados previamente com esquizofrenia e 85 indivíduos saudáveis de controle. O estudo ainda apresenta uma comparação de desempenho relativos ao tamanho da rede convolucional e o tamanho dos filtros utilizados, de modo a apresentar a rede que melhor se adéque ao problema.

É realizada ainda uma validação cruzada dos dados, utilizando um método de *holdout* com reamostragem aleatória com 530 iterações para cada predição e um método de *k-fold* com k=20, afim de medir e comparar os algoritmos de aprendizado para produzir um resultado mais confiável e reprodutível, estimando o desempenho e normalizando a generalização do

sistema. Em cada validação 70% das imagens foram utilizadas para treinamento e 30% para classificação e validação do sistema. Além disso, uma camada de *dropout* foi introduzida para prevenir a ocorrência de *overfitting*. Resultados utilizando *k-fold* apresentam uma acurácia média de 84% para uma rede convolucional de tamanho 3, com camadas de *dropout* antes e depois da camada de conexão.

Portanto, o uso de técnicas de *deep learning* para auxílio ao diagnóstico de esquizofrenia mostra-se promissor, onde houve um avanço nos resultados previamente obtidos utilizando o mesmo banco de dados. Desta forma evidenciando que com o avanço de técnicas de classificação de imagens, mais próximo será a utilização destes modelos de forma segura para o auxílio ao diagnóstico de doenças. Ainda, a região que maior apresentou interferência e peso para classificação mostrou compatibilidade com a literatura existente.

Palavras-chave: Esquizofrenia; Classificação; Redes Neurais; MRI;

Abstract

Schizophrenia is a severe psychiatric disorder that affects about 1% of the world's population and is diagnosed by a physician based on the Diagnostic and Statistical Manual of Mental Disorders DSM-5. However, this type of diagnosis usually happens belatedly, lowering treatment success.

Given the complexity of the classic diagnosis of schizophrenia presented by the manual and the discovery of anatomical changes in areas of the brain existing in patients with the disease, recent studies that used anatomical characteristics for classification have obtained promising results. Although studies are promising, only a few regions of the brain have been used for classification, but schizophrenia has anatomical changes in several areas, and there is no definitive pattern of choice for the problem.

On the other hand, there have been advances in machine learning techniques such as deep learning. In these techniques, there is no need to choose characteristics for the classification of the study; in other words, it is a technique which structures learn the best characteristics that describes the problem automatically, unlike the classical techniques of classification such as SVM (Support VectorMachine) which there is a need to choose these features as input.

In this context, the research proposes the development of a deep learning technique called the Convolutional Neural Network (CNN) for automatic classification of brain magnetic resonance imaging and diagnosis of schizophrenia, also extracting the learne characteristics in training for use in other classical classifiers for comparison.

The proposed method consists in developing a trellis structure based on CNN, producing performance metrics such as precision, accuracy and sensitivity for the diagnosis. An MRI database of the human brain T2-weighted of 87 individuals previously diagnosed with schizophrenia and 85 healthy control subjects was used. The study also shows a performance comparison for the size of convolutional network and filter size used to display the network that best describes the problem.

A cross-validation of the data is performed, using a holdout method with random subsampling of 530 iterations for each prediction and a k-fold using k=20, in order to measure and compare the learning algorithms to produce a more reliable and reproducible result, estimating the performance and normalizing the generalization of the system. In each validation, 70% of the images were used for training and 30% for system classification and validation. In addition, a dropout layer was introduced to prevent the occurrence of overfitting. Preliminary results have an average accuracy of 84% for a convolutional network of size 3, with dropout layers before and after the connection layer.

Therefore, the use of deep learning techniques to aid in the diagnosis of schizophrenia is promising, where there was an improvement in the results previously obtained using the same database, indicating that with the advancement of image classification techniques, the closer will be the use of these models in a safe way for diagnosis of diseases. Also, the region that presented the greatest interference and weight for classification was compatible with the existing literature.

Keywords: Schizophrenia; Classification; Neural Network; MRI

Sumário

Lista de Tabelas

Lista de Figuras

Lista de Símbolos, Nomenclaturas e Abreviações

CNN – Rede Neural Convolucional (do inglês, *Convolutional Neural Network*)
DL – Aprendizagem Profunda (do inglês, *Deep Learning*)
ML – Aprendizagem de Máquina (do inglês, *Machine Learning*)
MR – Ressonância Magnética (do inglês, *Magnetic Resonance*)
MRI – Imagem por Ressonância Magnética (do inglês, *Magnetic Resonance Imaging*)
SVM – Máquinas de Vetores de Suporte (do inglês, *Support Vector Machine*)
FP – Falso Positivo (do inglês, *False Positive*)
FN – Falso Negativo (do inglês, *False Negative*)
TP – Verdadeiro Positivo (do inglês, *True Positive*)
TN – Verdadeiro Negativo (do inglês, *True Negative*)
CID – Código Internacional de Doenças
OMS – Organização Mundial da Saúde

1 Introdução

Este trabalho aborda um problema de detecção e classificação de esquizofrenia utilizando técnicas de aprendizado profundo em imagens de ressonância magnética anatômica do cérebro humano. Dentro deste contexto, o presente capítulo apresenta um estudo da área de esquizofrenia, mostrando sua importância e lacunas que podem ser exploradas. É levantada uma problematização, evidenciando as dificuldades de diagnóstico, limitações das técnicas, inconsistências e imprecisões nas soluções existentes até o momento. São apresentadas também possíveis soluções para as lacunas observadas, levantando perguntas de pesquisa e salientando as contribuições geradas, com possíveis impactos no âmbito da saúde e pesquisa.

1.1 Contextualização

A introdução do imageamento médico em meados do século XIX abriu as portas para descobertas e diagnósticos de diversas doenças, e que hoje conduzem milhares de pesquisas mundo a fora. Com o uso de diferentes técnicas como tomografia computadorizada, radiografia por raio X, e ressonância magnética é possível criar representações visuais de diversas partes do interior do corpo humano e assim auxiliar no diagnóstico de inúmeras patologias como aneurismas cerebrais [62, 65], tumores [69, 24], osteoporose [20, 49], entre outras.

Existem ainda evidências que comprovam que doenças neurodegenerativas como Alzheimer [9], doença de Huntington [31], esclerose lateral amiotrófica [60], além de outras psicopatias como esquizofrenia [4], levam a alterações morfológicas do cérebro. Contudo, cada uma destas doenças apresentam alterações cerebrais em diferentes áreas e apresentam um quadro clínico muito específico em relação aos sintomas para possível diagnóstico. Como Greicius e Mulders descrevem [26, 44], diversas doenças psiquiátricas afetam múltiplas áreas do cérebro, apresentando mudanças estruturais em diversos níveis de ação, ao invés de apenas alterações isoladas.

Uma das doenças que apresentam mudanças estruturais no cérebro e que vem sendo estudada nos últimos anos é a esquizofrenia [4, 50, 57]. Transtornos psicóticos em um espectro completo são representados por um conjunto de psicopatias que contemplam além da esquizofrenia, outros tipos de transtornos, como transtorno de personalidade, transtorno esquizofreniforme, transtorno delirante, e afetam em sua totalidade cerca de 1% da população mundial [5, 43]. Só no Brasil, a psicopatia afeta cerca de 150 mil pessoas todo ano e segundo a Organização Mundial da Saúde (OMS). Trata-se de uma das mais severas desordens men-

tais existentes que pesquisadores ainda não identificaram apenas um simples fator causador da doença, e sim que diversos fatores genéticos, ambientais e fisiológicos contribuem para ela [5].

O diagnóstico do espectro da esquizofrenia é realizado por um médico, ou um grupo de médicos especialistas, baseado "Manual do Diagnóstico e Estatístico de Transtornos Mentais DSM-5" [5], que descreve as características essenciais para a definição e diagnóstico dos transtornos psicóticos. Essas características são separadas em dois grupos distintos, um primeiro descreve a presença de características e sintomas ativos no indivíduo, chamado de sintomas positivos, e outro onde há a ausência de características básicas inerente ao indivíduo, chamado de sintomas negativos [5]. Os sintomas positivos são referentes a delírios, alucinações, pensamento desorganizado, comportamento motor grosseiramente desorganizado ou anormal. Já os sintomas negativos, que são menos presentes em outros tipos de psicopatias, e mais difícil de ser diagnosticado, por serem menos objetivos, referem-se à diminuição das expressões emocionais, que englobam redução no contato visual, diminuição das expressões da face e da fala. Outras características que definem os sintomas negativos são a alogia, anedonia e falta de sociabilidade [21]. Este método de diagnóstico, apesar de ser bem aceito e bem definido, acontece geralmente após o paciente apresentar uma ou mais das características citadas no manual, podendo demorar de meses até alguns anos para o diagnóstico completo do paciente. Com o diagnóstico tardio, o início do tratamento acontece de forma demorada, consequentemente diminuindo as chances de sucesso do tratamento [43, 21].

As características e primeiros episódios de esquizofrenia diferem entre a população masculina e feminina. No sexo masculino o primeiro episódio psicótico ocorre em meados dos 20 anos, enquanto na população feminina ocorre no final dos 20 anos. Caraterísticas de esquizofrenia podem ser encontradas também em crianças, porém é mais difícil diagnosticar pelo fato de, que em crianças, alucinações visuais são mais comuns e as descrições das alucinações podem ser facilmente confundidas com jogos fantasiosos [5, 21].

As características essenciais para a definição da esquizofrenia, porém, podem ser mal interpretadas pelo fato de diferentes transtornos apresentarem características semelhantes. Transtornos como o psicótico breve, transtorno delirante, além do transtorno depressivo e bipolar, apresentam sintomas similares ao da esquizofrenia, portanto, em alguns casos, apenas o diagnóstico clínico não é capaz de afirmar a natureza do transtorno. Deste modo, exames adicionais seriam bem-vindos para um diagnóstico mais preciso e confiável, contudo, atualmente não existem exames laboratoriais, radiológicos ou testes psicométricos para o transtorno que sejam difundidos e de fácil acesso.

Desde o início do século passado, quando a doença foi identificada, até pouco tempo atrás, o diagnóstico da esquizofrenia era realizada puramente em sintomas clínicos evidenci-

ados pelo paciente, porém nos últimos anos, outros testes para auxílio ao diagnóstico foram desenvolvidos. Testes como o relatado por Lenton [39] em que indivíduos foram submetidos à testes para avaliar o movimento dos olhos, no qual indivíduos que apresentavam esquizofrenia possuíam uma velocidade menor de movimentação em relação aos indivíduos saudáveis e obteve uma taxa de acerto de 95%. Outros testes, como o relatado por Mota, utilizam a desorganização e aleatoriedade do discurso do indivíduo na primeira consulta clínica para avaliar a possibilidade da esquizofrenia, alcançando um sucesso de 91% de acerto [42]. Estes estudos evidenciam a existência de técnicas complementares ao diagnóstico clínico e que podem futuramente fazer parte das ferramentas utilizadas por médicos, porém são exames específicos, e não complementares a técnicas já existentes como em imagens de MR (Ressonância magnética, do inglês Magnetic Ressonance).

Neste contexto, pesquisadores encontraram fortes evidências que apontam alterações anatômicas em diversas áreas do cérebro em pacientes que apresentam um quadro de esquizofrenia [4, 61]. E ainda com o avanço de técnicas de aprendizado de máquina, que representa uma área da inteligência artificial responsável por descobrir padrões presentes em dados afim de prever novos dados de entrada, o auxílio do diagnóstico por meio destas características estruturais do cérebro mostra-se cada vez mais presente [64]. Um maior detalhamento acerca das regiões afetadas e dos estudos desenvolvidos será apresentado no capítulo 2.

Como base nesses fatos, algumas pesquisas começaram a abordar uma possível detecção da esquizofrenia por meio de imagens de ressonância magnética do cérebro, em conjunto com classificação automática utilizando teorias clássicas de classificação [16, 53] e teorias mais recentes de aprendizado de máquina [64, 3, 68].

Por outro lado, houve grandes avanços em métodos de classificação para problemas em geral, especialmente em técnicas de aprendizado de máquina modernos como o de aprendizagem profunda (do inglês, *deep learning*). O uso do *deep learning* para classificação se mostra conveniente pelo fato de que estas estruturas aprendem as melhores características que descrevem o sinal de forma automática [35], diferentemente de técnicas clássica de aprendizado de máquina que necessitam que as características dos dados sejam inseridas como entrada do algoritmo. Uma das técnicas dentro do *deep learning* que possui resultados surpreendentes é a Rede Neural Convolucional (do inglês *Convolutional Neural Network* - CNN). Apesar da CNN ter sido aprimorada em 1998 por LeCun [37], apenas recentemente ganhou atenção e visibilidade pelo fato de ter quebrado vários recordes de métricas na competição de reconhecimento visual *ImageNet Large-Scale Visual Recognition Challenge* com uma estrutura desenvolvido chamada AlexNet desenvolvida por [33]. Embora tenha obtido grande desempenho em problemas relacionados com segmentação de imagem [52, 14, 40], reconhe-

cimento de voz [29, 1, 2] e reconhecimento de objetos [13], o uso em aplicações médicas ainda está aquém do seu potencial. Estudos recentes utilizam CNN para aplicações médicas como detecção de tumores [28]. Essa técnica pode ser utilizada também em classificação de esquizofrenia como referenciado em [63], entretanto não foi encontrada na literatura muitas outras contribuições utilizando CNNs neste escopo.

1.2 Definição do Problema Científico e Proposta

Vários métodos e técnicas de aprendizado de máquina vêm sendo utilizados na última década em estudos relacionados à detecção de doenças psiquiátricas e apresentam um grande sucesso. Diversas pesquisas utilizam classificadores supervisionados clássicos como a SVM *(Support Vector Machine)*, que utiliza um hiperplano para separar duas classes de interesse [59, 12], porém um dos principais obstáculos na classificação clássica é a necessidade de selecionar manualmente características que representam verdadeiramente o escopo do problema em questão para que sejam classificadas de foram correta [35]. Uma escolha errada das características pode comprometer totalmente o estudo, e assim, não sendo de interesse científico.

Entretanto, outros métodos de aprendizado de máquina vêm ganhando força e apresentando resultados surpreendentes em relação à classificação de psicopatias, mais especificamente, em esquizofrenia. Método como o de aprendizado profundo (*deep learning*), diferentemente dos algoritmos de aprendizado clássico, não requer a escolha manual de características [7], diminuindo a interferência sobre os resultados, e identificando automaticamente as características que melhor representam os dados de entrada. Esta identificação é feita por meio da aplicação de diversas transformações não-lineares e que serão discutidos melhor no capítulo 2.

Das diversas técnicas de *deep learning*, uma que ainda pouco foi pouco explorada no diagnóstico da esquizofrenia e que possuem resultados surpreendentes em aplicação de outras áreas de reconhecimento visual [33] é a CNN. Esta técnica foi inicialmente desenvolvida para aplicações de processamento de vídeo e imagens, porém existem estudos recentes que apontam o seu uso em classificação de psicopatologias [51]. Entretanto, por se tratar de um estudo recente, poucas pesquisas a utilizam, deste modo proporcionando ainda muito espaço para ser desenvolvida.

Portanto, diante das lacunas apresentadas, o trabalho propõe a detecção de alterações anatômicas associadas à esquizofrenia com base em redes neurais convolucionais aplicadas a imagens de ressonância magnética com o objetivo de proporcionar auxílio ao diagnóstico e triagem, além de realizar uma transferência de conhecimento utilizando características

intermediárias da CNN com o intuito de utilizá-las em outras arquiteturas de aprendizado de máquina. Esta transferência de conhecimento é importante pois é possível observar se as características extraídas pela arquitetura são relevantes para outros classificadores, sendo assim, aumentando a confiabilidade dos resultados obtidos.

1.3 Objetivos

1.3.1 Objetivo Geral

Este projeto tem como objetivo o desenvolvimento de um programa para classificação e diagnóstico da esquizofrenia utilizando técnicas de CNN e produzir métricas de desempenho como precisão, acurácia, sensibilidade relativos ao diagnóstico. Faz parte desta análise o levantamento de características e áreas anatômicas enfatizadas pelos filtros treinados pelas redes convolucionais, comparando-as com as áreas previamente descritas na literatura. Com estas novas características, o projeto visa realizar uma transferência de conhecimento, utilizando-as em classificadores clássicos para efeito de comparação. A pesquisa pretende ainda observar as métricas de desempenho utilizando as novas características comparado com estudos anteriores que utilizaram o mesmo banco de dados.

A respeito da rede convolucional, a pesquisa tem como objetivo levantar uma analisar o desempenho considerando o tamanho da rede e tamanho dos filtros para produzir a rede que melhor solucione o problema.

1.3.2 Objetivos Específicos

Afim de contemplar os objetivos gerais do projeto, têm-se como objetivos específicos os seguintes procedimentos:

- Desenvolvimento de um programa para realização da classificação utilizando CNN.
- Avaliações sistemáticas utilizando *deep learning* e compará-las com teorias clássicas de classificação como *AdaBoost*, *Bagging*, Gradient Boost, e classificadores do tipo SVM.
- Utilização de características extraídas para outros classificadores.
- Extração de imagens obtidas por meio de máscaras de filtragem.

1.4 Justificativa e Contribuições

Com a aplicação de novas técnicas para extração de características e utilizando-as para classificação e diagnóstico de esquizofrenia, é possível realizar diagnósticos de forma mais precoce, deste modo aumentando as chances de tratamento da doença. Uma das grandes dificuldades de uma classificação utilizando técnicas clássicas de aprendizado de máquina, é a escolha precisa das características a serem utilizadas. Com a aplicação do algoritmo de CNN, características essenciais para a classificação são escolhidas de forma automática, características essas que podem ser extraídas dos filtros não-lineares e utilizadas como forma de entrada em outros algoritmos de classificação como a SVM, e até a utilização de outros conjuntos de algoritmos chamados *ensembles*.

Portanto, uma vez que os objetivos da pesquisa forem alcançados, resultados de classificação mais robustos, com maior precisão e mais confiável poderão ser obtidos, aumentando assim o sucesso do diagnóstico.

Da mesma forma, novas áreas do cérebro que antes não eram consideradas importantes para o estudo de esquizofrenia, podem se revelar essenciais por meio da extração automática de características do cérebro. Com isso, abrem-se portas para que outras pesquisam desenvolvam técnicas utilizando as novas características obtida pelas CNNs, desta forma melhorando o desempenho de seus classificadores e promovendo transferência de conhecimento.

Um dos principais obstáculos na classificação clássica é a dificuldade de selecionar características que representam verdadeiramente o escopo do problema em questão para que sejam classificadas de foram correta. Uma escolha errada das características pode comprometer totalmente o estudo, e assim, não sendo de interesse científico.

1.5 Estrutura da Dissertação

O restante do texto está organizado da seguinte maneira. O capítulo 2 apresenta a teoria básica acerca da esquizofrenia, levantando os sintomas, métodos de diagnóstico, tipo de tratamento e mudanças anatômicas geradas no encéfalo. O capítulo discute ainda o estado-da-arte em ressonância magnética, incluindo o método de funcionamento, técnicas de formação de imagem e codificação espacial. Ainda é discutida a teoria acerca da classificação automática e aprendizado de máquina, incluindo os avanços em *deep learning* e CNN, que são a base do método proposto.

Buscando critérios de repetibilidade do processo, o capítulo 3 aborda a metodologia de desenvolvimento aplicada para obter os objetivos apresentados. Aborda ainda os tipos de análise aplicados aos dados obtidos a fim de gerar uma métrica de desempenho. Ainda é

discutido o banco de dados utilizado, apresentando os critérios de exclusão do procedimento.

Deste modo, os resultados obtidos por meio da CNN e dos classificadores clássicos utilizando a metodologia apresentada é relatada no capítulo 4, além de uma discussão dos parâmetros obtidos, mostrando as vantagens e desvantagens do uso dos algoritmos desenvolvidos. O último capítulo deste trabalho apresenta as conclusões da dissertação, evidenciando as contribuições geradas para a saúde, pesquisa, entre outros. É apresentada possíveis lacunas ainda existentes e trabalhos futuros que podem se beneficiar do estudo.

2 Fundamentação Teórica

Este capítulo aborda as teorias e ferramentas necessárias para o desenvolvimento do projeto. São apresentadas as características da esquizofrenia e outros transtornos psicóticos, ilustrando os sintomas positivos e negativos das doenças. Dentro do contexto da esquizofrenia, são evidenciadas as características heterogêneas da psicopatia, fatores de risco, prognósticos e anormalidades em certas regiões anatômicas do cérebro relatadas pela literatura. As informações aqui apresentadas sobre a esquizofrenia se baseiam no "Manual do Diagnóstico e Estatístico de Transtornos Mentais DSM-5" [5] e no CID-10 [17].

O capítulo apresenta ainda a teoria sobre formação de imagens de ressonância magnética (MRI), exemplificando o princípio físico e análise da codificação espacial. Esta área é importante pelo fato de as imagens utilizadas no projeto serem parte de um banco de dados público de imagens de ressonância magnética estrutural de esquizofrenia.

Um outro ponto estudado neste capítulo é a utilização de técnicas de aprendizado de máquina para classificação de sinais. Dentro desta área de estudo e por se tratarem da base do método proposto, estão técnicas de *deep learning* e CNN, com destaque a vantagens e desvantagens em comparação com outras técnicas, como a SVM.

2.1 Espectro dos Transtornos Psicóticos

O espectro geral dos transtornos psicóticos abrange, além da esquizofrenia, outros transtornos psicóticos como o transtorno de personalidade, delirante, psicótico breve, esquizofreniforme, esquizoafetivo e psicótico induzido por medicamento. Estes transtornos psicóticos possuem características e sintomas que são divididos em duas categorias, sintomas positivos e sintomas negativos. O diagnóstico é realizado por meio de uma anamnese, onde os sintomas apresentados são detectados por um médico ou grupo de médicos com base nas características descritas no "Manual do Diagnóstico e Estatístico de Transtornos Mentais DSM-5" [5] e no CID-10 [17]. Neste contexto, os sintomas positivos são aqueles sintomas mais ativos e presentes, onde há evidências mais claras de seu aparecimento, e são divididos nas seguintes características:

Sintomas Positivos:

- Delírios: O delírio é uma certeza irredutível que o indivíduo possui mesmo quando evidências que comprovam o contrário são apresentadas. Estes delírios podem ser

divididos em dois tipos, bizarros e não-bizarros. Os bizarros são aqueles que não fazem parte da crença do povo em que o indivíduo habita, sendo surreal para todos naquela comunidade. Já os não-bizarros são decorrentes de teorias de conspiração, delírio de perseguição, entre outros.

- Alucinações: Alucinações são experiências extrassensoriais que o indivíduo vivencia, podendo ser evidenciadas em vozes, pessoas não presentes, e odores não existentes no local.

- Desorganização do Pensamento: Este sintoma está relacionado à ausência da sequência de um fluxo de pensamento, sendo que o indivíduo pode mudar completamente o tópico de discussão, além de elaborar respostas totalmente fora de sentido, sem relação alguma com a pergunta.

- Comportamento Motor Anormal: No comportamento motor anormal, o indivíduo tem dificuldades para realizar atividades cotidianas normais. Em seu caso extremo, o indivíduo para de responder totalmente a estímulos ambientais, podendo realizar movimentos aleatórios, ou nenhum movimento por um grande tempo.

Existem ainda os sintomas negativos, que são sintomas um pouco mais subjetivos em relação aos sintomas positivos, pois são sintomas em que o paciente deixa de apresentar emoções ou deixar de fazer algo que anteriormente gostava de fazer, porém são igualmente importantes na detecção do transtorno apresentado. Estas características são dividas nas seguintes categorias:

Sintomas Negativos:

- Expressão Emocional Diminuída: Está relacionada à diminuição de expressões do rosto e nos gestos, gerando uma apatia e dificuldade na expressão das emoções.

- Avolia: A avolia está relacionada à falta de desejo por tarefas motivadoras, em que o indivíduo pode ficar por muito tempo parado sem fazer um movimento, caracterizando uma falta de motivação para realizar algo.

- Anedonia: Este sintoma diz respeito à falta de prazer do indivíduo relacionado com atividades que antes produziam tal felicidade.

- Falta de Sociabilidade: Está associada à falta de relacionamentos interpessoais, mostrando desinteresse e evitando ao máximo contato com a sociedade.

Estes sintomas podem estar presentes em diversos transtornos psicóticos, porém o que difere um transtorno de outro é a duração e quantidade de sintomas positivos e negativos. Nesta pesquisa é apresentada como é realizado o diagnóstico, com a quantidade e tipo de sintomas para a esquizofrenia, fatores de risco, prognósticos e ainda anormalidades em estruturas no cérebro relativas às diferentes características do transtorno.

2.1.1 A esquizofrenia e seu diagnóstico

A esquizofrenia é um transtorno psicótico muito grave descoberto na virada do século XX e que atualmente atinge cerca de 1% da população mundial, porém sua origem e prevalência ainda não foram totalmente delimitadas. Em geral a esquizofrenia é caracterizada pela distorção da realidade, do pensamento e da percepção do indivíduo. Um dos grandes problemas da sua detecção precoce é pelo fato do indivíduo não ter noção de que a realidade que está vivendo não é a verdadeira, acreditando que os sintomas apresentados representam a realidade.

As características e primeiros episódios de esquizofrenia diferem entre a população masculina e feminina, uma vez que no sexo masculino o primeiro episódio psicótico ocorre em meados dos 20 anos, enquanto na população feminina ocorre no final dos 20 anos. Características de esquizofrenia podem ser encontradas também em crianças, porém é mais difícil diagnosticar pelo fato de que, em crianças, alucinações visuais são mais comuns e as descrições das alucinações podem ser facilmente confundidas com jogos fantasiosos. Por isso é feita, sempre que possível, uma comparação com irmãos não afetados, para se ter um parâmetro de medida.

O diagnóstico da esquizofrenia é realizado por meio da uma anamnese com o indivíduo e sua família conduzida por um médico especialista que, durante a entrevista, busca analisar os sintomas apresentados com base no "Manual do Diagnóstico e Estatístico de Transtornos Mentais DSM-5" e no CID-10, que foram explicados em 2.1.

Por ser um transtorno psicótico, os sintomas apresentados são os mesmos evidenciados por outras psicopatias, dividindo-se em positivos e negativos, porém a diferença está na quantidade destes sintomas e na sua duração. A esquizofrenia possui pelo menos dois sintomas, sendo ao menos um deles um sintoma positivo e um negativo, em que cada um deles tenha sido presente durante pelo menos 1 mês. Um outro fator observado é que desde o início das primeiras evidências de mudanças comportamentais, o indivíduo diminuiu a sua capacidade de relação interpessoal, levando à uma queda de rendimento no trabalho, estudo, ou outras áreas.

Deste modo, a esquizofrenia se diferencia de outros transtornos psicológicos pelo fato

de ser um transtorno heterogêneo, apresentando diferentes sintomas, podendo mudar de indivíduo para indivíduo, onde nenhum sintoma é próprio da esquizofrenia em si, mas sim uma análise do conjunto de todas características apresentadas.

2.1.2 Fatores de Risco e Prognóstico

Embora as causas da esquizofrenia ainda não serem totalmente descobertas e relatadas, alguns fatores de risco foram relatados como sendo os principais para a doença, e estão relacionados a três fatores: ambientais, fisiológicos e genéticos.

Sobre os fatores ambientais, existem evidências que crianças que crescem em ambientes urbanos possuem um maior fator de risco comparado àquelas que crescem longe deste ambiente. Este fator pode ser associado à densidade populacional, onde mudanças estruturais e mutações decorrentes do aglomerado levam a tais riscos. Evidências apontam ainda que o transtorno é mais frequente em grupos étnicos minoritários [48].

Apesar de não haver evidências que relacione a doença com parentes próximos, a mutação genética é um dos principais fatores de risco em indivíduos com esquizofrenia, decorrente do erro no processo de duplicação do DNA. Já em ralação aos fatores fisiológicos descritos na literatura estão relacionados aos de riscos da gestação com uma idade avançada dos pais, problemas durante a gestação, como estresse, infecção, entre outros [19, 48].

2.1.3 Anormalidades e Características da Esquizofrenia

Existem evidências em inúmeros estudos que comprovam que há diferenças em diversas regiões cerebrais de indivíduos esquizofrênicos comparados com indivíduos saudáveis. Apesar do conhecimento do transtorno desde o início do século XX, foi apenas a partir dos anos 90, com o avanço em técnicas de imageamento médico, que estas evidencias foram descobertas, e de lá pra cá, muitos estudos vêm tentando estipular um padrão para as anormalidades.

Apesar de diversos estudos reportarem anormalidades de estruturas cerebrais em indivíduos com esquizofrenia, um dos maiores e mais recentes até o momento, englobando 15 países ao redor do mundo, é o ENIGMA [61]. O grupo responsável pela pesquisa relatou a descoberta de diversas mudanças em regiões do encéfalo comparados com indivíduos saudáveis, entre estas mudanças estão a diminuição do hipocampo, das amídalas cerebelosas, da região do *thalamus*, do núcleo *accumbens*, e do volume intra cranial, além de um maior paleoestriado e volume do ventrículo lateral. Outros estudos apontam para uma mudança na substância branca e no volume da substância cinzenta em uma variedade de regiões, assim como a redução do volume cerebral total.

Apesar do ENIGMA ser um estudo muito vasto com diversos indivíduos para testes e

comparações, hoje ainda não existe um padrão para definir com precisão tais mudanças. Este fato é decorrente do transtorno não possuir apenas um sintoma, abrindo a possibilidade de diferentes indivíduos apresentarem mudanças em diferentes estruturas, o que é resultado da heterogeneidade da doença.

Um outro ponto é o fato de os indivíduos esquizofrênicos terem dificuldade de manter o olhar fixo em algum objeto, sendo assim possível detectar a doença por um sistema de rastreamento da pupila. Este estudo foi desenvolvido por Lenton e possui resultados surpreendentes, obtendo uma taxa de acerto de 95% [39].

2.2 Imagens por Ressonância Magnética

Atualmente MRI é a técnica de imageamento médico mais versátil e a que mais evoluiu na última década em comparação com outras técnicas, como a tomografia, computadorizada. O MRI se destaca pelo seu grande constaste, sendo possível diferenciar diversos tipos de tecido, órgãos, e identificar mudanças anatômicas com alta precisão. Apesar da técnica de ressonância magnética ter sido proposta em meados de 1940 [8], foi apenas nos anos 70, com inclusão do gradiente de campo magnético proposta por Lauterbur [34], que a técnica obteve mais sucesso e abriu as portas para mais pesquisas relacionadas.

Diferente de outras técnicas de imageamento que utilizam retro-projeção proveniente de raios de radiação capturados em diversas camadas e ângulos para reconstrução da imagem, a MR utiliza a propriedade natural dos átomos que constituem os tecidos. Tipicamente, utilizam-se propriedades magnéticas oriundas do hidrogênio, pelo fato de possuir apenas um próton, ser o menor núcleo existente, e ser constituinte de grande parte do corpo humano [66]. A principal propriedade que a MR utiliza é o *spin*, que se refere ao movimento de rotação das partículas sobre seu próprio eixo e variam de acordo com um campo magnético aplicado à ele. Uma vez que diferentes tecidos possuem diferentes densidades de prótons, a composição de *spins* presentes na aquisição de imagens de MR será diferenciado para cada parte do corpo, evidenciando diferentes estruturas.

Neste contexto, a técnica de MR utiliza um gigantesco ímã capaz de produzir um campo magnético uniforme estático $\mathbf{B_0}$, geralmente com magnitude variando de 1.5 tesla a 8 tesla em máquinas modernas (ou mais em máquinas utilizadas em pesquisa), para alinhar os momentos magnéticos dos *spins* μ com $\mathbf{B_0}$, gerando um momento magnético $\mathbf{M}$ chamado Magnetização. Uma vez que os *spins* estão alinhados, é aplicado sobre eles uma perturbação perpendicular ao campo magnético estático afim de medir o resultado gerado. Esta aplicação se dá na forma de um campo magnético de rádio frequência $\mathbf{B_1}$ circularmente polarizado, com uma frequência $\omega_0 = \gamma B_0$, conhecida como *Larmor Frequency*. Esta perturbação causa

então que $\mathbf{M}$ reflita em um plano transversal a $\mathbf{B_0}$ e ressonante em ω_0, induzindo tensão em uma bobina [11].

Este fenômeno pode ser descrito pela equação de *Bloch*, que descreve a dinâmica da magnetização em apenas uma simples equação [11, 66], e é a base da MR:

$$\frac{d\mathbf{M}(t)}{dt} = \mathbf{M}(t) \times \gamma\mathbf{B}(t) - \frac{M_x(t)}{T_2}\mathbf{i} - \frac{M_y(t)}{T_2}\mathbf{j} - \frac{(M_z(t) - M_0)}{T_1}\mathbf{k}, \tag{2.1}$$

em que $\mathbf{B}$ é o campo magnético efetivo total, γ é descrita como taxa giromagnética, cujo valor para átomos de hidrogênio é 42.6 MHz/tesla, T_1 e T_2 são os tempos de relaxação para os componentes longitudinal e transversal da magnetização.

Portanto, a base da MRI consiste em distinguir os tempos de relaxação T_1 e T_2 para diferentes tipos de tecidos, em que cada diferente tecido possui um tempo diferente de relaxação. Apesar de ser possível distinguir o tipo de tecido pelo período de cada tempo, para localizar a sua posição no espaço é necessário utilizar uma codificação espacial. Este tipo de codificação, assim como o método de reconstrução das imagens, não está no escopo deste trabalho, mas pode ser observado em mais detalhes em [11].

2.3 Aprendizado de máquina e classificação automática

Descobrir padrões em sinais vêm sendo um tópico de interesse por grande parte dos pesquisadores por muito tempo. Segundo Bishop [7], o estudo de reconhecimento de padrões está baseado em um conjunto de algoritmos capazes de encontrar características invariantes em um sinal para serem utilizadas na classificação em diferentes categorias. Foi isso que o criador de um dos primeiros algoritmos de aprendizado de máquina, Arthur Samuel, desenvolveu quando escreveu o primeiro algoritmo para jogar damas, melhorando o despenho do resultado a cada vez que que o programa jogava com ele mesmo, refinando os parâmetros de aprendizado, o que vinha a se tornar o pioneiro no ramo de aprendizado de máquina [10].

Para realizar uma classificação, um dos maiores desafios é construir um bom modelo a partir do conjunto de características de entrada do algoritmo e usá-las para reconhecimento de padrões. Dentro destes modelos de classificação, os mais usados são os modelos de aprendizado supervisionado e modelo de aprendizado não-supervisionado.

Nos modelos supervisionados há a criação de rótulos para os grupos de classificação, onde geralmente em classificação binária são rotuladas em classes positivas e negativas. Deste modo, novas características de entrada para validação e teste serão avaliadas nestas classes. Já nos modelos de treinamento não-supervisionados, durante a fase de treinamento não é fornecido o rótulo de entrada de cada elemento para a classificação, visto que seu objetivo é

encontrar um padrão de distribuição inerente do banco inicial de características.

Um dos grandes desafios da aprendizagem de máquina tradicional é em relação à extração de características essenciais do sinal de entrada para servir de entrada para o classificador. Este tipo de extração é uma parte fundamental do problema, em que um especialista da área necessita de total domínio do problema para encontrar as melhores características para utilizar. Apesar de serem raros, existem relatos de estudos utilizando o sinal em sua forma original em problema de aprendizagem de máquina tradicional como em SVM.

Com isso, por se tratar de uma arquitetura rasa, algoritmos tradicionais de aprendizagem de máquina necessitam de um bom extrator de características para tentar resolver problemas de classificação. Dentro deste aspecto, arquiteturas de aprendizagem profunda tentam abordar um espectro mais amplo, sem necessidade de seletividade de características iniciais do sinal de entrada, em que automaticamente tentam encontrar as representações essenciais para a classificação do problema, seja ele supervisionado ou não.

Como um dos objetivos deste trabalho é a transferência de conhecimento para outros classificadores clássicos, uma breve introdução dos classificadores tipo SVM e do tipo *ensemble* será apresentada para uma maior familiarização com as técnicas utilizadas na metodologia. Estes tipos de aprendizado de máquina são compostos por arquiteturas rasas, ou seja, arquiteturas que possuem apenas uma camada de aprendizado, porém são muito eficientes para determinados casos onde as características relevantes ao problema são bem conhecidas, facilitando sua classificação.

2.3.1 SVM

Máquina de vetor de supoerte (SVM, do inglês, *Support Vector Machine*) é um tipo de classificador supervisionado desenvolvido por Vapnik em 1979, e que é utilizado amplamente em diversos problemas de classificação. Ele é considerado um classificador raso, pois não possui camadas de aprendizado com diversas funções não-lineares que auxiliam na extração das informações invariantes presentes no sinal.

Em sua forma mais simples, a SVM utiliza um hiperplano para separar duas classes de interesse [12] [59], em que este hiperplano pode ser descrito por,

$$w_0.x + b_0 = 0, \tag{2.2}$$

em que w representa um vetor de pesos distribuídos ao longo do sinal, e b um escalar de compensação. Se pegarmos como exemplo um conjunto de sinais $(y_0, x_0), ..., (y_k, x_k)$ em que y_k representa o rótulo binário (-1,1) do elemento x_k, eles serão linearmente separáveis se

existir um w e um b capaz de resolver,

$$w.x_k + b \geq 1, \text{ se } y_k = 1, \tag{2.3}$$

$$w.x_k + b \leq -1, \text{ se } y_k = -1, \tag{2.4}$$

Na Figura 2.1 é possível observar os dois vetores chamados de vetores de suporte, que separam as duas classes e o hiperplano ao meio. O objetivo então é separar as duas classes de interesse com o máximo valor de margem possível entre entes dois vetores, fazendo com que a miniminize o erro de classificação [15].

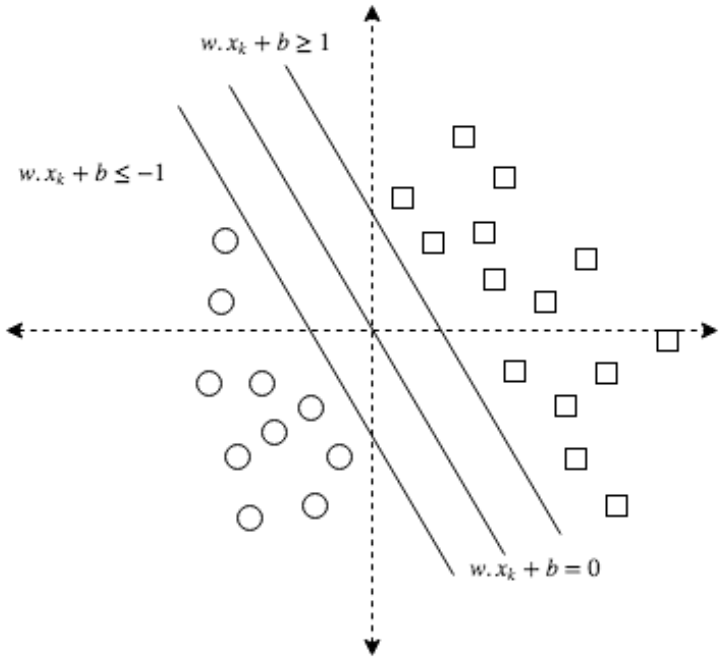

Figura 2.1. Exemplo de duas classes separadas com funções do tipo linear

Podemos perceber pela Equação 2.3 e 2.4 que a distância entre os dois vetores de suporte é

$$\frac{2}{\|w\|}, \tag{2.5}$$

portanto para maximar a distância entre os vetores, temos que minimizar o valor de $\|w\|$.

2.3.2 Classificadores do tipo *ensemble*

Classificadores do tipo *ensemble* são, como o seu nome induz, um agrupamento de classificadores com a finalidade de diminuir o erro, que pode ser de forma sequencial ou paralela.

Os classificadores que possuem agrupamento sequencial são chamados de *boost*, com o exemplo do AdaBoost [70]. Nele, uma sequência de classificadores é utilizado para treinar os dados de entrada, em que os classificadores subsequentes tentam focar nos erros dos prévios [70]. Combinando estes classificadores, onde se explora a dependência de um classificador com o outro, obtém-se um classificador final robusto.

Já no caso dos agrupamento paralelo, o objetivo é explorar a independência entre os classificadores. Um exemplo de agrupamento paralelo é o algoritmo de Bagging, em que é realizado uma média da predição dos classificadores individuais, diminuindo assim a variância entre os modelos [7].

Estes classificadores, assim com a SVM, serão utilizados com o intuito de comparar os resultados da arquitetura de CNN que será apresentada a seguir.

2.4 Aprendizagem profunda

Historicamente, os algoritmos de aprendizagem de máquina e classificação automática de sinais eram limitados na sua forma de interpretar dados de entrada completos, sem extração prévia de características relevantes para o processamento. Por meio dessas técnicas de aprendizagem profunda, busca-se identificar características invariantes do sinal e classificar suas representações abstratas de forma automática, sendo que estas abstrações são mais resistentes a possíveis ruídos no sinal de entrada do sistema, deste modo, tornando o modelo de aprendizado mais robusto e sofisticado.

Estes tipos de métodos de classificação e aprendizagem profunda, que possuem diversos níveis de aprendizado, diferentemente de arquiteturas rasas com apenas um nível, criam automaticamente diversos modelos multinível de modo hierárquico em que a entrada é o sinal original normal sem modificações, e, em cada estágio e camada da arquitetura, se torna mais abstrato, chegando finalmente na camada final de saída em que há o resultado da classificação propriamente dita. Estes níveis de abstração podem ser compreendidos se pegarmos, por exemplo, uma imagem de rosto humano, em que o sinal de entrada seria a imagem completa, seguido de características mais abstratas como o contorno da cabeça, olhos, boca, etc, seguido de uma camada ainda mais abstrata como a intensidade dos *pixels*, e assim por diante.

A Figura 2.2 apresenta um exemplo das camadas de uma arquitetura de aprendizagem profunda, indicando as camadas escondidas em dois níveis assim como as conexões entre elas. Este tipo de conexão e os níveis de elementos de aprendizagem podem se tornar muito complexos de acordo com a quantidade de camadas de aprendizado, elementos de ativação, e profundidade da rede.

Dentro da ampla área de aprendizagem profunda, existem diversos algoritmos que tentam melhorar e otimizar certos casos de estudo, como por exemplo para imagens e vídeos. Um deste tipos de algoritmos otimizados para imagens são as Redes Neurais Convolucionais, algoritmos estes que foram utilizados no desenvolvimento do presente trabalho e são explicados na seção seguinte.

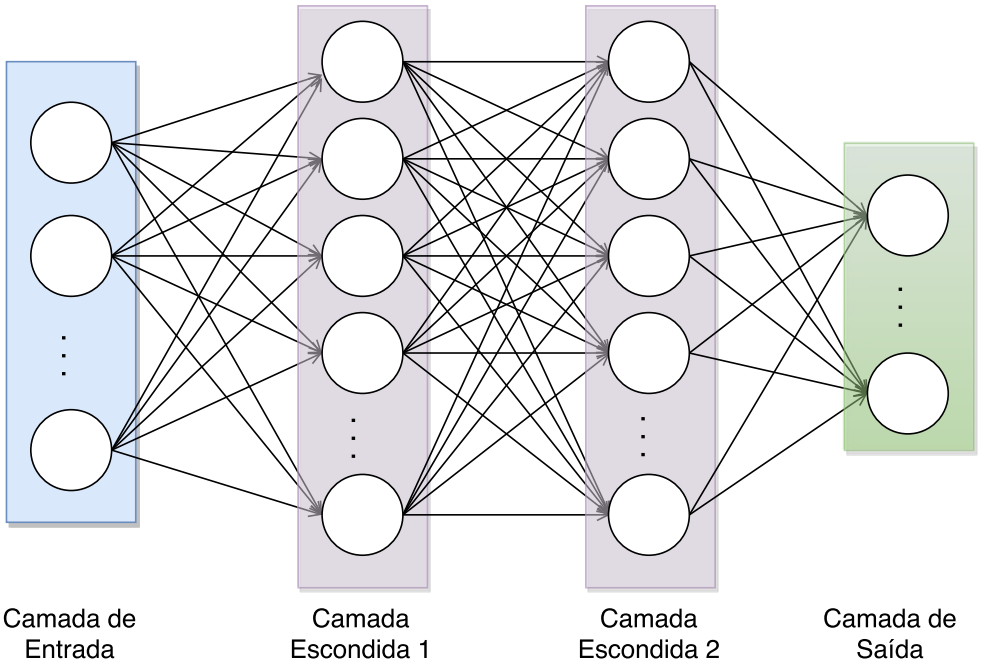

Figura 2.2. Exemplo das camadas de uma arquitetura de Aprendizagem Profunda apresentando as camadas escondidas assim como as conexões entre elas. É possível perceber que o nível de complexidade entre a interconexões aumenta bastante dependendo da quantidade de elementos em cada camada, podendo em alguns casos ser impraticável a sua utilização

2.4.1 Redes Neurais Convolucionais - CNN

Apesar da *Convolutional Neural Network*(CNN) ter sido desenvolvida em 1989 por LeCun [36] e aprimorada por [37], foi apenas recentemente que a técnica ganhou visibilidade com resultados expressivos em competições de classificação de imagem [33]. Trata-se de uma técnica similar à técnica de Rede Neural tradicional, com características únicas que fazem que tenha um resultado melhor em imagens. Um outro fator que fez com que a CNN obtivesse excelentes resultados recentemente é o desenvolvimento de melhores GPU's (unidade de processamento gráfico) e CPU's (unidade de processamento central), fazendo com que o tempo de treinamento, que nos anos 90, era impraticável se tornasse viável.

As Redes Neurais tradicionais não são otimizadas para realizar treinamento e classificação de grandes imagens pelo fato de que os neurônios das camadas de aprendizado estão

diretamente conectados com todos parâmetros de entrada, gerando uma grande quantidade de parâmetros que possivelmente produziria um *overfitting*, isto é, quando o modelo aprende apenas singularidades e características relativas especificamente ao conjunto de treinamento. Deste modo, a CNN assume que o sinal de entrada é uma imagem, fazendo com que os neurônios das camadas de aprendizado se conectem apenas à uma região delimitada da camada anterior, onde cada neurônio possui uma codependência e correlação com os outros neurônios de sua camada, diferentemente da RN tradicional, fazendo com que a quantidade de parâmetros de aprendizado seja reduzida. Com menos conexões, a rede de treinamento fica mais fácil de ser realizada, porém, em casos que a construção da arquitetura não seja realizada de forma adequada, existe a possibilidade destas conexões não representarem um modelo geral de aprendizado.

Assim como em RN tradicionais, a CNN possui várias camadas de aprendizado aumentando o nível de abstração a cada nível de profundidade da arquitetura. Sendo assim as CNNs são composta basicamente por 3 componentes: uma camada convolucional que possui filtros com pesos variados (nas redes tradicionais este filtro é chamado de *Kernel*), seguido de uma propriedade não-linear e uma camada de *pooling* assim como descrito em [38]. A partir destas três propriedades, é possível criar mais camadas de aprendizado conectando-as quantas vezes for necessário.

Um exemplo de como é formada a base da CNN é descrita na Figura 2.3, onde é possível notar que um filtro de tamanho especificado ($m \times m$) convolui uma subamostragem da imagem de entrada por n vezes, sendo n o número total de kernels gerados para a camada. A descrição detalhada de cada camada assim como outras propriedades para otimização da rede serão apresentadas neste trabalho.

Além das camadas de aprendizado, as CNNs modernas utilizam algoritmos de otimização numérica para minimizar os erros de saída e convergir para um valor determinado. Normalmente é utilizada uma forma de *gradient descent* para convergir em determinadas épocas e com uma certa taxa de aprendizado para o valor determinado na construção do treinamento supervisionado.

O passo-a-passo para a construção da rede, explicando o funcionamento de cada camada, será apresentado a seguir.

Convolucional

A primeira camada da CNN é a camada convolucional, ou camada de banco de filtros. Esta camada é responsável por aplicar uma convolução de determinados filtros com uma região da imagem de entrada afim de extrair características que melhor descrevem a região, realizando um produto escalar com os filtros inicias. Para tal, alguns hiper-parâmetros ajustáveis são

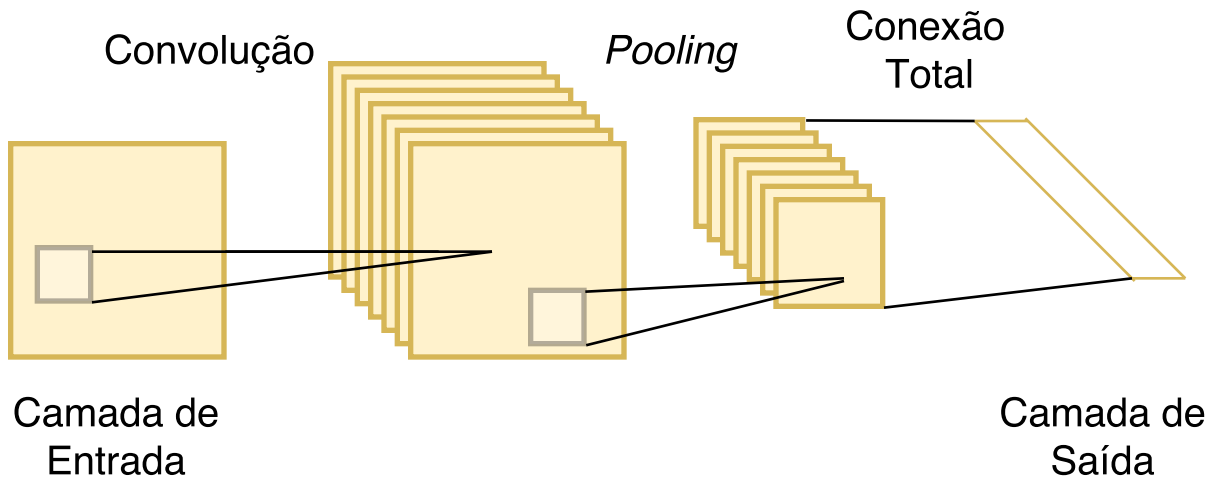

Figura 2.3. Exemplo das camadas de um Rede Neural Convolucional composta por uma camada de convolução por um filtro (mxm), seguido por uma propriedade não-linear e uma cada de subamostragem. Este tipo de estrutura pode ser utilizado diversas vezes para criação de uma arquitetura mais profunda.

selecionados. Dentro destes parâmetros ajustáveis, encontram-se a quantidade, tamanho, e tipo de inicialização dos filtros, assim como o passo (*stride*) e *zero-padding*.

A seleção do tamanho e quantidade de filtros convolucionais na primeira camada é de fundamental importância para que não haja uma quantidade exagerada de parâmetros de aprendizado nas camadas seguintes. Outros parâmetros importantes durante a criação de uma arquitetura é o tamanho do passo que o filtro terá nos *pixels* adjacentes ao inicial. Usualmente utiliza-se um passo de 2 [56], ou seja, a cada convolução há um salto de duas unidades para realização da próxima convolução com o intuito de reduzir pela metade a resolução da rede e evitar a possibilidade de *overfitting*. Caso o tamanho do filtro e o passo não forem proporcionais ao tamanho inicial da imagem, é aplicado um *zero-padding* que adiciona zeros nas bordas da imagem para que os filtros consigam percorrer toda imagem. Porém existem casos em que a imagem de entrada é cortada nas bordas para que o mesmo aconteça.

Para exemplificar a camada convolucional, um exemplo simples é descrito na Figura 2.4. Nele é possível observar a utilização de um passo de 3 em um filtro de tamanho (3x3), gerando uma saída de tamanho (5x5). Para o cálculo do tamanho do filtro T, passo P e *zero-padding* Z, a quantidade de neurônios resultantes da rede tem que ser um valor inteiro de

$$N = \frac{I - T + 2Z}{P} + 1, \tag{2.6}$$

em que I é o tamanho inicial da imagem de entrada. Este cálculo geralmente apenas é

utilizado para se ter uma ideia do número de elementos de saída da rede, dado que, em algoritmos otimizados, mesmo que o tamanho de neurônios não seja inteiro, há um corte da imagem de entrada para não ocorrer sobreposição de filtros.

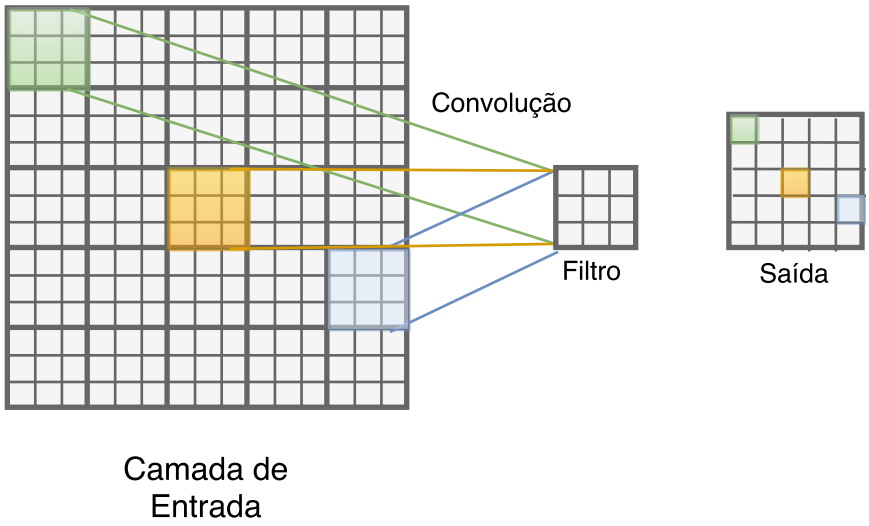

Figura 2.4. Camada convolucional. Nela está presente uma imagem de entrada de tamanho (15x15) e aplicado um filtro de tamanho (3x3) com um passo de 3 e 0 *zero-padding*, gerando assim uma saída de tamanho (5x5).

Apesar da Figura 2.4 apresentar uma imagem $2D$, a ideia em casos em que a imagem tenha três níveis (RGB), isso ocorra para toda profundidade da imagem de entrada. Contudo, como as imagens utilizadas neste trabalho são em escala de cinza, a convolução é realizada apenas em um canal.

Os filtros utilizados na primeira camada convolucional são inicializados geralmente com medidas de uma distribuição normal com ganho unitário, em que cada filtro percorre toda imagem, gerando assim uma saída com um tamanho da quantidade de filtros utilizada. Este tipo de inicialização se mostra adequado ao problema segundo [23] e serão ilustrados no resultados do trabalho.

Rectifier Linear Unit (ReLU)

Após a camada convolucional, vem a aplicação de uma unidade não-linear. Nos trabalhos iniciais propostos por LeCun, utilizava-se a tanh como não-linearidade, porém estudos recentes mostram que a sua convergência de otimização é mais lenta comparadas com outras funções não-lineares. Uma outra função mais eficiente e amplamente utilizada é chamada de Unidade Linear Retificante (ReLU), que foi definida por [27], utilizada em vários estudos como em [45] e comprovado sua eficiência por [33]. Na Figura 2.5 podemos ver a comparação da ReLU com a tanh, onde é possível observar uma melhora da convergência em até 6 vezes.

Nesta abordagem todos valores negativo de neurônios são zerados nas camadas de aprendizado, que é definida por

$$\varphi(\mathbf{x}) = \max(0, x). \tag{2.7}$$

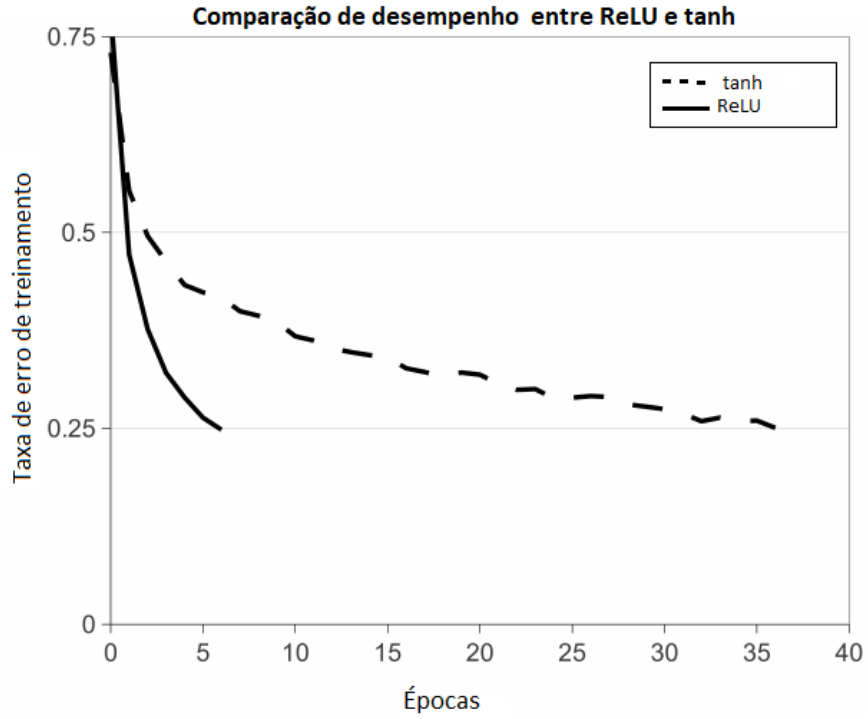

Figura 2.5. Adaptado de: [33]. Comparação entre ReLU e tanh realizada por Krizhevsky para descrever a melhor eficiência da unidade retificadora. Nela é possível observar que para chegar em uma mesma taxa de erro, a tanh demora cerca de 6 vezes na resolução do mesmo problema.

MaxPooling

Saindo da unidade não-linear, é aplicado uma camada de *pooling*, que em estudos recentes é comumente utilizado um *MaxPool*. A camada de *MaxPooling* serve para diminuir a dimensionalidade da camada anterior, reduzindo a quantidade de parâmetros de aprendizado conforme o tamanho da janela. Esta redução da dimensionalidade é realizada conforme o exemplo da Figura 2.6. Neste exemplo, é utilizado um *maxpool* de tamanho (2x2) com um passo de 1 e pega-se o maior valor em cada área de corte, tendo em vista que, na maioria dos casos reais, este é o tamanho mais utilizado para a redução. Deste modo, a dimensionalidade da camada é reduzida pela metade, reduzindo a complexidade de aprendizado dos parâmetros e diminuindo a possibilidade de *overfitting*.

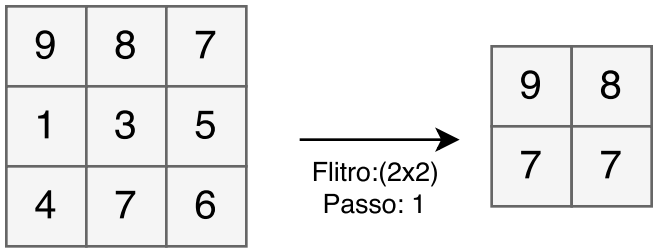

Figura 2.6. Exemplo de *MaxPool* utilizando um filtro de tamanho (2x2) e um passo de 1. Neste exemplo simples se observa que a dimensionalidade da da camada foi reduzida pela metade ao se escolhes os maiores elementos de cada corte da matriz. É importante observar ainda que apenas a dimensionalidade da imagem é reduzida, e não o tamanho da profundidade da camada, ou seja, a quantidade de filtros treinados na camada convolucional continua a mesma

Função de perda - Conexão total

Depois do uso de sucessivas multicamadas de aprendizado, a última camada de uma arquitetura de aprendizado utilizando CNN é a camada de conexão total, chamada de *Fully-Connected*. Essa camada serve para nos informar a qualidade da rede, apontando uma probabilidade de acerto para a tarefa designada. Para avaliar a qualidade do treinamento e da resposta do classificador, utiliza-se uma função de perda P, comparando o resultado final com o rótulo designado, ou seja, todos os neurônios da camada anterior são alimentados nessa função, que fornece um solução de compromisso quantificável entre os valores previstos pelo algoritmo e os valores reais. Podemos descrever este comportamento como,

$$P(y, x) = \frac{1}{m} \sum_{i=1}^{m} (x_i - y_i)^2, \tag{2.8}$$

em que x é o valor que nós queremos de saída, e y como sendo o valor que realmente obtivemos da rede.

Um algoritmo bastante utilizado na literatura é a função de perda de *Softmax*, que pode ser descrita como a generalização da Regressão Logística binária para problemas multi-classe, e é definida por,

$$P(\mathbf{x})_i = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}}, \tag{2.9}$$

em que a saída do *Softmax* é resultado da distribuição normalizada de probabilidade ao longo dos K valores de saída, sendo o somatório de todas probabilidades igual a 1.

Dropout

Para evitar o treinamento excessivo do classificador, algumas técnicas foram introduzidas por [30] com o intuito de reduzir a quantidade de elementos de aprendizado. Esta técnicas consiste em zerar a saída de cada neurônio da camada de aprendizado com uma certa probabilidade, portanto não contribuindo para a sequência de classificação da arquitetura. Um exemplo $2d$ desta técnica é descrito na Figura 2.7, onde é possível analisar que com a redução de neurônios nas camadas escondidas, a complexidade da rede diminui consideravelmente.

Está técnica se baseia no fato que os pesos em cada camada de aprendizado dividem informação entre eles, havendo uma dependência em cada camada. Ao atribuir alguns neurônios para 0, os neurônios remanescentes não terão mais uma relação com os zerados, forçando-os a representar características mais robustas. Usualmente a camada de *Dropout* é utilizada antes e depois da camada de conexão total, evitando o *overfitting*.

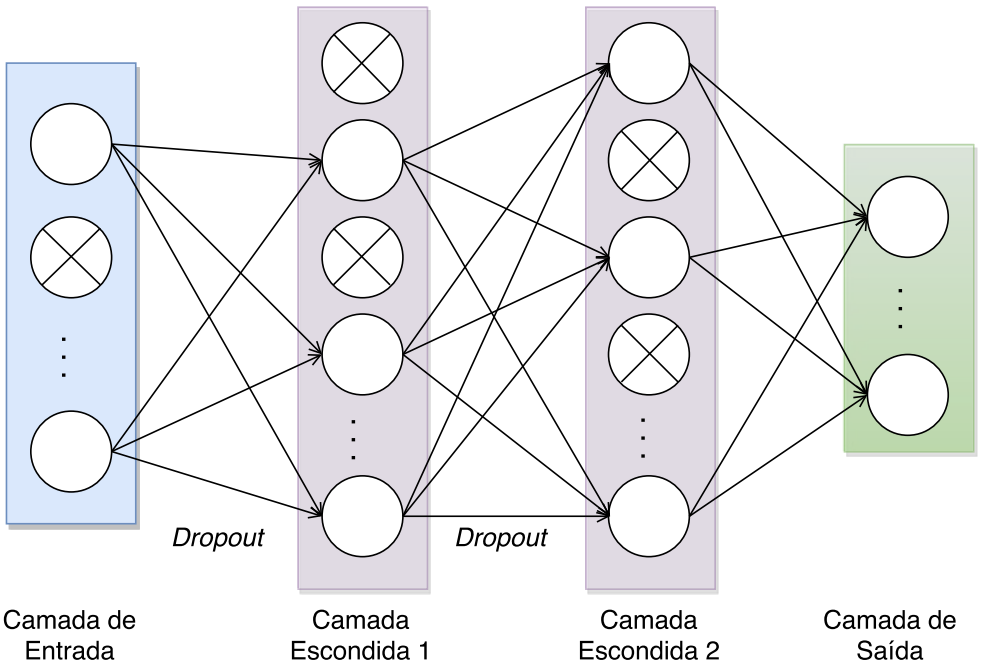

Figura 2.7. Exemplo do uso de *Dropout* em um sistema $2d$ com um probabilidade de 0.5 de exclusão dos neurônios. Com isso a complexidade e números de elementos de aprendizado diminui pela metade, evitando um *overfitting*.

2.4.2 Treinamento

Ao se iniciar a CNN, os pesos θ do filtro são inicializados de forma aleatória, que no caso do presente trabalho é iniciado com uma distribuição normal com ganho unitário. Após passar por todas camadas de aprendizado descritas (Convolução, Propriedade não-linear e Pooling), chega-se na última camada em que estão presentes os elementos de ativação, ou seja, as características que a rede encontrou como sendo relevantes para o problema. Para avaliar a qualidade do treinamento e da resposta do classificador, utiliza-se uma função de perda P, comparando o resultado final com o rótulo designado,

$$P(x, y) = \frac{1}{m} \sum_{i=1}^{m} (x_i - y_i)^2 \tag{2.10}$$

em que m representa o número de exemplo de cada treinamento, x a os elementos de treinamento e y o respectivo rótulo.

Em algoritmos de aprendizado de máquina, o objetivo é minimizar a função de perda para diminuir a probabilidade de erro. Uma função comumente utilizada é o *Gradient Descent*. Assim como em outros algoritmos de otimização numérica, o *gradient descent* tem como objetivo encontrar um mínimo de uma função P. Este mínimo é computado ajustando os pesos referentes aos erros e ao gradiente do erro computados nas saídas. Assim, este cálculo é repetido durante uma quantidade determinada de épocas κ à uma taxa de aprendizado α computando a cada iteração um novo valor para o erro [25]. Este tipo de método pode ser sumarizado pela expressão,

$$\theta = \theta - \alpha . \nabla_\theta P(\theta). \tag{2.11}$$

No caso da CNN, esta taxa de aprendizagem geralmente é reduzida a cada iteração produzida durante um número fixo de épocas, em que seu valor inicial nos casos mais recentes está em torno de $\alpha = 0.001$, assim como em[41].

Porém, este tipo de otimização numérica não é recomendada, pois para cada atualização nos pesos do algoritmo, é necessário calcular o gradiente de todo conjunto de dados, e isso possui um custo computacional muito elevado, em muitos casos não sendo possível realizar por falta de memória nos dispositivos.

Deste modo, utiliza-se um método de *gradient descent* em mini-lotes. Neste caso, ao invés de calcular o gradiente da função de perda de todos elementos de entrada, calcula-se apenas de uma parte dos dados. Em um caso específico do *gradient descent* em mini-lotes e que o o número de elementos do lote $m = 1$, chamamos de *gradient descent* estocástico (SGD)[25].

O algoritmo pode ser ilustrado pelo seguinte pseudocódigo, em que temos um mini-lote

de tamanho m do grupo de treinamento $(x_1, x_2, ..., x_m)$ com o respectivo rótulo y_i.

---

function GRADIENT DESCENT ESTOCÁSTICO($\alpha > 0$)
  $n \leftarrow 1$
  while $n < \kappa$ **do**
    $g \leftarrow \frac{1}{m}\nabla_\theta \sum_{i=1} P(f(x_i;\theta), y_i)$
    $\theta \leftarrow \theta - \alpha.g(\kappa)$
    $\kappa \leftarrow \kappa + 1$
  end while
  return θ
end function

Apesar desta técnica ser muito bem fundamentada e utilizada amplamente, em alguns momentos, como em caso de grandes curvaturas, ela pode ser lenta. Portanto, ainda é possível utilizar técnicas para acelerar os vetores gradientes para a direção correta, acelerando a convergência da função de perda para o valor mínimo, aumentando a precisão do algoritmo com mais rapidez [25].

Uma dessas técnicas chama-se momento. Para tal, é introduzida uma variável de momento σ que multiplica o decaimento negativo do gradiente, deste modo aumentando exponencialmente a convergência da função. Essa função pode ser descrita por,

$$g = \frac{1}{m}\nabla_\theta \sum_{i=1} P(f(x_i;\theta), y_i), \tag{2.12}$$

em que a velocidade v é atualizada a cada época, é descrita por,

$$v = \sigma.v - \alpha.g(\kappa), \tag{2.13}$$

atualizando os parâmetros de aprendizado, ou pesos, θ,

$$\theta = \theta + v. \tag{2.14}$$

Um exemplo do momento está representado na Figura 2.8, em que é possível observar que em regiões de grande curvatura, a abordagem tradicional de SGD, à esquerda, move-se em pequenos passos, sendo lento para algumas aplicações. Já no caso da aplicação de momento para o SGD, esse decaimento é muito mais rápido, convergindo para o valor esperando com menos iterações.

O algoritmo pode ser ilustrado pelo seguinte pseudocódigo, em que temos um mini-lote de tamanho m do grupo de treinamento$(x_1, x_2, ..., x_m)$ com o respectivo rótulo y_i, além do valor do momento σ.

function GRADIENT DESCENT ESTOCÁSTICO COM MOMENTO($\alpha > 0, \sigma > 0$)
  $n \leftarrow 1, v \leftarrow init$
  while $n < \kappa$ **do**
    $g \leftarrow \frac{1}{m}\nabla_\theta \sum_{i=1} P(f(x_i;\theta), y_i)$
    $v \leftarrow \sigma.v - \alpha.g(\kappa)$
    $\theta \leftarrow \theta + v$
    $\kappa \leftarrow \kappa + 1$
  end while
  return θ
end function

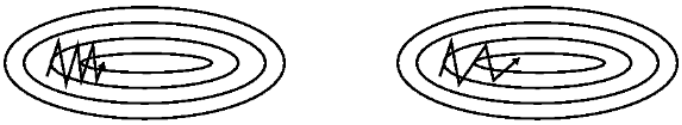

Figura 2.8. Exemplo do Momento para SGD. À esquerda é possível observar a convergência do algoritmo de SGD sem o uso do momento. Observa-se que em casos em que a região possui grande curvatura, a abordagem tradicional de SGD, move-se em pequenos passos, sendo lento para algumas aplicações. À direta observamos a aplicação do momento para SGD, em que a convergência ocorre de forma mais rápida. Adaptado de [47]

É necessário observar o tamanho de σ, pois em casos de sigma muito grande, o passo pode ser muito maior que a possível convergência do SGD, ou seja, em casos em que o valor do gradiente da função de perda é muito baixo, porém sem chegar no mínimo local da função. Nas arquiteturas atuais, utiliza-se um momento de $\sigma = 0.9$, ou seja, para cada nova iteração, a convergência acontece 10 vezes mais rápida que se não usarmos o valor de momento.

Como discutido no decorrer desta seção, o SGD com momento primeiramente calcula-se o gradiente na posição atual e então aplica-se um momento σ na direção atualizada do gradiente acumulado. Porém, recentemente, [55] introduziu uma nova técnica para auxiliar a otimização com um modelo mais eficiente. Esta nova técnica para aumentar a eficiência do momento chamado de momento de Nesterov, em que a principal diferença é onde o cálculo do gradiente é realizado. Ao contrário do momento tradicional, no momento de Nesterov, o gradiente é calculado após o salto do momento, isto é, a velocidade do decaimento do gradiente é atualizado após a aplicação do momento. Então, temos que a velocidade v atualizada a cada época, é descrita por,

$$v = \sigma.v - \alpha\nabla_\theta[\frac{1}{m}\sum_{i=1} P(f(x_i;\theta + \sigma.v), y_i)], \tag{2.15}$$

atualizando os parâmetros de aprendizado θ,

$$\theta = \theta + v. \tag{2.16}$$

A diferença do comportamento entre o momento clássico e o momento de Nesterov é representada na Figura 2.9. Á esquerda, é possível observar o momento clássico, em que primeiramente é calculado o gradiente no ponto θ_t e então o momento na direção v_t com uma taxa σ. No momento de Nesterov, primeiramente é realizado um salto na direção v_t com uma taxa σ e então é aplicado um gradiente resultante de $\theta_t + \sigma.v_t$.

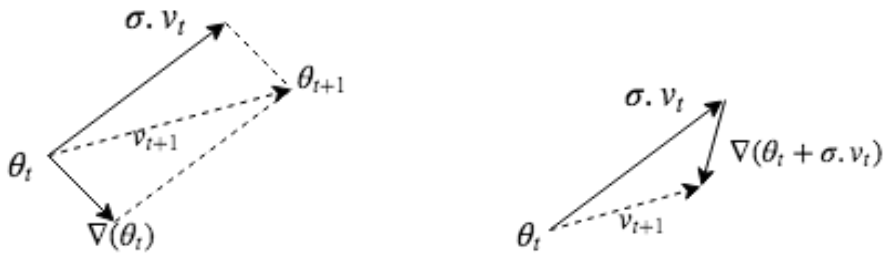

Figura 2.9. Exemplo do Momento Clássico e do Momento de Nesterov. Á esquerda, é possível observar o momento clássico, em que primeiramente é calculado o gradiente no ponto θ_t e então o momento na direção v_t com uma taxa σ. Á direita é possível observar o momento de Nesterov, em que primeiramente é realizado um salto na direção v_t com uma taxa σ e então é aplicado um gradiente resultante de $\theta_t + \sigma.v_t$. Adaptado de [55]

O algoritmo pode ser ilustrado pelo seguinte pseudocódigo, em que temos um mini-lote de tamanho m do grupo de treinamento$(x_1, x_2, ..., x_m)$ com o respectivo rótulo y_i, além do valor do momento σ.

---

function GRADIENT DESCENT ESTOCÁSTICO COM MOMENTO DE NESTEROV($\alpha > 0, 0 < \sigma < 1$)
  $n \leftarrow 1, v \leftarrow init$
  while $n < \kappa$ **do**
    $\hat{\theta} \leftarrow \theta + \sigma.v$
    $g \leftarrow \frac{1}{m}\nabla_\theta \sum_{i=1} P(f(x_i; \hat{\theta}), y_i)$
    $v \leftarrow \sigma.v - \alpha.g(\kappa)$
    $\theta \leftarrow \theta + v$
    $\kappa \leftarrow \kappa + 1$
  end while
  return θ
end function

Portanto, segundo [25], ao se utilizar um valor de aprendizado muito grande, o valor do

gradiente pode flutuar, por isso com um valor baixo de aprendizado, combinado com um valor alto para o momento, resulta em uma proporção eficaz para a resolução do problema que se queira resolver. Sendo assim, veremos que no caso da resolução do problema deste trabalho utilizamos uma taxa de aprendizado $\alpha = 0.014$ e um momento de Nesterov $\sigma = 0.9$

3 Metodologia

Este capítulo descreve a metodologia de desenvolvimento do sistema de aprendizado de máquina e classificação de imagens, especificando a arquitetura desenvolvida por meio de diagrama de blocos, pseudocódigo e análise visual do sistema.

Apresenta ainda o passo-a-passo dos procedimentos experimentais para regularização das imagens de entrada com objetivo de atender às necessidades da arquitetura desenvolvida e gerar reprodutibilidade da metodologia. Um outro ponto apresentado é o tipo de banco de dados utilizado no projeto, descrevendo a quantidade de indivíduos em cada grupo de controle e esquizofrênico, incluindo os critérios de inclusão e exclusão para o processo de seleção.

De posse dos resultados obtidos nos experimentos, o capítulo aponta as métricas de qualidade utilizadas, descrevendo os testes estatísticos desenvolvidos para comparação, e ainda a quantidade de iterações para cada grupo de treinamento e testes, assim como a proporção de dados de cada grupo.

3.1 Desenvolvimento do sistema de aprendizado de máquina e classificação de imagens

3.1.1 Arquitetura Geral

A arquitetura do sistema proposto está apresentado de forma reduzida na Figura 3.2. Por se tratar de um banco de dados com poucas imagens para treinamento da rede, optou-se por construir uma rede não muito profunda como desenvolvido em outros trabalhos de aprendizado profundo [56], evitando o *overfitting*. A arquitetura consiste em 4 camadas de aprendizado com pesos, sendo 2 camadas de convolução seguidas por funções não-lineares, e duas camadas de conexão total. O resultado da última camada de conexão total alimenta a camada de saída com uma função de *softmax* de 2 variáveis que queremos identificar, esquizofrênico ou não-esquizofrênico.

Assim como descrito na Seção 3.2 e observado na Figura 3.1, as imagens de entrada do sistema possuem tamanho 100x100, mas uma vez que passam pela camada de convolução com 32 filtros de tamanho 5x5, passo de 1 e *zero-padding* de 0, o tamanho é reduzido para 96x96 para que seu arranjo espacial seja mantido, eliminando *pixels* das bordas. Uma vez que as imagens de MRI utilizadas possuem bordas pretas irrelevantes ao problema, este

corte não influencia no resultado, mas sim diminui os parâmetros de aprendizado, reduzindo o *overfitting*.

Ainda na primeira camada de convolução, o resultado passa por uma função não-linear. A não-linearidade utilizada foi a de Unidade Linear Retificada (ReLU), que assim como mostrada em 2.4.1, possui um treinamento muito mais rápido em comparação com outras funções não-lineares como a *tanh*. O resultado então passa por uma redução na dimensionalidade utilizando a função de *maxpool* de tamanho 2x2, reduzindo pela metade os parâmetros de aprendizado e deixando a rede com um tamanho de 32x48x48.

Seguindo nas camadas de aprendizado, outra camada de convolução de tamanho 3x3 é aplicada ao resultado da camada anterior, resultando com 64 imagens filtradas e utilizando a função de ReLU. Deste modo resultando em uma dimensionalidade de 64x46x46, diminuindo em 2 o resultante da imagem pelo mesmo fato de manter o arranjo espacial. Na sequência é aplicado um *maxpool* de 2x2 e passo de 1.

Assim como discutido em 2.4.1, um fato comum à redes neurais é a presença de *overfitting* pelo excesso de parâmetros de treinamento, deste modo, foi introduzido uma camada de *dropout* com probabilidade de 0.5, excluindo aleatoriamente metade dos parâmetros de aprendizado. Esta técnica reduz a dependência entre os neurônios das camadas, ou parâmetros de aprendizado, forçando cada parâmetro conter informações mais relevantes e robustas, uma vez que não há mais codependência com as características dos neurônios excluídos.

É introduzida então a primeira camada densa de conexão completa, contendo 100 unidades de ativação. A camada densa é parecida com a camada de conexão total descrita na Seção 2.4.1, no qual é realizada uma operação linear em que todos elementos da camada anterior são conectados com todos unidades de ativação por um certo peso. Estes pesos são resultantes de uma desatribuição normal com ganho de 1. A camada densa ainda é seguida por uma operação não-linear de ReLU.

Um outro *dropout* de 0.5 é utilizado para reduzir os parâmetros e aumentar a robustez dos neurônios de aprendizado. A saída desta camada passa finalmente por um *softmax*, gerando dois resultados. Estes resultados apresentam um peso para cada uma das variáveis que queremos identificar, apontando para a solução do sistema.

Finalmente, para o algoritmo de otimização numérica, utilizamos um método de *Gradient Descent* estocástico combinado com um momento de Nesterov, utilizando 250 épocas, uma taxa de aprendizado $\alpha = 0.014$ e um momento de Nesterov $\sigma = 0.9$.

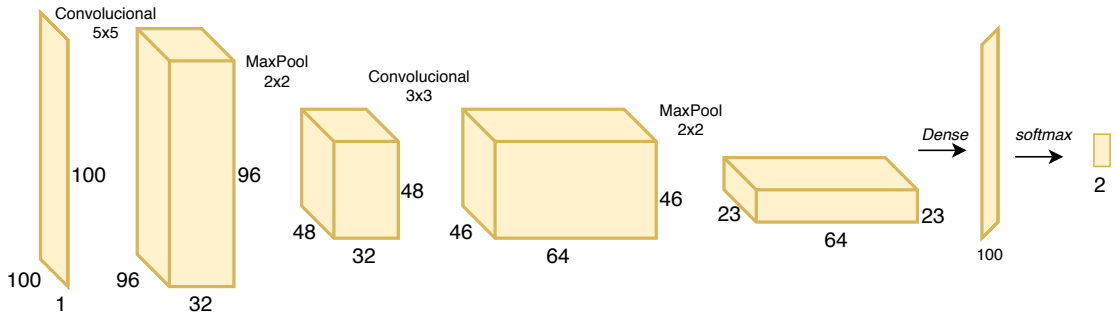

Figura 3.1. FayNet: Uma adaptação do sistema proposto por LeCun utilizando a LeNet [38]. A arquitetura utiliza camadas de *dropout* antes e depois da camada densa de conexão para evitar *overfitting*, além de utilizar ReLUs como propriedades não-lineares nas camadas convolucionais.

3.2 Metodologia Experimental

Uma vez que a arquitetura da rede de treinamento foi desenvolvida, é necessário organizar e regularizar as imagens de entrada afim de atender às necessidades do sistema. Para tal, é desenvolvido uma série de procedimentos experimentais para gerar reprodutibilidade além de resultados regulares e coesos.

Dada a posse do banco de dados, selecionamos os cortes axiais 13 e 14, que são os cortes que melhore evidenciam os ventrículos laterais, de todos indivíduos de controle e esquizofrênicos. A escolha dos cortes foi uma primeira tentativa de solucionar o sistema, uma vez que tais cortes representam de forma mais clara os ventrículos e caixa cerebral, apontados pela literatura como sendo os maiores responsáveis por produzir resultados relevantes, deste modo totalizando 152 imagens de controle e 160 imagens de indivíduos esquizofrênicos.

A fim de diminuir a complexidade e demora no treinamento da rede, a dimensionalidade das imagens foram reduzidas para 100x100, mantendo a proporção e utilizando funções de *antialias* com a filtragem *Lanczos*. Durante testes descobrimos que a utilização de tamanhos maiores não interferiam no resultado final do treinamento, porém deixam a rede mais pesada, ou seja, com mais dados a serem treinados e funções a serem realizadas, além de um consumo maior de memória RAM.

Com as imagens reduzidas e em tom de cinza (apenas um canal), podemos selecionar os dados que serão utilizados para treinamento e validação da arquitetura. Em uma primeira analise, utilizamos o método de *holdout*, ou seja, criamos dois vetores que recebem a seleção aleatória de 70% das imagens para treinamento das duas classes, e mais dois vetores rotulando 1 para a classe dos indivíduos com esquizofrenia e 0 para a classe de controle. Uma vez que temos vetores de treinamento criados, concatenamos os valores das duas classes para que haja um vetor único com pacientes esquizofrênicos e de controle assim como um vetor com

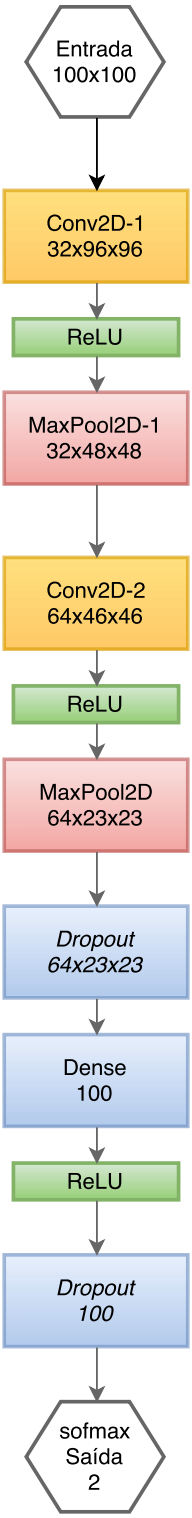

Figura 3.2. Descrição completa da arquitetura proposta. Nela é possível observar a dimensionalidade de cada camada da rede, demostrando a quantidade e tamanho dos filtros utilizados na convoluções, assim como o uso da propriedades não-lineares como a ReLU e a *softmax*. O resultado final é o diagnóstico entre esquizofrênico ou não, representado por pesos para cada classe.

seus devidos rótulos.

Com o vetor de treinamento criado, rotulado e embaralhado, normalizamos os valores, subtraindo a média das imagens indicadas pelo vetor e dividindo pelo desvio padrão, obtendo assim imagens entre entre -1 e 1, diminuindo a oscilação dos valores, consequentemente deixando a rede mais fácil de ser treinada. Finalmente os vetores são embaralhamos aleatoriamente utilizando funções que mantêm os rótulos iniciais, de forma que os indivíduos fiquem misturados sem perder o rótulo, colocando-os na arquitetura proposta. Os mesmos procedimentos foram adotados para o vetor de validação, utilizando os 30% restantes das imagens de entrada.

Um segundo método experimental desenvolvido foi a utilização de *k-fold* com k=20 e sua construção será descrita em 3.3. Neste método, 20 grupos com a mesma quantidade de imagens de esquizofrênicos e controle são separados para realizar o treinamento. São realizados 20 treinamentos distintos, em cada um deles um grupo dos 20 é separado para teste enquanto os outros são utilizados para treinamento. A forma de criação dos vetores, rotulação, normalização e embaralhamento é realizada da mesma maneira com a descrita para o método de *holdout*.

Para cada repetição do modelo de aprendizagem, seja ele por meio de *holdout* ou *k-fold*, a arquitetura utilizou 250 épocas, sendo utilizado em cada época um lote de tamanho 128 com 80% dos valores para o treinamento interno da rede e otimizando os pesos com os 20% restantes, para tentar diminuir possíveis ruídos e convergir para o resultado exato.

Foi desenvolvido ainda um algoritmo para a transferência de conhecimento para outros classificadores. Para tal, criamos dois vetores com as saídas das camadas internas de Maxpool e densa para servirem de entrada de treinamento nos novos classificadores. Os resultados destes classificadores, então, são comparadas com os mesmos vetores de validação utilizados no *k-fold* da CNN para que se possa ter uma comparação dos dois métodos.

Todos algoritmos foram desenvolvidos em *Python* utilizando um o compilador para expressões matemáticas chamado *Theano*. Ele é utilizado como uma representação matemática simbólica parecida com a representação padrão *NumPy*, porém as expressões são otimizadas para utilização de aprendizado de máquina, sendo o seu uso com GPU's e CPU's mais eficiente do que a utilização padrão [6].

Outros pacotes como o *Lasagne* [18] e *nolearn* [46] foram utilizados para criação e treinamento da rede neural, e extração das características das camadas internas da rede. Lasagne é um framework dedicado ao treinamento de redes neurais utilizando a biblioteca de funções matemáticas Theano, que por sua vez serve para definir e otimizar eficientemente funções matemáticas multi dimensionais utilizando GPU [58].

Por fim, utilizou-se o SciKit para auxiliar o desenvolvimento das métricas de análise como

a matriz de confusão e para a construção dos classificadores SVM e do tipo *ensemble*.

Algorithm 1 Algoritmo de exemplo para a criação dos vetores de treinamento e validação

function GET IMAGES I(Proportion P)
  $N \leftarrow size_{set}$
  $rows \leftarrow 100$
  $columns \leftarrow 100$
  $d \leftarrow 0$
  while $j \geq N$ **do**
    $I_d \leftarrow j$
    $I_d \leftarrow resize((rows, columns))$
    $d \leftarrow d_{++}$
    $I_{total}(d, 0) \leftarrow I_d$
  end while
  $I_{total}(N, 0) \leftarrow shuffle$
  $numTrain \leftarrow N * P$
  $trainSet \leftarrow I_f(: numTrain)$
  $validationnSet \leftarrow I_f(numTrain :)$
  return $trainSet$, $validationSet$
end function

3.2.1 Banco de Dados

A iniciativa norte-americana chamada *Function Biomedical Informatics Research Network* (FBIRN) foi um programa fundado pelo Instituto Nacional de Saúde (do inglês NIH) com o intuito de desenvolver métodos, técnicas, ferramentas e protocolos para auxiliar na aquisição de imagens de ressonância magnética anatômica do cérebro em diversos centros de pesquisa, gerando assim um banco de dados diverso capaz de ser compartilhado com a comunidade científica e abrindo portas para pesquisas neste ramo.

A FBIRN consiste em 3 fases destinas. Na primeira fase, o objetivo foi encontrar diferenças entre os diversos centros de aquisição de imagens para que assim fosse possível desenvolver métodos de aperfeiçoamento, diminuindo as variações no protocolos de cada centro e aumentando a confiabilidade e homogeneidade das imagens obtidas [22]. As fases II e III são estudos realizados com indivíduos com esquizofrenia ou desordem esquizoafetiva e indivíduos saudáveis executados em diversos centros de aquisição. Nestes estudos, devido ao aperfeiçoamento dos métodos e protocolos desenvolvidos na fase I do projeto, a diferença de imagens de fMRI obtidos entre os diversos centros foi menor, criando um banco de dados uniforme e capaz de ser utilizado em pesquisas científicas.

As imagens utilizadas no presente trabalho fazem parte do banco de imagens de fMRI

obtidos na fase II do projeto FBIRN. Esta fase consiste de 87 indivíduos previamente diagnosticados clinicamente com esquizofrenia ou desordem esquizoafetiva por um médico, ou grupo de médicos, com base nos sintomas apresentados no manual DMS-IV, sendo 59 homens e 28 mulheres. Outros 85 indivíduos saudáveis, sendo 70 homens, foram utilizados para controle e comparação das imagens. Todos participantes tinham idade entre 18 e 70 anos, não possuíam contraindicações para estudos de MRI. Além de haver critérios de exclusão referentes aos indivíduos esquizofrênicos que possuíam histórico de outras doenças complexas, sintomas extrapiramidais significativos ou que apresentavam mudanças consideráveis em seu quadro clínico nos últimos dois meses. O critério de exclusão dos indivíduos de controle estão relacionados à doenças neurológicas graves relatados em seu histórico ou em parentes de primeiro grau.

Nos diversos centros de aquisição, os *scanners* de MRI possuíam campos variando de 1.5T a 4T, e cada sessão de captura foi realizada durante 1 hora e 30 minutos, além de ser repetida entre 24 horas e 3 semanas depois da primeira consulta. Estas imagens consistem em 27 cortes axiais do encéfalo humano ponderadas em T_2 com resolução de 256x256.

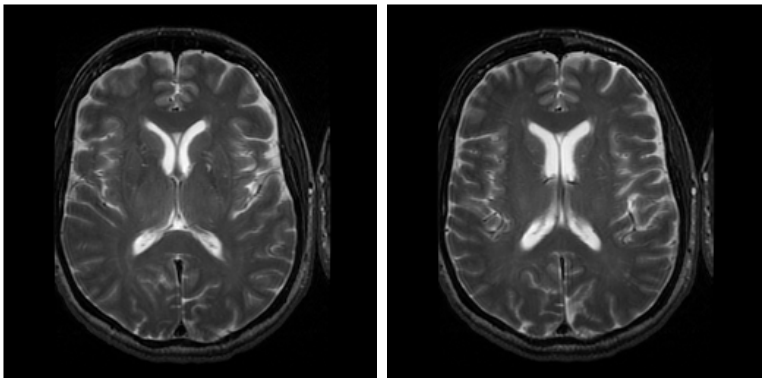

Figura 3.3. Exemplo de imagens do banco de dados dos cortes axiais 13 e 14 de um indivíduo de controle. Nele é possível notar os ventrículos laterais e outras partes do encéfalo.

3.3 Métodos de avaliação do modelo e estimação da acurácia

Métodos de avaliação para modelos de aprendizagem supervisionados são essenciais para medir, validar e comprovar a eficiência e acurácia da predição. Neste tipo de aprendizagem tem-se como objetivo comparar o resultado da predição com o valor esperado do rotulo em questão. Neste sentido, este capítulo visa apresentar os métodos de avaliação utilizados para validação da acurácia da arquitetura desenvolvida a fim de diminuir a variância e aumentar a confiabilidade final. A avaliação e analise de desempenho de uma arquitetura de classificação se dá pela capacidade do modelo de generalizar o problema, podendo ser usado para aplicação e classificação de novos dados.

Porém, um problema recorrente que aos algoritmos de aprendizado de máquina em geral é a presença de *overfitting*, isto é, o modelo aprende apenas singularidades e características relativas especificamente ao conjunto de treinamento, irrelevante à classificação desejada, devido ao uso de várias transformações não-lineares [64], podendo levar classificações erradas ao se deparar com um novo conjunto de dados [54]. Portanto, um sistema de validação consistente mostra-se necessário para validar os dados de saída do sistema.

Pelo fato da quantidade de imagens para treinamento e validação ser pequena, o projeto utilizou duas metodologias de análise não-exaustivas de validação-cruzada para estimar a acurácia do resultado obtido, são elas: *holdout* e *k-fold*.

3.3.1 *Holdout*

No método de avaliação de *holdout*, o conjunto de dados é divido em dois grupo mutualmente exclusivas chamadas grupo de treinamento e grupo de validação, ou *holdout*. Segundo [32], usualmente atribui-se à classe de treinamento $2/3$ do total de conjunto de dados e o restante é utilizado para validar o modelo treinado, mantendo a proporcionalidade de ambas classes a serem classificadas. A definição matemática da acurácia deste método é definida por,

$$acu_t = \frac{1}{t} \sum_{(v_i, y_i) \epsilon D_h} \delta(I(D_t, v_i), y_i), \tag{3.1}$$

assumindo D_h como grupo de *holdout* de tamanho h e D_t como grupo de treinamento, em que $\delta(i, j) = 1$ se $i = j$, caso contrário $\delta(i, j) = 0$. A medida v_i representa um valor não rotulado para um valor rotulado de y_i e I representa o modelo de treinamento para classificação.

Como apresentado por [32], o método de *holdout* é um estimador pessimista, pelo fato de quanto maior for o tamanho do grupo de *holdout*, maior será o valor de viés, ou seja, maior

será a distorção aleatória da amostra estatística. Por outro lado, quanto menor for o grupo de *holdout*, maior será o dispersão do valor da acurácia, ou seja, em um caso extremo de *leave-one-out* em que são utilizado todos - 1 elementos do conjunto de dados para treinamento e apenas 1 para validação, o resultado será 100% ou 0% de acurácia. Para este método ser treinado e validado com diferente grupos de treinamento e teste, é necessário a repetição de n vezes, em que cada nova etapa de treinamento é realizada com um grupo aleatório do conjunto de dados de entrada, e realizar a média do valor total da acurácia pelo número de repetições n. Este método é chamado de *holdout* com reamostragem aleatória, e mostra-se muito eficiente em estimar a acurácia real do modelo de aprendizagem [32].

Este trabalho utilizou o método de *holdout* com reamostragem aleatória empregando uma divisão de 70% da imagens para treinamento e 30% para validação, além de um $n = 530$, que segundo [32], um valor de $n = 500$ é o padrão ouro para testes estatísticos. Desta forma é possível levantar um histograma dos resultados das predições e obter um resultado mais robusto, em que é possível observar o valor da convergência das médias dos resultados de acurácia.

Esta média simples é calculada utilizando o resultado da validação de cada um dos métodos de *holdout*, R_i, da arquitetura proposta com 250 épocas, e μ é a média destes i resultados, ou seja

$$\mu = \frac{1}{n} \sum_{i} R_i \tag{3.2}$$

sendo n é o número total de treinamentos realizados, e σ é o desvio padrão destas i validações, sendo

$$\sigma = \sqrt{\sum_{i} \frac{(R_i - \mu)^2}{n(n-1)}}. \tag{3.3}$$

O histograma completo dos resultados obtidos pela classificador utilizando a arquitetura proposta está representado na Figura 4.2. Outras arquiteturas foram construídas a fim de observar variâncias nos resultados da acurácia utilizando tamanhos e quantidades variados de filtros. O histograma destes resultados estão descritos nas Figuras 4.3, 4.4 e 4.5 .

3.3.2 *k-fold*

Outra metodologia de análise desenvolvida e utilizada neste trabalho é a de validação-cruzada com *k-fold*. Este tipo de método assim como o de *holdout* é chamado de método não-exaustivo pois não computa todas possibilidade de distribuição do conjunto de dados, e sim utiliza um estimação da probabilidade da acurácia. Neste tipo de método, o conjunto de dados é aleatoriamente dividido em k grupos mutualmente excludentes de tamanhos iguais, mantendo a proporcionalidade de ambas classes a serem classificadas. Após essa divisão,

são utilizados $k-1$ grupos para treinamento do classificado e o grupo remanescente para validação. Este processo é repetido k vezes para contemplar todas possibilidades de uso dos grupos divididos, em que valor estimado da acurácia do método acu_m é dado pela média do número total de acertos de todos grupos pelo número N de elementos no conjunto de dados. Este comportamento pode ser definido por,

$$acu_m = \frac{1}{N} \sum_{(v_i, y_i) \epsilon D} \delta(I(D \setminus D_i, v_i), y_i), \tag{3.4}$$

assumindo D_i como grupo de treinamento da veze D como sendo o valor total conjunto de dados, ou seja, a cada k grupo o modelo de treinamento I treina a rede utilizando todos valores D excluindo D_i e testa para os valores dos elementos não rotulados v_i. O somatório dos acertos de todos k grupos de treinamento é dividido então pelo número total de elementos no conjunto de dados. Um exemplo de validação cruzada de 5 grupos está explícito na Figura 3.4.

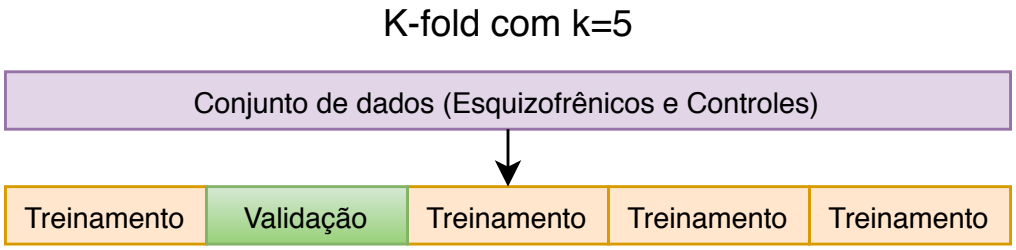

Figura 3.4. Exemplo de Validação-Cruzada com k=5. Ela é composta por quatro camadas de treinamento de uma camada de validação repetido por k vezes, em que cada vez um grupo diferente é utilziado para validação e os outros para treinamento.

Para validação do modelo desenvolvido no trabalho, utilizou-se um *k-fold* de tamanho 20, ou seja, 20 grupos de tamanhos iguais com os valores do conjunto de dados distribuídos de forma aleatória e mantendo a proporcionalidade dos indivíduos esquizofrênicos e de controle. Dentro destes grupos, 19 são utilizados para treinamento e 1 grupo para validação, repetidos por k vezes a fim de contemplar a utilização de todos grupos para treinamento.

3.3.3 Métricas de desempenho

Computar métricas de desempenho de um dado classificador de aprendizagem supervisionada é uma tarefa importante para avaliar o algoritmo desenvolvido. Além da acurácia citada anteriormente, outras métricas de desempenho como sensibilidade, especifidade e precisão são medidas importantes para se avaliar quando pretende-se validar e avaliar o resultado de forma completa.

Estas métricas podem ser retiradas de uma matriz chamada matriz de confusão, que representa o quanto o algoritmo "confunde" os resultados, sendo eles realizar a predição de um valor positivo quando na verdade era negativo (FP), predição de um valor positivo quando realmente era positivo (TP), predição de um valor negativo quando na verdade era positivo (FN) e finalmente a predição de um valor negativo quando realmente era negativo (TN).

Tabela 3.1. Tabela da Matriz de Confusão

		Previsão	
		Negativo	**Positivo**
Rótulo	**Negativo**	TN	FP
	Positivo	FN	TP

Neste contexto, a métrica de acurácia (AC) representa a proporção total das predições corretas, podendo ser descrita por,

$$AC = \frac{TN + TP}{n}, \tag{3.5}$$

sendo n o valor total do conjunto de dados, ou $TN + TP + FN + FP$.

A sensibilidade (S) pode ser descrita pela proporção de TP em relação à todos valores que realmente eram positivos, ou seja, a capacidade de prever corretamente nos casos em existe a presença da doença, descrita por,

$$S = \frac{TP}{TP + FN}. \tag{3.6}$$

Uma outra métrica relevante é a especifidade (E), que descreve o contrário da sensibilidade, ou seja, a capacidade de prever corretamente quando não existe a presença da doença em casos que realmente não há, e é descrito por,

$$E = \frac{TN}{TN + FP}. \tag{3.7}$$

Finalmente a métrica da precisão (P) é definida como a proporção de TP em relação à todos valores que o modelo descreveu como positivo, ou seja,

$$P = \frac{TP}{TP + FP}. \tag{3.8}$$

.

4 Resultados e Discussões

Este capítulo apresenta os resultados obtidos através da arquitetura desenvolvida em 3.1.1 utilizando os métodos de análise descritos em 3.3. Para tal, são apresentados histogramas da acurácia resultante de arquiteturas variadas, utilizando *holdout* com reamostragem aleatória para cada novo treinamento, métricas de desempenho utilizando *k-fold* para CNN, além da métricas de desempenho da SVM e outros classificadores do tipo *ensemble* com a transferência de conhecimento. São ainda apresentadas as imagens das saídas das camadas convolucionais a fim de observar as abstrações de cada nível.

A Figura 4.1 representa o fluxo geral da arquitetura implementada. Nela é possível observar exemplos das imagens filtradas em cada camada convolucional e ainda a diminuição de parâmetros de aprendizado com a utilização da operação de *maxpool* após as camadas convolucionais.

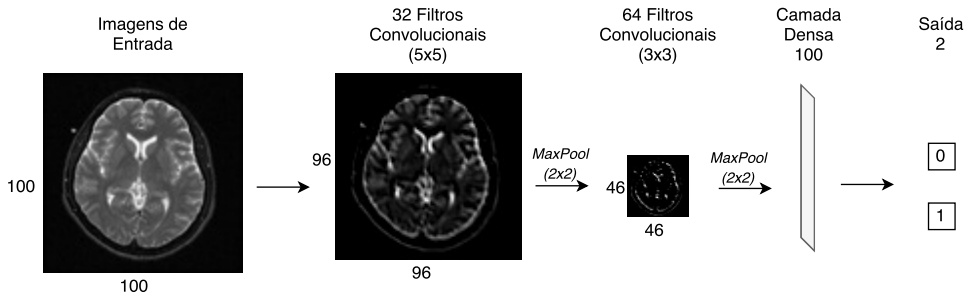

Figura 4.1. Resultado geral da arquitetura. Exemplo com uma imagem utilizada no treinamento da arquitetura onde é possível observar a continua abstração da imagem a medida em que a arquitetura se torna mais profunda, em que cada camada convolucional e de maxpool, a imagem se torna mais abstrata e com uma menor dimensionalidade, chegando finalmente à saída binária para determinar se o indivíduo possui a doença ou não.

Na Tabela 4.1 temos o resultado do tamanho de cada camada de aprendizado assim como o tamanho e quantidade de filtros utilizados em cada convolução. É possível observar o número total de parâmetros que foram aprendidos durante a classificação.

Utilizando a formula descrita em 2.6 é possível constatar o fato do tamanho da imagem para a primeira camada convolucional ter reduzido para 96×96, pois com um passo de 1 e um kernel convolucional de tamanho 5x5 temos que, $I = \frac{(100-5+2\times0)}{1} + 1 = 96$

Tabela 4.1. Tabela de informações da Arquitetura. É possível notar o tamanho e quantidade de cada filtros nas camadas convolucionais, assim como o passo e *zero padding* aplicados para cada filtro

Nome	Tamanho Filtro/Passo/ *zero padding*	Tamanho Saída	Dimensão Camada	Parâmetros de Aprendizado
input		1x100x100	10000	
conv2d1	5x5/1/0	32x96x96	294921	
maxpool1	2x2/1/0	32x48x48	73728	
conv2d2	3x3/1/0	64x46x46	135424	
maxpool2	2x2/1/0	64x23x23	33856	
dropout1		64x23x23	33865	
dense		100	100	
dropout2		100	100	
output		2	2	
Total				3405230

4.1 Resultado da Arquitetura Proposta

Em um primeiro momento, durante a escolha da arquitetura final do problema, foram levantas curvas prévias da acurácia com diferentes hiper-parâmetros a fim de comparar resultados e avaliar o melhor modelo. Os resultados destas arquiteturas foram gerados por meio de um *holdout* com reamostragem aleatória utilizando 200 iterações. O primeiro histograma gerado para a arquitetura proposta está apresentado na Figura 4.2.

Estas arquiteturas utilizadas durante a criação da arquitetura final utilizaram menos filtros convolucionais, uma camada densa com menos parâmetros de ativação, além de filtros maiores que percorriam uma quantidade maior da imagem, sendo menos específico em relação à uma área menor. Parâmetros como o tamanho do *maxpool* e quantidade de *dropout* foram iguais aos utilizados na arquitetura principal. Do mesmo modo da arquitetura original, o histograma foi gerado por 200 iterações na classificação de cada *set* de imagens e foi utilizado 200 épocas no algoritmo de otimização. Os histogramas representados nas Figuras 4.3,4.4 e 4.5 apresentam o tamanho e quantidade de filtros para cada caso.

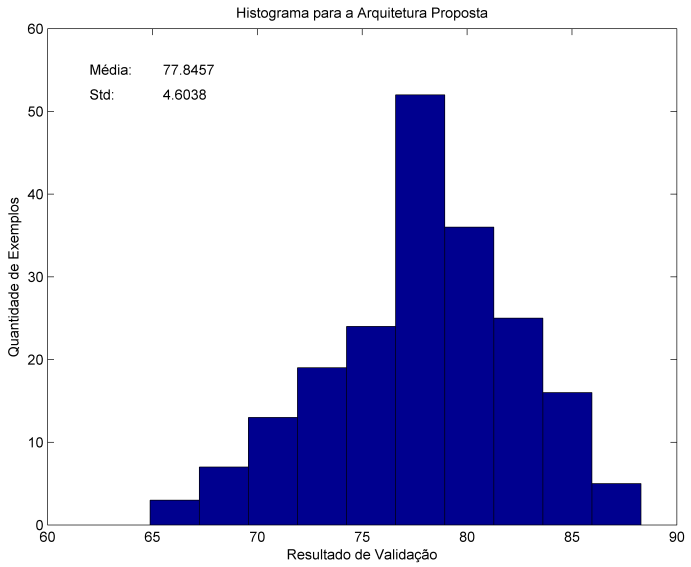

Figura 4.2. Histograma resultante das acurácias para arquitetura final utilizando um *holdout* de 200 iterações com reamostragem aleatória. Esta arquitetura resultou uma média de 77.85% de acurácia com um desvio padrão de 4.6

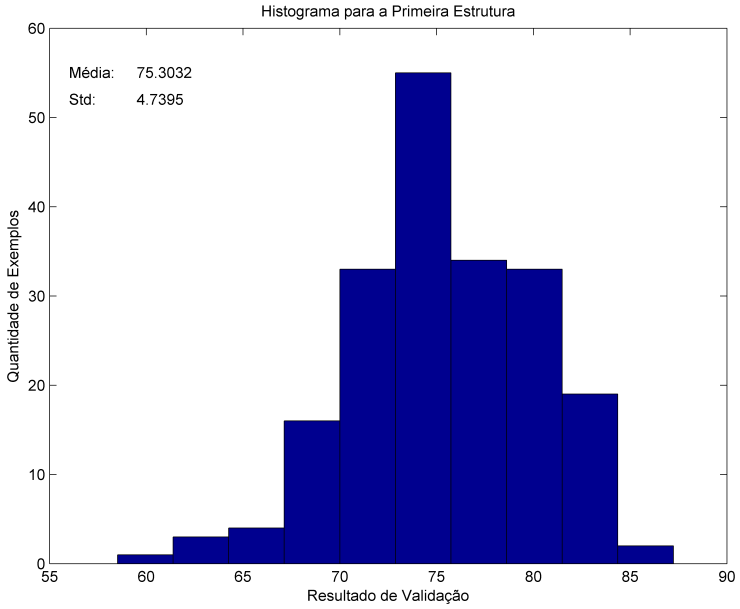

Figura 4.3. Histograma resultante das acurácias para a segunda arquitetura desenvolvida utilizando um *holdout* de 200 iterações com reamostragem aleatória. Esta arquitetura resultou uma média de 75.30% de acurácia com um desvio padrão de 4.74. Características da arquitetura: Primeiro Filtro Convolucional = 10x5x5 - Segundo Filtro Convolucional 8x5x5 - Camada densa de 20 unidades - 200 épocas.

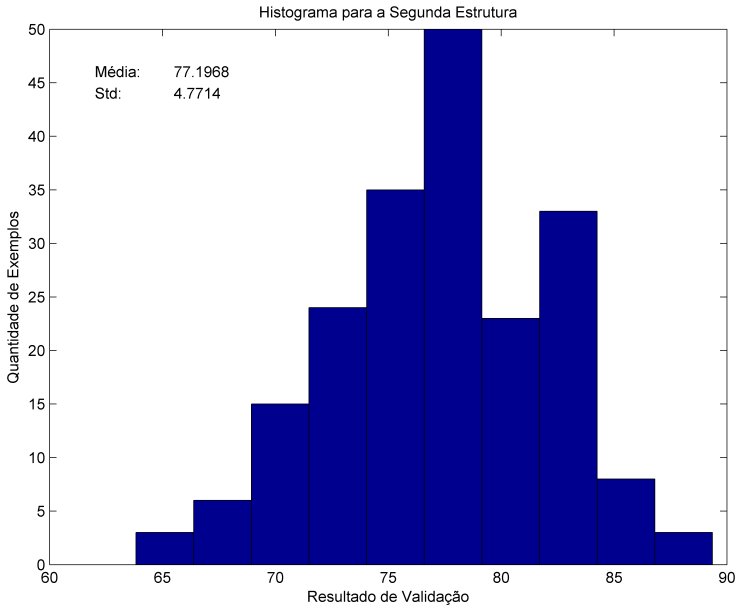

Figura 4.4. Histograma resultante das acurácias para a terceira arquitetura desenvolvida utilizando um *holdout* de 200 iterações com reamostragem aleatória. Esta arquitetura resultou uma média de 77.17% de acurácia com um desvio padrão de 4.77. Características da arquitetura: Primeiro Filtro Convolucional = 15x5x5 - Segundo Filtro Convolucional 10x5x5 - Camada densa de 30 unidades - 200 épocas

Analisando as Figuras 4.3, 4.4, 4.5 e 4.2 observa-se que o resultado tende a convergir para a média resultante na arquitetura orignal, sendo possível notar que o aumento na quantidade de filtros está diretamente proporcional à melhora da média de classificação, o que indica que com mais informações acerca das imagens de entrada, maior é a confiabilidade do sistema.

A quantidade de unidades de ativação na camada densa não parece ter causado melhoras na classificação, uma vez que a arquitetura descrita na Figura 4.5 possui metade de unidades de ativação comparados com a arquitetura original sendo que a média de classificação não apresentou melhora expressiva.

Estes resultados estão resumidos na Tabela 4.2 em que observa-se o melhor desempenho de classificação para a arquitetura proposta e discutida na metodologia deste trabalho. Este foi o indício que levou ao desenvolvimento e utilização de outros métodos de avaliação sistemática utilizando a mesma arquitetura proposta, como um *holdout* com mais reamostragem e a utilização do *k-fold*. Esta arquitetura ainda foi utilizada para a transferência de conhecimento para outros classificadores, em que estes classificadores são alimentados com dados de camadas intermediárias com o intuito de comparar o resultado com a classificação da CNN e avaliar o desempenho de cada classificador.

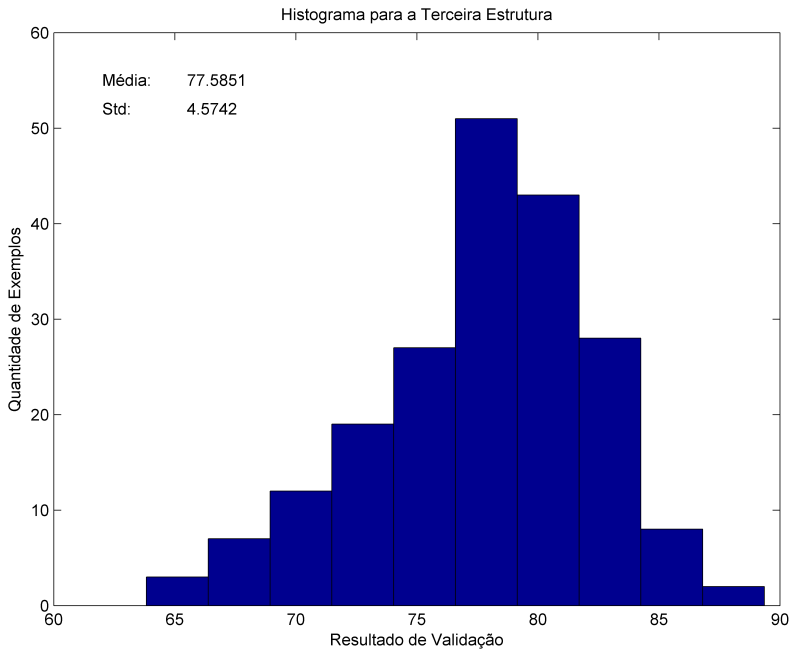

Figura 4.5. Histograma resultante das acurácias para a quarta arquitetura desenvolvida utilizando um *holdout* de 200 iterações com reamostragem aleatória. Esta arquitetura resultou uma média de 77.59% de acurácia com um desvio padrão de 4.57. Características da arquitetura: Primeiro Filtro Convolucional = 20x7x7 - Segundo Filtro Convolucional 15x7x7 - Camada densa de 50 unidades - 250 épocas

Tabela 4.2. Tabela das médias das acurácias resultantes das diferentes arquiteturas

	Média (%)	**Desvio Padrão**
Arquitetura Proposta	77.85	4.60
Arquitetura 2	75.30	4.74
Arquitetura 3	77.17	4.77
Arquitetura 4	77.59	4.57

Como comentado anteriormente, a fim de descrever melhor a resposta do sistema e aumentar a cofiabilidade do projeto, foram utilizados outros métodos de avaliação e análise dos resultados como os descritos em 3.2. Deste modo, foi levantado um novo resultado da arquitetura proposta utilizando um *holdout* com reamostragem aleatória de 530 iterações, além do método de *k-fold* com um K=20.

O resultado do novo *holdout* está descrito na Figura 4.6, em que é possível observar que houve um pequeno aumento em relação ao resultado anterior, porém como é possível analisar em Figura 4.6, depois de 300 iterações a média começa a convergir para o resultado, ou seja, com uma grande quantidade de repetições é possível observar a estabilidade dos resultados,

levando a uma confiabilidade. Este novo *holdout* resultou uma acurácia de 77.97% com um desvio padrão de 4.53%, o que superou o *holdout* anterior em 0.12%.

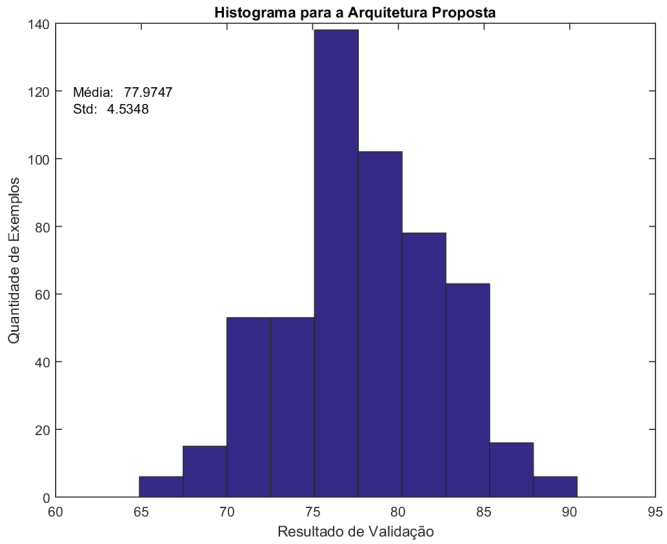

Figura 4.6. Histograma resultante das acurácias para arquitetura final utilizando um *holdout* de 530 iterações com reamostragem aleatória. Esta arquitetura resultou uma média de 77.97% de acurácia com um desvio padrão de 4.53

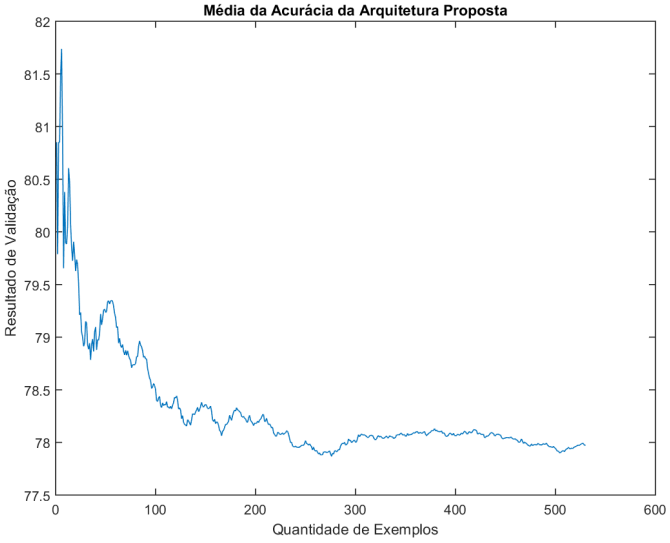

Figura 4.7. Média das acurácias para arquitetura final utilizando um *holdout* de 530 iterações com reamostragem aleatória. É possível observar que a partir de 300 iterações a média começa a estabilizar e convergir para o resultado final, diminuindo o erro aleatório e levando a crer que o resultado obtido se trata do resultado real.

Uma pergunta de pesquisa que foi levada durante a elaboração do projeto foi se a transferência de conhecimento para outros classificadores aumentaria as métricas de análise, e ainda se outros métodos de análise como o *k-fold* melhorariam o resultado. Como o método de validação-cruzada *k-fold* tem um trabalho computacional inferior comparado ao *holdout* com várias iterações, o seu uso é recomendável pelo fato de que, como citado em [32], é possível obter um resultado equivalente ao padrão ouro (500 iterações do *holdout*) utilizando um $K = 20$.

As Figuras 4.8, 4.9, 4.10, 4.11 representam respectivamente a acurácia média, especificidade média, precisão média e sensibilidade média ao longo dos 20 grupos de *k-fold* para o CNN e utilizando transferência de conhecimento para os classificadores SVM, Adaboost, Bagging e Gradient Boost.

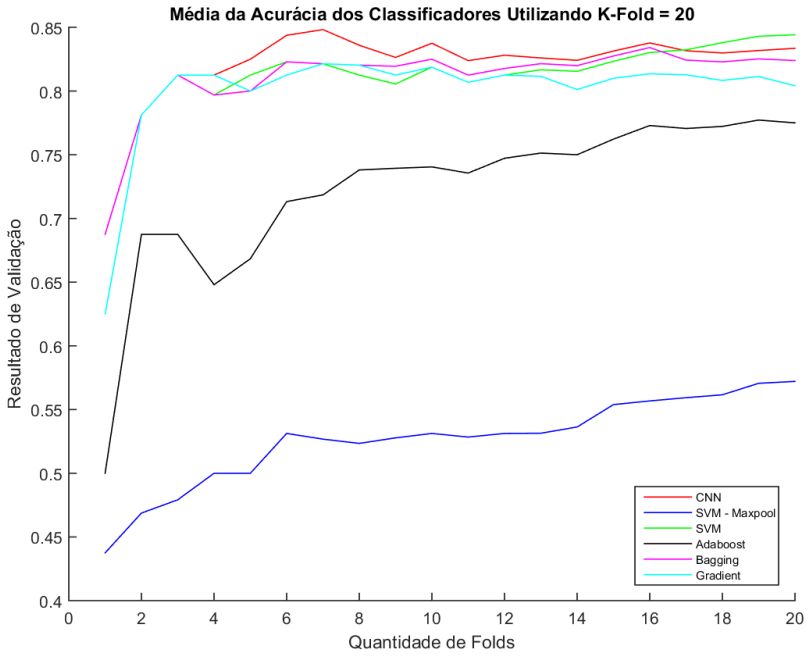

Figura 4.8. Resultado da média das acurácias ao longo dos 20 grupos. É possível observar que o resultado médio da acurácia do classificador SVM utilizando a primeira camada de Maxpool como alimentação do modelo não apresentou um resultado compatível com os que utilizaram a camada densa. Isso se acontece pois na saída da primeira camada de Maxpool, as características de aprendizado ainda não foram identificadas, prejudicando assim o desempenho da classificação. É possível observar também que a acurácia média do classificador Adaboost foi em média 4% menos do que os outros. Esse fato aconteceu pois neste classificador utilizamos um número de estimadores muito alto ($n = 100$), além de uma grande profundidade ($depth = 5$) o que possivelmente gerou overfit no treinamento, prejudicando o resultado da classificação.

Tabela 4.3. Tabela de Resultados para a acurácia média da CNN e de outros classificadores tipo *ensemble* utilizando um método de validação de *k-fold* com k=20.

	Acurácia média (%)
CNN	83.35
SVM - Maxpool	57.21
SVM	84.42
Adaboost	77.50
Bagging	82.40
Gradient Boost	80.42

Os resultados para a média da acurácia estão descritos na Figura 4.8 e sumarizados na Tabela 4.3. Para efeitos de comparação, foram utilizadas duas camadas diferentes de saída da arquitetura (Maxpool e Densa) como entrada da SVM. Podemos perceber pelas imagens e tabelas de desempenho que as métricas de validação utilizando a primeira camada de Maxpool como entrada para o classificador não apresenta um resultado consistente, com uma acurácia de 57.21%. Comparado com a mesma SVM utilizando a camada densa como entrada de alimentação do classificador obteve-se uma acurácia de 84.42%. Esta diferença se dá pois na primeira de Maxpool, a arquitetura ainda não conseguiu encontrar padrões nos dados de entrada, porém com o uso da camada densa estes padrões e abstrações já estão presentes e então obtém-se um melhor resultado.

É possível observar também que a acurácia média do classificador Adaboost foi em média 4p.p. menor do que os outros. Esse fato ocorreu pois neste classificador utilizamos um número de estimadores muito alto ($n = 100$), além de uma grande profundidade ($depth = 5$) o que possivelmente gerou overfit no treinamento, prejudicando o resultado da classificação.

Os resultados para a média da especifidade estão descritos na Figura 4.9 e sumarizados na Tabela 4.4. Assim como descrito em 3.3.3, a sensibilidade é a capacidade do algoritmo detectar quando não existe a presença da doença, portanto podemos perceber pelos resultados que a predição utilizando primeira camada de Maxpool está inclinada aos resultados positivos, tendo só 29.64% de média ao decorrer dos 20 folds de treinamento.

Tabela 4.4. Tabela de Resultados para a especifidade média da CNN e de outros classificadores tipo *ensemble* utilizando um método de validação de *k-fold* com k=20.

	Especifidade média (%)
CNN	84.91
SVM - Maxpool	29.64
SVM	85.18
Adaboost	79.38
Bagging Boost	82.95
Gradient Boost	80.80

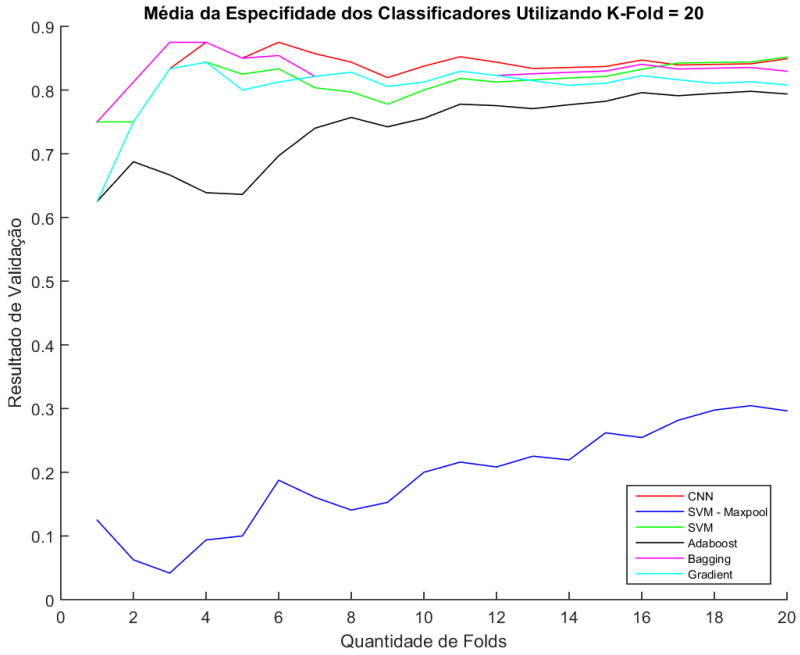

Figura 4.9. Resultado da média das especifidades ao longo dos 20 grupos. É possível observar que o resultado médio da especifidade do classificador SVM utilizando a primeira camada de Maxpool como alimentação do modelo não apresentou um resultado compatível com os que utilizaram a camada densa. Isso se acontece pois na saída da primeira camada de Maxpool, as características de aprendizado ainda não foram identificadas, prejudicando assim o desempenho da classificação.

Os resultados para a média da precisão estão descritos na Figura 4.10 e sumarizados na Tabela 4.5. Assim como descrito em 3.3.3, a precisão é a capacidade do algoritmo de detectar corretamente a presença da doença comparado com todos resultados que apontaram como positivo. Esta métrica é importante de ser avaliada pois aponta se o modelo de aprendizado de possui bias, ou seja, se o modelo apontasse todos resultados como sendo positivo (possuindo a doença), a precisão seria de em média 50%, mas como é possível observar pelos resultados, a precisão dos classificadores utilizando a camada densa possui a média maior que as outras métricas.

Os resultados para a média da sensibilidade estão descritos na Figura 4.11 e sumarizados na Tabela 4.6. Assim como descrito em 3.3.3, a sensibilidade é a proporção dos valores que o algoritmo detectou como positivo, e que realmente eram, com todos valores que possuíam o rotulo de positivo. Podemos perceber que o classificador SVM utilizando a primeira camada de Maxpool atingiu um valor similar aos outros classificadores, mas este resultado não pode ser avaliado separadamente das outras métricas de desempenho. No começo da abstração da arquitetura, o algoritmo de classificação interpreta a maioria dos elementos de entrada

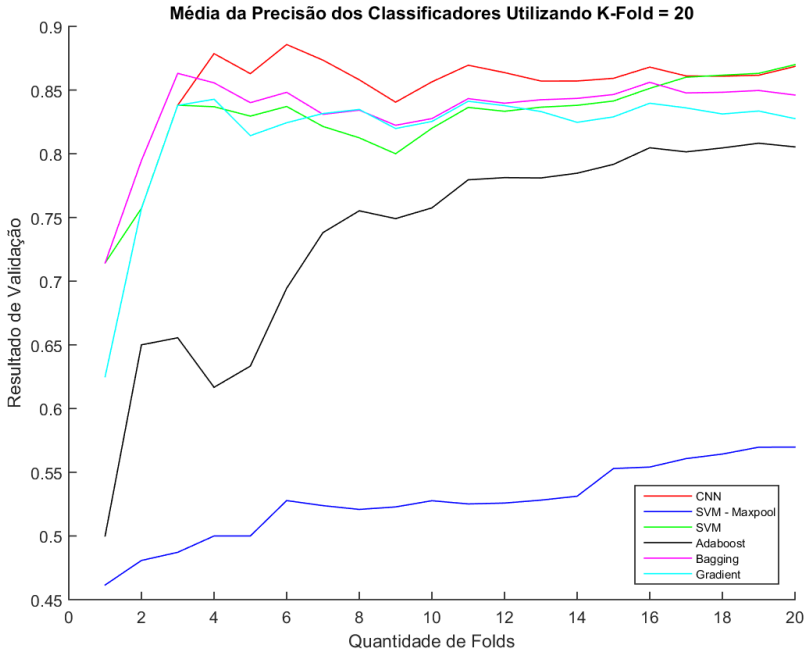

Figura 4.10. Resultado da média das precisões ao longo dos 20 grupos. É possível observar que o resultado médio da especifidade do classificador SVM utilizando a primeira camada de Maxpool como alimentação do modelo não apresentou um resultado compatível com os que utilizaram a camada densa. Isso se acontece pois na saída da primeira camada de Maxpool, as características de aprendizado ainda não foram identificadas, prejudicando assim o desempenho da classificação.

Tabela 4.5. Tabela de Resultados para a precisão média da CNN e de outros classificadores tipo *ensemble* utilizando um método de validação de *k-fold* com k=20.

	Precisão média (%)
CNN	86.85
SVM - Maxpool	56.58
SVM	86.99
Adaboost	80.54
Bagging Boost	84.61
Gradient Boost	82.76

como sendo positivo, ou seja, a taxa de acurácia é baixa, porém a sensibilidade é alta.

Estes valores podem ser avaliados e melhor compreendidos analisando a matriz de confusão, que será apresentada a seguir.

Podemos notar que o método de k-fold melhorou a acurácia da CNN com uma diferença em pontos percentuais de 5.35% comparada ao método de *holdout*. Isso indica que com mais dados para treinamento, o classificador consegue identificar com maior exatidão as características invariantes que definem os indivíduos com esquizofrenia, assim como os de

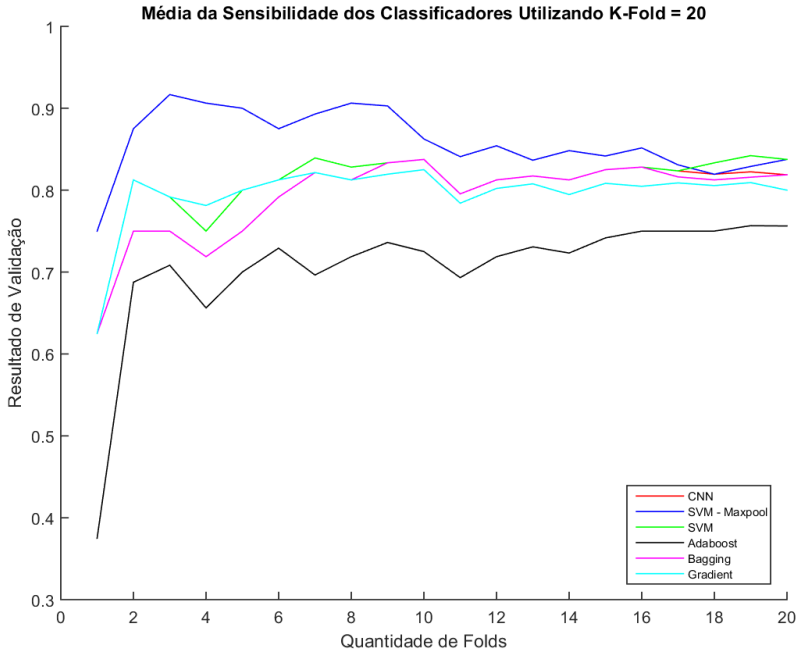

Figura 4.11. Resultado da média das sensibilidades ao longo dos 20 grupos. É possível observar que o resultado médio da especifidade do classificador SVM utilizando a primeira camada de Maxpool como alimentação do modelo apresentou um resultado similar aos outros classificadores, pois no começo da abstração da arquitetura, o algoritmo de classificação interpreta a maioria dos elementos de entrada como sendo positivo, ou seja, a taxa de acerto da acurácia é baixa, porém a sensibilidade é alta.

Tabela 4.6. Tabela de Resultados para a sensibilidade média da CNN e de outros classificadores tipo *ensemble* utilizando um método de validação de *k-fold* com k=20.

	Sensibilidade média (%)
CNN	81.87
SVM - Maxpool	83.75
SVM	83.75
Adaboost	75.62
Bagging Classifier	81.87
Gradient Boost	80.00

controle.

Uma pergunta de pesquisa que foi levantada no começo deste trabalho era se a transferência de conhecimento melhoraria as métricas de desempenho. Podemos observar pelos resultados descritos que a utilização da SVM com a camada densa resulta em um melhor desempenho para todas métricas de validação, com uma diferença em pontos percentuais de 1.07% na acurácia média, porém os outros classificadores não apresentam a mesma melhora. Um fator positivo acerca dos classificadores tipo *ensemble* e da SVM é o tempo necessário

para realizar o treinamento e validação, que foi em média 50% mais rápido que a CNN.

Resultado Matriz de Confusão

Como descrito anteriormente, a camada que possivelmente levaria para um melhor resultado na transferência de conhecimento seria a camada densa. Porém foi utilizada também a saída da primeira camada de Maxpool para fins de comparação. É possível observar que os resultados utilizando a camada de Maxpool foram piores comparados aos que utilizaram a camada densa como entrada para classificação.

Um outro modo de analisar os resultados obtidos é observar a matriz de confusão de cada classificador. Estas matrizes estão descritas nas Tabelas 4.7 4.8 4.9 4.10 4.11 e 4.12

Tabela 4.7. Tabela da matriz de confusão para o CNN

		Previsão	
		Negativo	**Positivo**
Rótulo	**Negativo**	129	23
	Positivo	29	131

Tabela 4.8. Tabela da matriz de confusão para a SVM utilizando a camada Maxpool

		Previsão	
		Negativo	**Positivo**
Rótulo	**Negativo**	44	108
	Positivo	26	134

Tabela 4.9. Tabela da matriz de confusão para a SVM utilizando a camada densa

		Previsão	
		Negativo	**Positivo**
Rótulo	**Negativo**	129	23
	Positivo	26	134

Tabela 4.10. Tabela da matriz de confusão para o Adaboost utilizando a camada densa

		Previsão	
		Negativo	**Positivo**
Rótulo	**Negativo**	121	32
	Positivo	39	121

Uma analise interessante que podemos tirar dos dados da matriz de confusão é que mesmo em casos que os classificadores não acertam o valor correto da predição, é favorável errar em casos que o indivíduo não possua a doença e o algoritmo identifica como tendo (FP), pois como se trata de um auxilio ao diagnóstico, exames adicionais possivelmente serão realizados

Tabela 4.11. Tabela da matriz de confusão para o Bagging utilizando a camada densa

		Previsão	
		Negativo	**Positivo**
Rótulo	**Negativo**	126	26
	Positivo	29	131

Tabela 4.12. Tabela da matriz de confusão para o Gradient Boost utilizando a camada densa

		Previsão	
		Negativo	**Positivo**
Rótulo	**Negativo**	123	29
	Positivo	32	128

para a identificação correta da doença. Portanto, uma outra métrica que é possível ser retirada das matrizes de confusão é uma taxa de "confiabilidade" C, em que pegam-se todos TP, TN e FP em relação ao número total n. Esta taxa para a SVM utilizando a camada densa é 91.66%.

Assim como as outras métricas de desempenho descritas anteriormente, não podemos analisar este resultado isoladamente, pois mesmo a SVM utilizando a camada Maxpool teria um resultado similar aos outros, porém sua acurácia, precisão e especifidade são extremamente baixas comparadas aos outros, portanto não representaria um resultado fidedigno.

Para se ter uma ideia de onde o presente trabalho se enquadra em relação à outros trabalhos com a mesma proposta de implementação, utilizando modelos de *deep learning* para realizar diagnósticos de esquizofrenia, apresentamos uma comparação da acurácia média entre estes trabalhos. Apesar dos bancos de dados serem diferentes, não sendo possível realizar uma comparação direta dos resultados, é ainda interessante compará-los para analisar onde o nosso trabalho se enquadra em relação à outros presentes na literatura.

Tabela 4.13. Tabela de comparação entre os trabalhos de esquizofrenia

	Acurácia média (%)
Nosso Trabalho	84.42
Pinaya - 2017[63]	70.00
Zeng - 2018[68]	81.00

4.1.1 Visualização de Camadas - Controle

Está seção apresenta a visualização das camadas de aprendizado da arquitetura, começando com uma imagem aleatória do grupo de controle para exemplificar o caminho seguido pela rede neural. São apresentados os pesos iniciais com distribuição normal utilizados no processo, seguido das imagens filtrados nas camadas de convolução, subamostragem, e seguindo por camadas mais abstratas até chegar na camada de saída onde há a ativação de um dos rótulos classificação resultante do processo. Na saída das camadas de convolução, poderemos observar a presença dos mapas de características, que são os resultados das convoluções dos *n kernels* com a imagem de entradas.

Imagem de Entrada

A imagem de exemplo ilustrada na Figura 4.12 pertence à um indivíduo de controle e é utilizada como entrada na arquitetura. A figura ainda apresenta os 32 *kernels* iniciais com pesos distribuídos de forma normal para dar início ao processo.

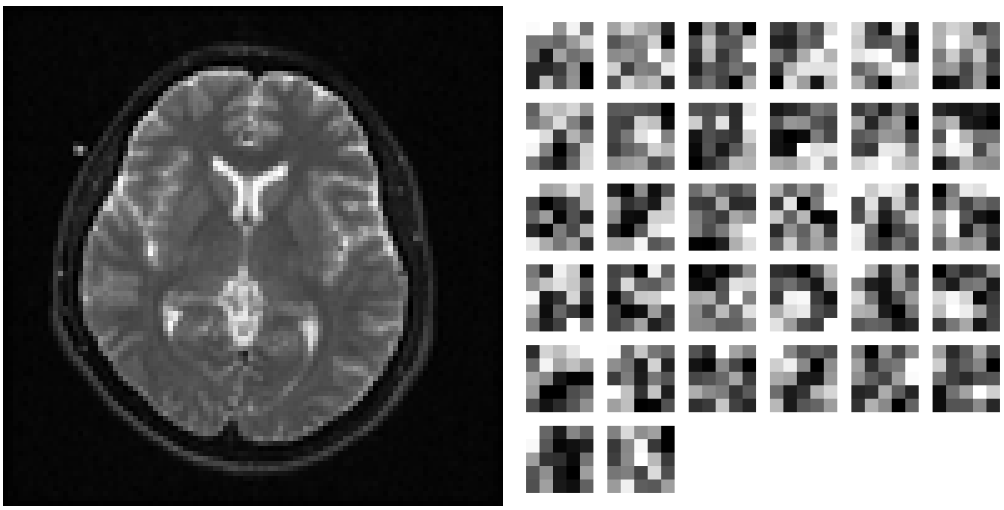

Figura 4.12. À esquerda: Exemplo de imagem de controle para processamento e classificação. À direita: Resultado da distribuição normal para inicialização dos 32 *kernels* de convolução

Os pesos apresentados na Figura 4.12 são resultantes de uma destribuição normal com ganho de 1. Este tipo peso para inicialização foi proposto por Glarot [23] e amplamente utilizado em algoritmos de aprendizado de máquina.

Mapa de características para a primeira camada convolucional

A Figura 4.13 apresenta o mapa de característica resultante da primeira camada convolucional para cada um dos 32 *kernels* de entrada da arquitetura aplicados à um exemplo de entrada. As informações presentes em cada mapa são informações de traços, bordas, retas e possíveis estruturas encontradas em cada área da imagem. Em alguns casos há a presença de relevo, enfatizando a região dos ventrículos do cerebelo.

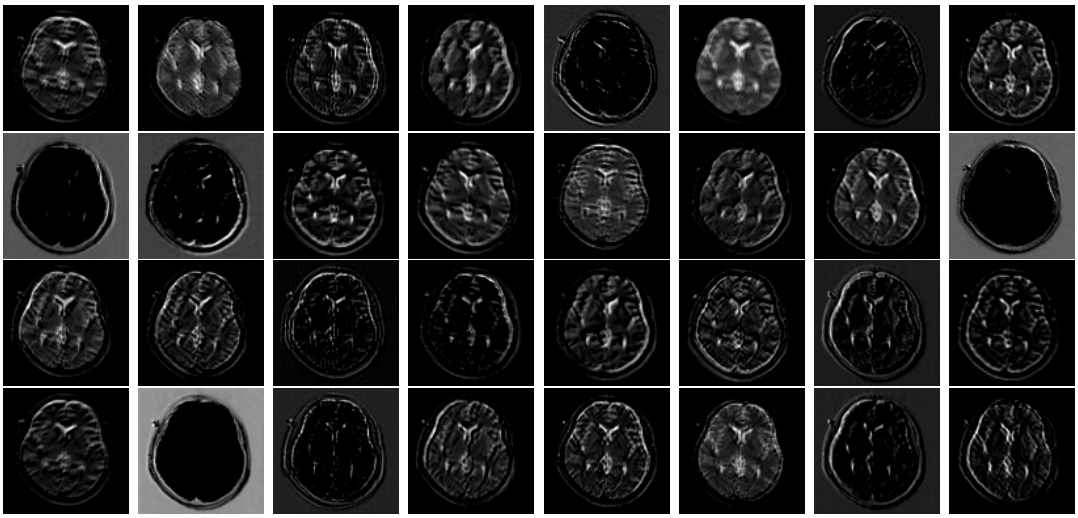

Figura 4.13. Conjunto de imagens resultante de indivíduos de controle para a primeira camada convolucional dos 32 *kernels* de entrada. As informações presentes em cada mapa são informações iniciais de traços, bordas, retas e possíveis estruturas encontradas em cada área da imagem. Em alguns casos há a presença de relevo, enfatizando a região dos ventrículos do cerebelo.

Mapa de características para a segunda camada convolucional

A Figura 4.14 apresenta o mapa de característica resultante da segunda camada convolucional para cada um dos 64 *kernels* da camada atual aplicados à um exemplo da camada anterior. As informações presentes em cada mapa já estão um pouco mais abstratas, com maior nível de características comparados com os mapas da primeira camada. Estes mapas são de um tamanho menor pelo fato de ter passado anteriormente por uma camada de *maxpool* diminuindo em dois a sua dimensionalidade.

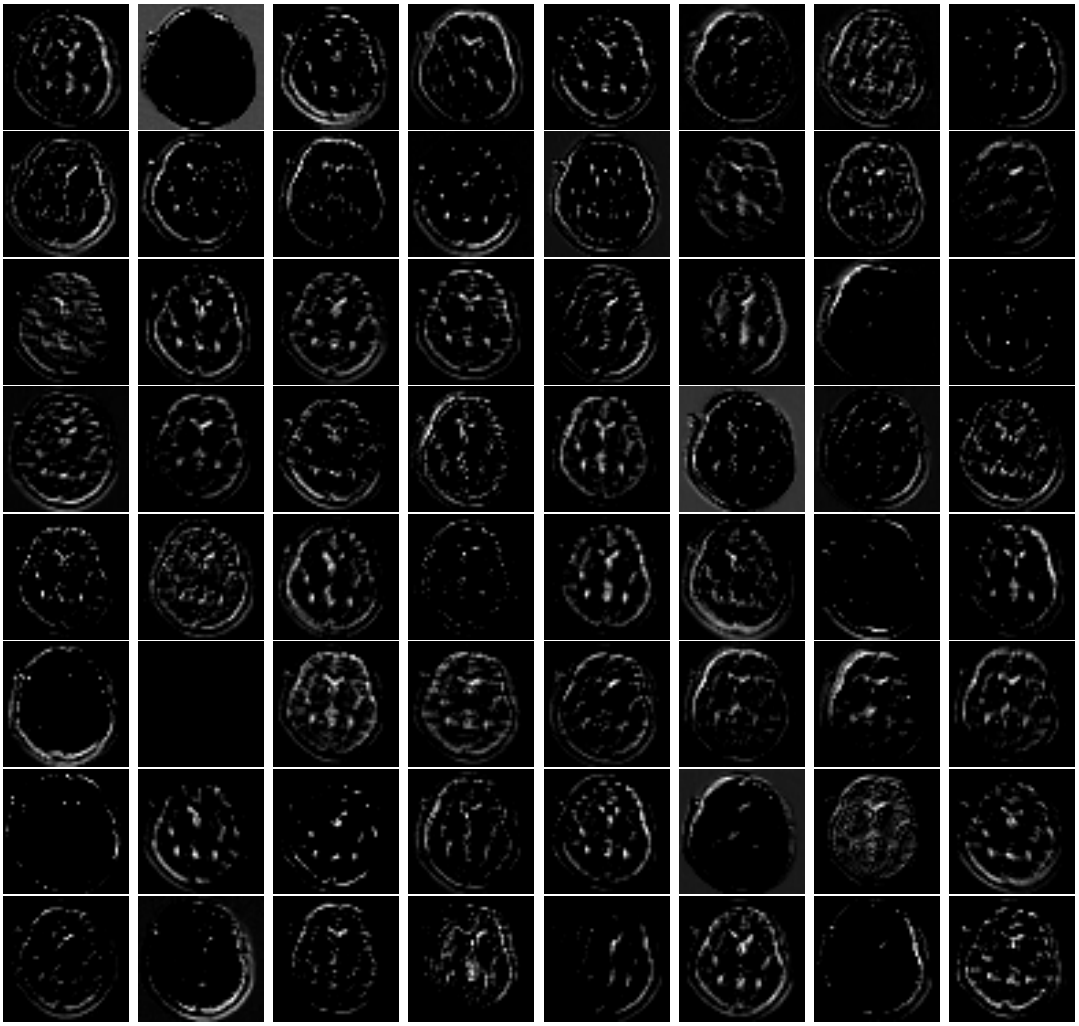

Figura 4.14. Conjunto de imagens resultante de indivíduos de controle para a segunda camada convolucional dos 64 *kernels* aplicados à um exemplo da camada anterior. Nesta camada, as informações estão um pouco mais abstratas, com maior nível de características comparados com os mapas da primeira camada. Estes mapas são de um tamanho menor pelo fato de ter passado anteriormente por uma camada de *maxpool* diminuindo em dois a sua dimensionalidade.

Elementos de ativação da camada densa

Como discutido anteriormente, a arquitetura desenvolvida pode ser utilizada com um extrator de características para utilização em outro classificadores. Além das características dos filtros das duas camadas convolucionais, outras características como as ativações na camada densa podem ser utilizadas como entrada em um classificador clássico. Neste contexto, a Figura 4.15 apresenta as 100 unidades de ativação da camada densa antes de alimentar a saída geral da arquitetura.

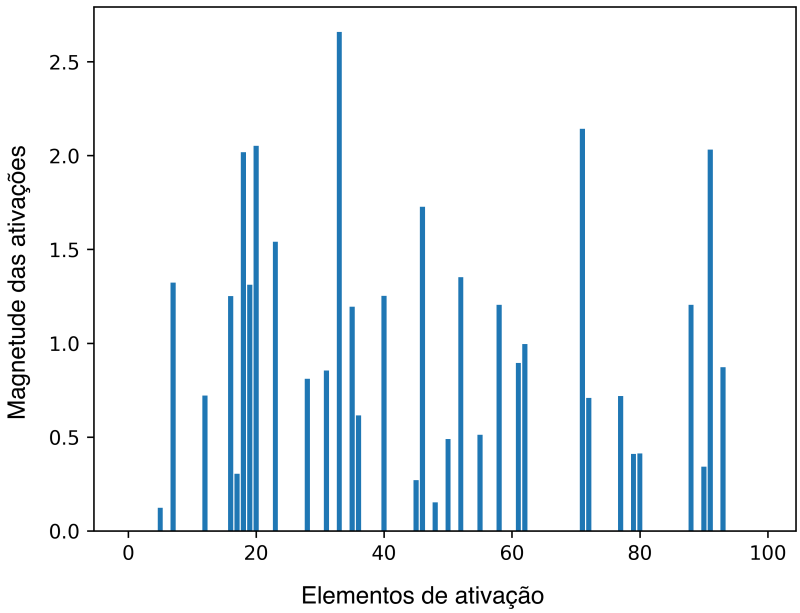

Figura 4.15. Elementos de ativação da camada densa de um exemplo de indivíduo de controle. Com estas ativações é possível realizar uma transferência de conhecimento e aplicá-las em outros classificadores.

Resultado do aprendizado

Finalmente a última camada representa a camada de saída, tendo com maior abstração o resultado de aprendizado com a classificação da imagem em um dos rótulos inicias. O 0 representa indivíduos de controle e o 1 representa indivíduos esquizofrênicos. A Figura 4.16 mostra que de fato o indivíduo de teste representa o grupo de controle pois a resultante de maior ativação acontece no 0.

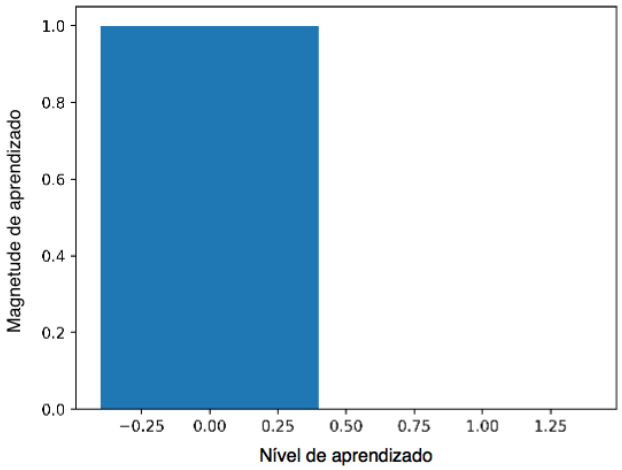

Figura 4.16. Resultado da saída de um indivíduo de controle, destacando o valor onde acontece a maior ativação, comprovando que o treinamento foi correto.

4.1.2 Visualização de Camadas - Esquizofrenia

Está seção apresenta uma visualização similar à apresentada anteriormente, mostrando as camadas de aprendizado da arquitetura, começando com uma imagem aleatória do grupo de esquizofrênicos para exemplificar o caminho seguido pela rede neural. São apresentados os pesos iniciais com distribuição normal utilizados no processo, seguido dos mapas de características resultantes das camadas de convolução, subamostragem, e seguindo por camadas mais abstratas até chegar na camada de saída onde há a ativação de um dos rótulos classificação resultante do processo.

Imagem de Entrada

A imagem de exemplo ilustrada na Figura 4.17 pertence à um indivíduo esquizofrênico e é utilizada como entrada na arquitetura. A figura ainda apresenta os 32 *kernels* iniciais com pesos distribuídos de forma normal para dar início ao processo.

É possível perceber que os filtros inicializados são os mesmos *kernels* representados na 4.12, isto acontece pelo fato dos filtros de entrada para ambos casos serem inicializados na mesma rede, onde as mudanças acontecem no decorrer da arquitetura, buscando as características invariantes em ambos os casos.

Mapa de características para a primeira camada convolucional

A Figura 4.18 apresenta o mapa de características resultante da primeira camada convolucional para cada um dos 32 *kernels* de entrada da arquitetura aplicados à um exemplo de

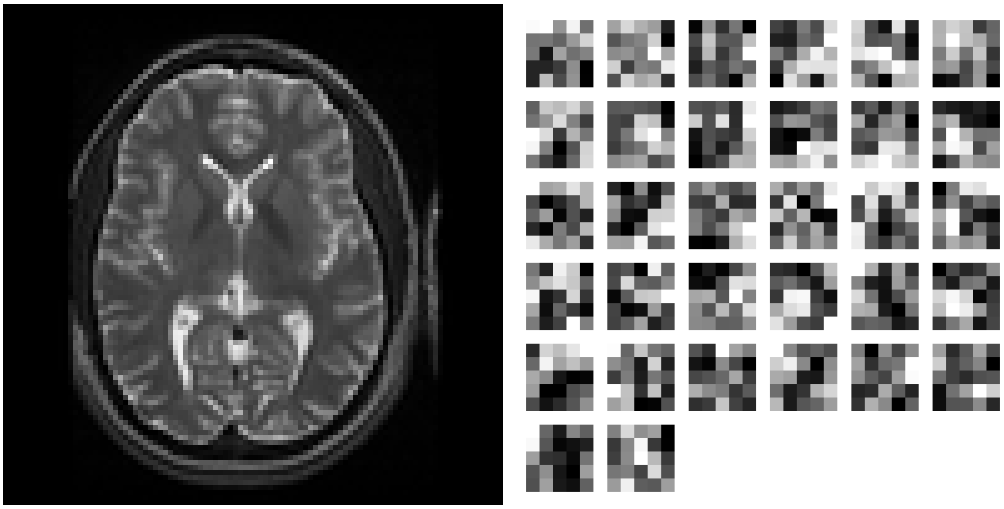

Figura 4.17. À esquerda: Exemplo de imagem de esquizofrenia para processamento e classificação. À direita: Resultado da distribuição normal para inicialização dos 32 *kernels* de convolução

entrada de uma paciente esquizofrênico. Nota-se que as informações presentes em cada mapa são parecidas com aquelas apresentadas em indivíduos de controle, que são informações de traços, bordas, retas e possíveis estruturas encontradas em cada área da imagem. Em alguns casos há a presença de relevo, enfatizando a região dos ventrículos do cerebelo.

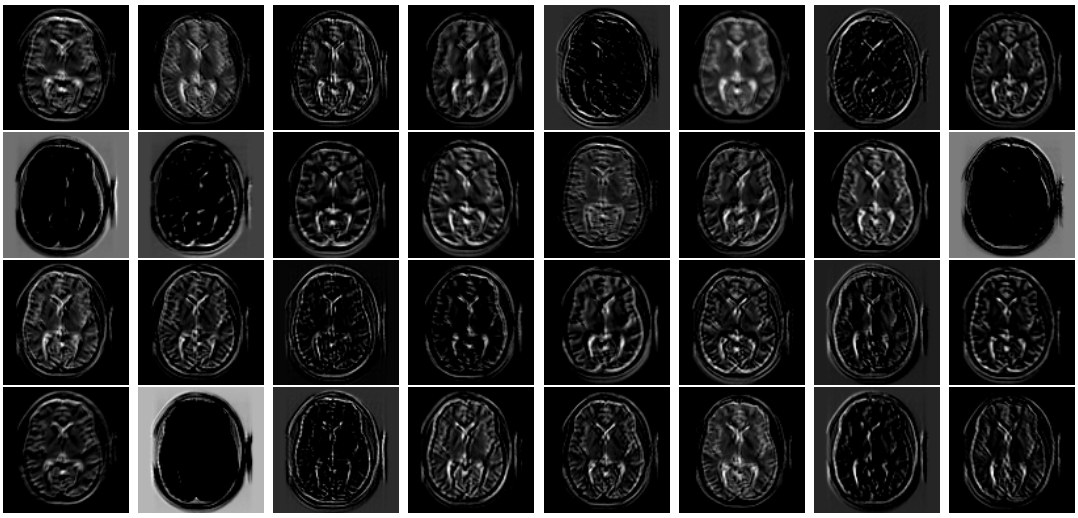

Figura 4.18. Conjunto de imagens resultante de indivíduos esquizofrênicos para a primeira camada convolucional dos 32 *kernels* de entrada. As informações presentes em cada mapa são informações iniciais de traços, bordas, retas e possíveis estruturas encontradas em cada área da imagem. Em alguns casos há a presença de relevo, enfatizando a região dos ventrículos do cerebelo.

Mapa de características para a segunda camada convolucional

A Figura 4.14 apresenta o mapa de característica resultante da segunda camada convolucional para cada um dos 64 *kernels* da camada atual aplicados à um exemplo da camada anterior. As informações presentes em cada mapa são novamente similares às informações de controle, tendo um nível de abstração um pouco maior em relação ao mapa anterior, com maior nível de características comparados com os mapas da primeira camada. Estes mapas são de um tamanho menor pelo fato de ter passado anteriormente por uma camada de *maxpool* diminuindo em dois a sua dimensionalidade.

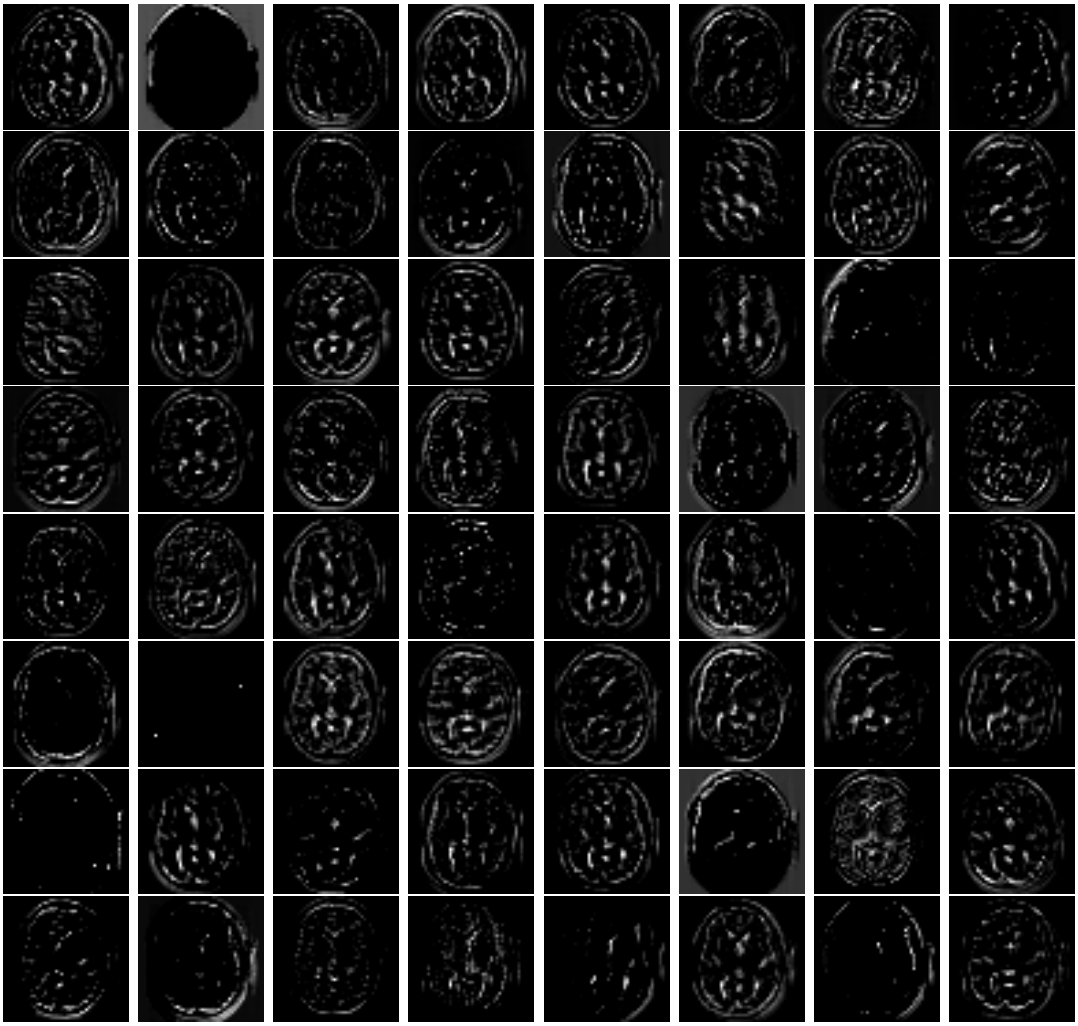

Figura 4.19. Conjunto de imagens resultante de indivíduos esquizofrênicos para a segunda camada convolucional dos 64 *kernels* aplicados à um exemplo da camada anterior. Nesta camada, as informações estão um pouco mais abstratas comparados com os mapas da primeira camada.

Elementos de ativação da camada densa

Do mesmo modo que aconteceu no indivíduos de controle, as características dos esquizofrênicos servem também para alimentar outros tipos de classificadores. Deste modo a Figura 4.20 apresenta as 100 unidades de ativação da camada densa antes de alimentar a saída geral da arquitetura.

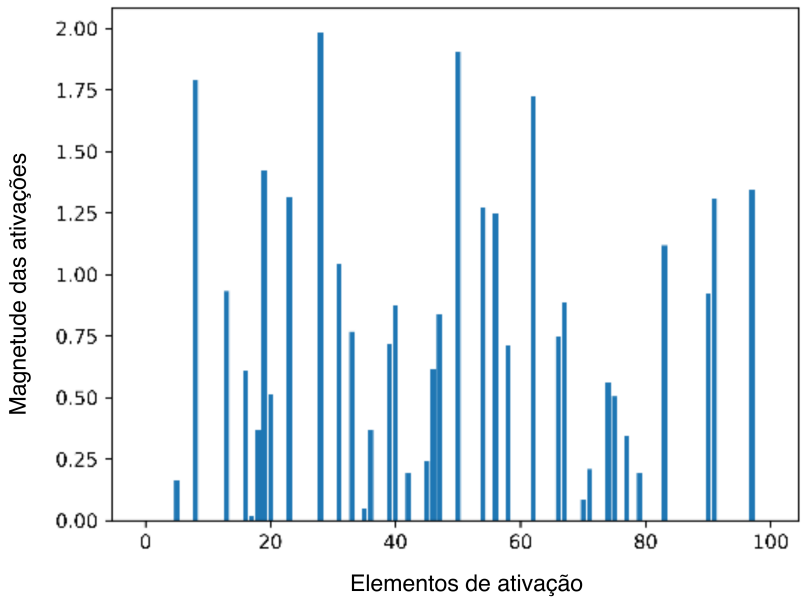

Figura 4.20. Elementos de ativação da camada densa de um exemplo de indivíduo esquizofrênico. Com estas ativações é possível realizar uma transferência de conhecimento e aplicá-las em outros classificadores

Resultado do aprendizado

Finalmente a última camada representa a camada de saída, tendo com maior abstração o resultado de aprendizado com a classificação da imagem em um dos rótulos inicias. O 0 representa indivíduos de controle e o 1 representa indivíduos esquizofrênicos. A Figura 4.21 mostra que de fato o indivíduo de teste representa o grupo de controle pois a resultante de maior ativação acontece no 1.

Estes resultados mostram com maior clareza o processo em que a arquitetura passa, saindo de uma imagem de entrada até chegar ao nível mais elevado de abstração e a saída da função de perda para cada exemplo.

Podemos observar que a função de perda possui um valor muito baixo de erro, em que ambos casos (controle e esquizofrênico) tiveram uma magnitude do resultado de aprendizado

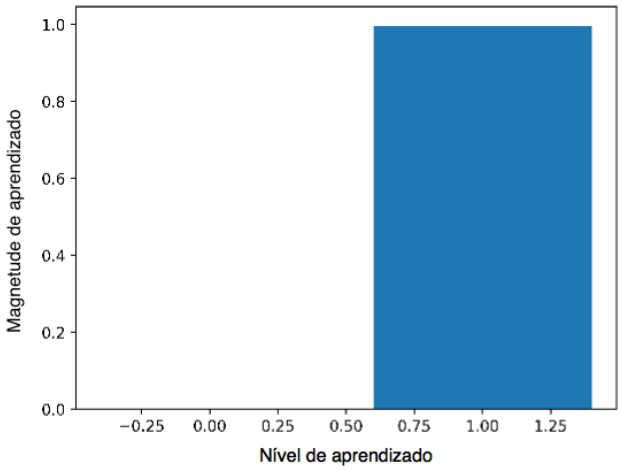

Figura 4.21. Resultado Saída Esquizofrênico. Nota-se que a ativação maior está presente no 1, representando a saída correta do classificador

em quase 100%. Estes exemplos tiveram um resultado positivo, porém em casos em que o algoritmo não detecta corretamente a classe correta, o valor da função de perda pode apresentar valores errôneos, variando com uma precisão reduzida.

Visualização das áreas mais relevantes para classificação

Uma outra visualização interessante de se fazer como referenciado em [67]. A Figura 4.22 representa um exemplo das partes que a arquitetura utilizou como mais importante e tendo mais peso para a classificação. Este tipo de visualização é importante para analisar se a rede de aprendizado buscou características importantes para a classificação e compará-las com as referenciados na literatura. Observa-se que na sobreposição das imagens, esta visualização apresenta as regiões onde a rede reconheceu como sendo as mais importantes para a classificação, e quando colocadas sobre à imagem original, as características destacadas apresentam as regiões essenciais para o treinamento.

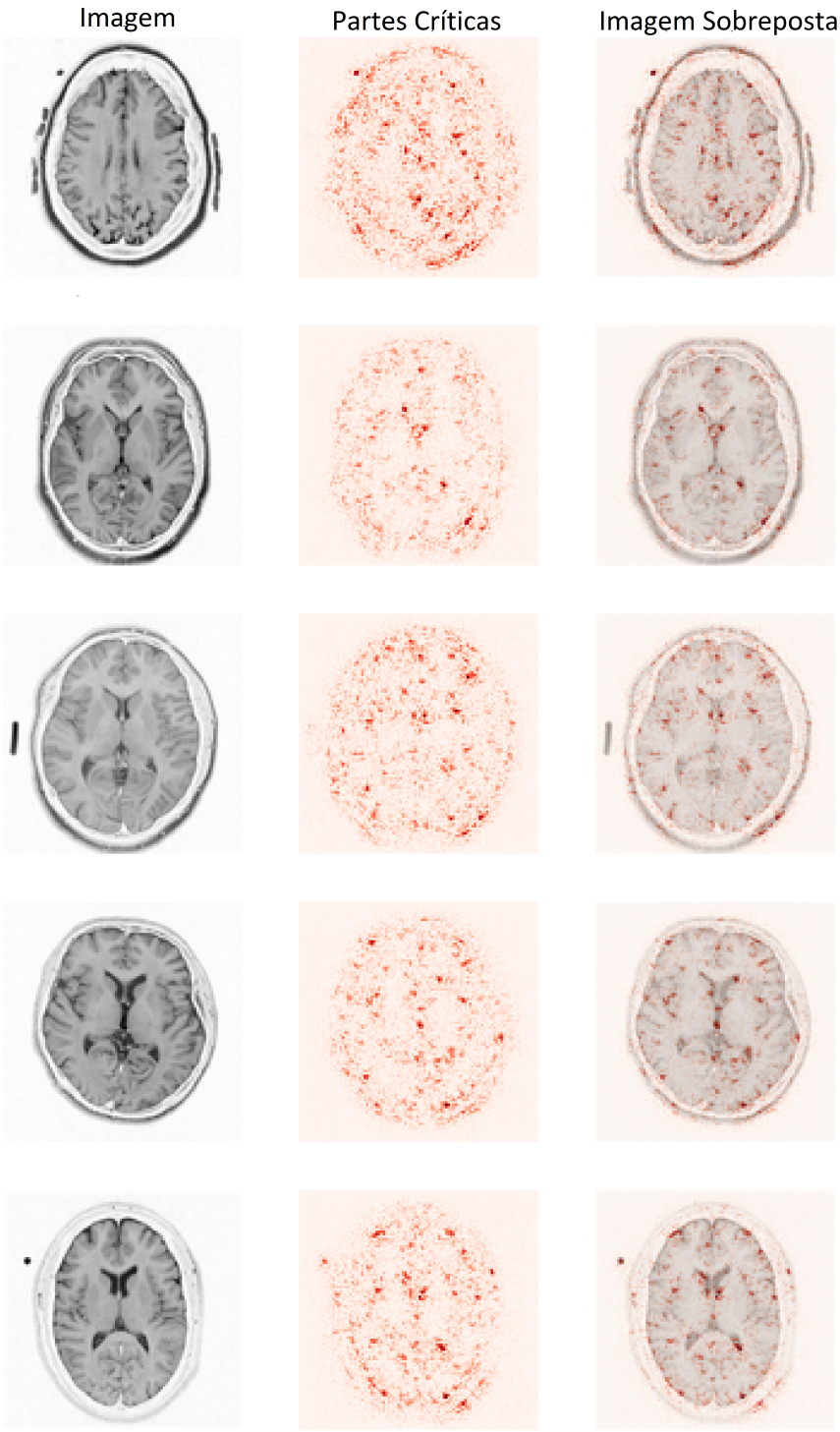

Figura 4.22. Regiões mais relevantes para o treinamento da CNN. Esta visualização apresenta as regiões onde a rede reconheceu como sendo as mais importantes para a classificação, e quando colocadas sobre à imagem original, as características destacadas apresentam as regiões essenciais para o treinamento.

5 Conclusão e Trabalhos Futuros

Desde o início deste trabalho, o objetivo que guiava a pesquisa era a busca por uma metodologia e ferramentas capazes de realizar um auxílio ao diagnóstico da esquizofrenia utilizando características do encéfalo humano em imagens de ressonância magnética. Neste sentido foi desenvolvido um sistema a fim de extrair as características invariantes de indivíduos com esquizofrenia e de indivíduos de controle para realizar um treinamento capaz de distinguir estas duas classes.

Para tal, algumas perguntas de pesquisa foram levantadas com o intuito de avaliar o treinamento e sua capacidade de avaliação e diagnóstico. Perguntas como com qual precisão, acurácia, sensibilidade e especifidade a arquitetura desenvolvida de CNN conseguiria identificar corretamente as classes designadas. E se a transferência de conhecimento para outros classificadores melhoraria estas métricas de desempenho. Observamos que a transferência para a SVM melhorou o desempenho da acurácia média em 1.07 pontos percentuais.

A arquitetura desenvolvida consiste em 4 camadas de aprendizado, sendo 2 camadas convolucionais e duas camadas de conexão total. A primeira camada de convolução possui 32 filtros 5x5 e um *maxpool* de 2x2, a segunda camada possui 64 imagens filtradas por filtros 3x3 e outro *maxpool* de 2x2, além de *dropouts* de 0.5 antes e depois da camada densa, sendo esta camada com 100 unidades de ativação. Utilizando esta arquitetura e o banco de dados fBIRN, gerou-se um histograma com 530 iterações do método de *holdout* com reamostragem aleatória, obtendo uma acurácia média de 77.97%($\pm$4.53), e 83.35% para o método de *k-fold* com k=20. Esta melhora indica que com mais imagens para treinamento, o classificador consegue identificar com maior exatidão as características invariantes que definem os indivíduos de controle, assim como os de controle.

Apesar da arquitetura apresentada não ser tão profunda como em diversos casos descritos na literatura, o desempenho alcançado foi acima do resultados que utilizam uma abordagem de CNN para disgnóstico de esquizofrenia, apontando que há ainda espaço para esse ramo ser aprofundado com redes mais sofisticadas. Utilizando esta proposta, o diagnóstico da esquizofrenia pode ser realizada com resultados mais concretos, ao invés do diagnóstico subjetivo apresentado atualmente, desde modo gerando benefícios à pacientes que estão sofrendo com este transtorno.

Esta pesquisa ainda aponta que existe uma grande oportunidade de crescimento na área de classificação de imagens médicas utilizando técnicas de deep learnig para auxílio ao diagnostico de esquizofrenia, e que há a possibilidade de que mais experimentos possam ser realizados com o intuito de melhorar ainda mais as métricas de desempenho. Alguns possíveis

trabalhos futuros são descritos na Seção 5.1.

5.1 Trabalhos Futuros

1. Utilização da arquitetura desenvolvida para testes em outros bancos de dados;
2. Utilização do sistema proposto em casos reais para detectar a presença da psicopatia e auxiliar médicos na caracterização do diagnóstico;
3. Refinamento e utilização da arquitetura para diagnósticos e predição de outros tipos de doenças;

Referências Bibliográficas

[1] Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., Deng, L., Penn, G., and Yu, D. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing 22*, 10 (2014), 1533–1545.

[2] Abdel-Hamid, O., Mohamed, A.-r., Jiang, H., and Penn, G. Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on* (2012), IEEE, pp. 4277–4280.

[3] Amin, M. F., and Plis, e. a. Reading the (functional) writing on the (structural) wall: Multimodal fusion of brain structure and function via a deep neural network based translation approach reveals novel impairments in schizophrenia. *NeuroImage* (2018).

[4] Andreasen, N. C., Flashman, L., Flaum, M., Arndt, S., Swayze, V., O'leary, D. S., Ehrhardt, J. C., and Yuh, W. T. Regional brain abnormalities in schizophrenia measured with magnetic resonance imaging. *Jama 272*, 22 (1994), 1763–1769.

[5] Association, A. P., et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[6] Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. Theano: A cpu and gpu math compiler in python. In *Proc. 9th Python in Science Conf* (2010), pp. 1–7.

[7] Bishop, C. M. *Pattern recognition and machine learning*. springer, 2006.

[8] Bloch, F. Nuclear induction. *Physical Review 70*, 1 (1946), 460–473.

[9] Braak, H., and Braak, E. Neuropathological stageing of alzheimer-related changes. *Acta neuropathologica 82*, 4 (1991), 239–259.

[10] Brink, H., Richards, J. W., and Fetherolf, M. *Real-World Machine Learning*. Manning Publications, 2016.

[11] Bryan, R. N. *Introduction to the Science of Medical Imaging*, 1st ed. Cambridge University Press, New York, USA, 2010.

[12] Burges, C. J. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery 2*, 2 (1998), 121–167.

[13] Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915* (2016).

[14] Ciresan, D. C., Meier, U., Gambardella, L. M., and Schmidhuber, J. Convolutional neural network committees for handwritten character classification. In *Document Analysis and Recognition (ICDAR), 2011 International Conference on* (2011), IEEE, pp. 1135–1139.

[15] Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[16] Cruz, B. F. d. Classificação de esquizofrenia com base em máquinas de suporte vetorial aplicadas a características de imagens de ressonância magnética. *University of Brasilia* (2016).

[17] DE SAUDE, O. *Cid-10-Vol.3 - Classificaçação Estatística Internacional de Doenças Vol. 3*. No. v. 2 in CID-10: classificação estatística internacional de doenças e problemas relacionados à saúde. Edusp, 1998.

[18] Dieleman, S., Schlüter, J., Raffel, C., Olson, E., Sønderby, S. K., Nouri, D., et al. Lasagne: First release., Aug. 2015.

[19] D'Onofrio, B. M., Rickert, M., and Frans, E. Paternal age at childbearing and offspring psychiatric and academic morbidity. *JAMA Psychiatry 71*, 4 (2014), 432–438.

[20] Engelke, K., Adams, J. E., Armbrecht, G., Augat, P., Bogado, C. E., Bouxsein, M. L., Felsenberg, D., Ito, M., Prevrhal, S., Hans, D. B., et al. Clinical use of quantitative computed tomography and peripheral quantitative computed tomography in the management of osteoporosis in adults: the 2007 iscd official positions. *Journal of Clinical Densitometry 11*, 1 (2008), 123–162.

[21] Freedman, R. Schizophrenia. *New England Journal of Medicine 349*, 18 (2003), 1738–1749. PMID: 14585943.

[22] Friedman, L., and Glover, G. H. Report on a multicenter fmri quality assurance protocol. *Journal of Magnetic Resonance Imaging 23*, 6 (2006), 827–839.

[23] Glorot, X., and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (2010), pp. 249–256.

[24] Goehde, S. C., Hunold, P., Vogt, F. M., Ajaj, W., Goyen, M., Herborn, C. U., Forsting, M., Debatin, J. F., and Ruehm, S. G. Full-body cardiovascular and tumor mri for early detection of disease: feasibility and initial experience in 298 subjects. *American Journal of Roentgenology 184*, 2 (2005), 598–611.

[25] Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[26] Greicius, M. D., Flores, B. H., Menon, V., Glover, G. H., Solvason, H. B., Kenna, H., Reiss, A. L., and Schatzberg, A. F. Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biological psychiatry 62*, 5 (2007), 429–437.

[27] Hahnloser, R. H., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature 405*, 6789 (2000), 947.

[28] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., and Larochelle, H. Brain tumor segmentation with deep neural networks. *Medical image analysis 35* (2017), 18–31.

[29] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine 29*, 6 (2012), 82–97.

[30] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

[31] Hodges, A., Strand, A. D., Aragaki, A. K., Kuhn, A., Sengstag, T., Hughes, G., Elliston, L. A., Hartog, C., Goldstein, D. R., Thu, D., et al. Regional and cellular gene expression changes in human huntington's disease brain. *Human molecular genetics 15*, 6 (2006), 965–977.

[32] Kohavi, R., et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (1995), vol. 14, Montreal, Canada, pp. 1137–1145.

[33] Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (2012), pp. 1097–1105.

[34] Lauterbur, P. Image formation by induced local interactions: examples employing nuclear magnetic resonance. *Nature* (1973).

[35] LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature 521*, 7553 (2015), 436–444.

[36] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural computation 1*, 4 (1989), 541–551.

[37] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (1998), 2278–2324.

[38] LeCun, Y., Kavukcuoglu, K., and Farabet, C. Convolutional networks and applications in vision. In *Circuits and Systems (ISCAS), Proceedings of 2010 IEEE International Symposium on* (2010), IEEE, pp. 253–256.

[39] Lenton, D. Eye-tracking test for schizophrenia wins business award. *Engineering and Technology* (2013).

[40] Long, J., Shelhamer, E., and Darrell, T. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3431–3440.

[41] Mishkin, D., Sergievskiy, N., and Matas, J. Systematic evaluation of convolution neural network advances on the imagenet. *Computer Vision and Image Understanding 161* (2017), 11–19.

[42] Mota, N. B., Copelli, M., and Ribeiro, S. Thought disorder measured as random speech structure classifies negative symptoms and schizophrenia diagnosis 6 months in advance. *npj Schizophrenia 3*, 1 (2017), 18.

[43] Mueser, K. T., and McGurk, S. R. Schizophrenia. *The Lancet 363*, 9426 (2004), 2063 – 2072.

[44] Mulders, P. C., van Eijndhoven, P. F., Schene, A. H., Beckmann, C. F., and Tendolkar, I. Resting-state functional connectivity in major depressive disorder: a review. *Neuroscience & Biobehavioral Reviews 56* (2015), 330–344.

[45] Nair, V., and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)* (2010), pp. 807–814.

[46] Nouri, D. nolearn: scikit-learn compatible neural network library, 2014.

[47] Orr, G. Momentum and learning rate adaptation. https://www.willamette.edu/~gorr/classes/cs449/momrate.html. Acessado: 20-05-2018.

[48] Os, J. v., Kenis, G., and Rutten, B. The environment and schizophrenia. *Nature 468*, 7321 (2010), 203–212.

[49] Pickhardt, P. J., Pooler, B. D., Lauder, T., del Rio, A. M., Bruce, R. J., and Binkley, N. Opportunistic screening for osteoporosis using abdominal computed tomography scans obtained for other indications. *Annals of internal medicine 158*, 8 (2013), 588–595.

[50] Rodrigues-Amorim, D., Rivera-Baltanás, T., López, M., Spuch, C., Olivares, J. M., and Agís-Balboa, R. C. Schizophrenia: A review of potential biomarkers. *Journal of Psychiatric Research* (2017).

[51] Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015), Springer, pp. 234–241.

[52] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[53] Siqueira, P. G., and Vergara, R. F. Segmentação e parametrização de imagens de ressonância magnética do cérebro: método semi-automático de extração de características para apoio a diagnóstico de pacientes com esquizofrenia. *University of Brasilia* (2016).

[54] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *Journal of machine learning research 15*, 1 (2014), 1929–1958.

[55] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *International conference on machine learning* (2013), pp. 1139–1147.

[56] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1–9.

[57] Takayanagi, Y., Takahashi, T., Orikabe, L., Mozue, Y., Kawasaki, Y., Nakamura, K., Sato, Y., Itokawa, M., Yamasue, H., Kasai, K., et al. Classification of first-episode schizophrenia patients and healthy subjects by automated mri measures of regional brain volume and cortical thickness. *PloS one 6*, 6 (2011), e21047.

[58] Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints abs/1605.02688* (May 2016).

[59] Tong, S., and Koller, D. Support vector machine active learning with applications to text classification. *Journal of machine learning research 2*, Nov (2001), 45–66.

[60] van der Burgh, H. K., Schmidt, R., Westeneng, H.-J., de Reus, M. A., van den Berg, L. H., and van den Heuvel, M. P. Deep learning predictions of survival based on mri in amyotrophic lateral sclerosis. *NeuroImage: Clinical 13* (2017), 361–369.

[61] van Erp, T. G., Hibar, D. P., Rasmussen, J. M., Glahn, D. C., Pearlson, G. D., Andreassen, O. A., Agartz, I., Westlye, L. T., Haukvik, U. K., Dale, A. M., et al. Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the enigma consortium. *Molecular psychiatry 21*, 4 (2016), 547–553.

[62] Vernooij, M. W., Ikram, M. A., Tanghe, H. L., Vincent, A. J., Hofman, A., Krestin, G. P., Niessen, W. J., Breteler, M. M., and van der Lugt, A. Incidental findings on brain mri in the general population. *New England Journal of Medicine 357*, 18 (2007), 1821–1828.

[63] Vieira, S., Pinaya, W. H., and Mechelli, A. Su82. using convolutional neural networks to examine structural abnormalities in schizophrenia. *Schizophrenia Bulletin 43*, suppl_1 (2017), S190–S190.

[64] Vieira, S., Pinaya, W. H., and Mechelli, A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews* (2017).

[65] Wiebers, D. O., of Unruptured Intracranial Aneurysms Investigators, I. S., et al. Unruptured intracranial aneurysms: natural history, clinical outcome, and risks of surgical and endovascular treatment. *The Lancet 362*, 9378 (2003), 103–110.

[66] Wright, G. A. Magnetic Resonance Imaging. *Signal Processing Magazine, IEEE 14*, 1 (1997), 56–66.

[67] Zeiler, M. D., and Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision* (2014), Springer, pp. 818–833.

[68] Zeng, L.-L., Wang, H., Hu, P., Yang, B., Pu, W., Shen, H., Chen, X., Liu, Z., Yin, H., Tan, Q., et al. Multi-site diagnostic classification of schizophrenia using discriminant deep learning with functional connectivity mri. *EBioMedicine 30* (2018), 74–85.

[69] Zhou, Z., Kong, B., Yu, C., Shi, X., Wang, M., Liu, W., Sun, Y., Zhang, Y., Yang, H., and Yang, S. Tungsten oxide nanorods: an efficient nanoplatform for tumor ct imaging and photothermal therapy. *Scientific reports 4* (2014), 3653.

[70] Zhou, Z.-H. *Ensemble methods: foundations and algorithms*. CRC press, 2012.

A Apendice

A.1 Instalação Theano e Lasagne

Lasagne é um framework dedicado ao treinamento de redes neurais utilizando a biblioteca de funções matemáticas Theano, que por sua vez serve para definir e otimizar eficientemente funções matemáticas multi dimensionais utilizando GPU.

Para a instalação dos Theano, alguns pré-requisitos são necessários, como:

- Python versões 2.7 ou 3.4 até 3.6
- NumPy
- SciPy
- BLAS

Estes pacotes podem ser instalados utilizando o pacote de instalação conda:

- conda install numpy scipy mkl

Em seguida pode ser instalado o Theano com o comando,

- conda install theano pygpu

Após a instalação de todos pacotes acima, pode-se instalar o Lasagne utilizando os comandos:

1. pip install -r https://raw.githubusercontent.com/Lasagne/Lasagne/master/requirements.txt
2. pip install https://github.com/Lasagne/Lasagne/archive/master.zip