University of Brasilia at Gama – FGA/UnB
Biomedical Engineering Graduate Program

Detection of Schizophrenia Based on Brain Structural Analysis Using Machine Learning Applied to Different Combinations of Multi-Slice Magnetic Resonance Images

J.S. Avelar Filho

Advisor: Dr. Cristiano Jacques Miosso

University of Brasilia at Gama

Programa de
Pós-Graduação em
ENGENHARIA BIOMÉDICA

Detection of Schizophrenia Based on Brain Structural Analysis Using Machine Learning Applied to Different Combinations of Multi-Slice Magnetic Resonance Images

J.S. Avelar Filho

Advisor: Dr. Cristiano Jacques Miosso

Master Degree Thesis on
Biomedical Engineering

Publication: 152A/2022

Brasilia/DF, julho de 2022

UNIVERSITY OF BRASILIA AT GAMA

GRADUATE PROGRAM

DETECTION OF SCHIZOPHRENIA BASED ON BRAIN STRUCTURAL ANALYSIS USING MACHINE LEARNING APPLIED TO DIFFERENT COMBINATIONS OF MULTI-SLICE MAGNETIC RESONANCE IMAGES

J.S. AVELAR FILHO

MASTER THESIS SUBMITTED TO THE BIOMEDICAL ENGINEERING GRADUATE PROGRAM, AS A PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER IN BIOMEDICAL ENGINEERING

APPROVED BY:

____

Dr. Cristiano Jacques Miosso

(Advisor)

____

Dr. Fabricio Ataides Braz

(Internal examiner)

____

Dr. Joalbo Matos Andrade

(External examiner)

Brasilia/DF, julho de 2022

Catalog Card

Avelar Filho, J.S.
Detection of Schizophrenia Based on Brain Structural Analysis
Using Machine Learning Applied to
Different Combinations of Multi-Slice Magnetic Resonance Images
[Distrito Federal], 2019.
82p., 210 × 297 mm (FGA/UnB Gama, Mestrado em Engenharia Biomédica, 2022).
Dissertação de Mestrado em Engenharia Biomédica, Faculdade UnB Gama, Programa de Pós-Graduação em Engenharia Biomédica.

1. Machine Learning	2. Magnetic Resonance Imaging
3. Schizophrenia	4.
I. FGA UnB/UnB.	II. Título (série)

Reference

Avelar Filho, J.S. (2022). Detection of Schizophrenia Based on Brain Structural Analysis Using Machine Learning Applied to Different
Combinations of Multi-Slice Magnetic Resonance Images. Master thesis in Biomedical Engineering, Publication 152A/2022, Biomedical Engineering Graduate Program, University of Brasilia at Gama, Brasilia, DF, 82p.

Copyright

Autor: J.S. Avelar Filho

Título: Detection of Schizophrenia Based on Brain Structural
Analysis Using Machine Learning Applied to Different
Combinations of Multi-Slice Magnetic Resonance Images

Grau: Mestre

Ano: 2022

jeronimo.filho@unb.br

Brasília, DF – Brasil

Agradecimentos

Quero agradecer inicialmente aos meus pais, Jeronimo e Maria Maia, que me incentivaram ao estudo e me mostraram as oportunidades que ele proporcionaria em minha vida. Agradeço a minha esposa Josiane, pelo carinho e suporte em todos os momentos dessa jornada. Agradeço ao meu orientador Professor Cristiano Miosso pela dedicação e contribuição na elaboração deste trabalho e ao Professor Nílton Silva pelas contribuições relevantes para a manutenção do curso deste trabalho. Aos meus filhos pelo apoio e compreensão. A todos aqueles que direta ou indiretamente contribuíram para este trabalho.

Detecção de Esquizofrenia com Base em Análise Estrutural do Cérebro Usando Aprendizagem de Máquina Aplicada a Combinações de Cortes em Imagens Volumétricas de Ressonância Magnética

Resumo Estendido

A Esquizofrenia é uma doença mental com muitas manifestações clínicas, que transformam o diagnóstico em um grande desafio. Até que um diagnóstico seja finalizado, o paciente passa por muitos episódios de sofrimento mental que podem redundar em conflitos sociais, acidentes involuntários e até suicídios.

Apesar da complexidade clínica, um diagnóstico nos estágios iniciais da doença é de grande relevância. Vários estudos recentes, com foco na análise das modificações estruturais do cérebro, encontraram correlações com a esquizofrenia, e sugerem que a esquizofrenia pode ser diferenciada do caso controle com base em imagens anatômicas de ressonancia magnética.

Pesquisas anteriores aplicando aprendizagem de máquina a estas imagens de ressonância magnética apresentaram resultados promissores. Apesar dos resultados, o escopo destas pesquisas estava limitado a um ou poucos cortes do cérebro e também não utiliza os mais recentes algoritmos de aprendizagem de máquina nos seus classificadores. Em consequência, o uso de poucas fatias ou algoritmos mais simples pode levar a perda de informação devido à extração de caracteristicas em nível abaixo do desejado.

No presente estudo, criamos modelos de aprendizagem de máquina baseados em Redes Neurais Convolucionais (Convolutional Neural Networks - CNN), e avaliamos os parâmetros para treinamento. Para tanto, utilizamos um conjunto de dados para treinamento, correspondente a imagens de Ressonância Magnética, com imagens de um grupo de controle (com pessoas sem diagnóstico de distúrbio mental) e um grupo experimental (com portadores de esquizofrenia). Avaliamos também critérios para a seleção dos cortes a serem utilizados para compor o conjunto de dados (dataset) e as diversas combinações que podem levar a um melhor desempenho do classificador.

Obtivemos as imagens dos cortes pela extração uma a uma da estrutura 3D correspondente a um volume do crânio humano, em cada imagem. Os cortes são numerados usando os índices do eixo axial do volume mapeado. Experimentamos selecionar as fatias utilizando métricas como covariância e entropia, e os melhores resultados foram obtidos quando utilizamos o conceito de entropia para avaliar as imagens dos cortes.

Os cortes foram ordenados pelo critério de maior entropia. Com esse critério, fizemos a avaliação individual dos cortes pelo modelo de aprendizagem de máquina e a seleção dos conjuntos de cortes. Nossa abordagem foi criar um dataset com adição incremental dos cortes ordenados pela entropia e usá-lo como conjunto de treinamento do modelo de aprendizado de máquina. Primeiramente, foi treinado o modelo com apenas uma fatia. No próximo, passo foi treinado com duas fatias para compor o conjunto de dados e assim por diante, até criarmos um dataset com todas as imagens extraídas do volume representado.

Cada dataset criado foi submetido ao treinamento no modelo de aprendizado de máquina e foram obtidas as métricas de desempenho do sistema.

Nossos resultados sugerem que é possível obter do classificador acurácia próximo de 80% quando treinado com um conjunto de cortes previamente selecionado.

Neste trabalho, também exploramos o uso de Inteligência Artificial Explicável (Explainable Artificial Intelligence - XAI), para compreender o resultado da classificação do modelo.

Palavras-chave: Esquizofrenia, Deep Learning, Machine Learning, Aprendizagem de Máquina. Neurociência, Imageamento por Ressonância Magnética, Mudanças Estruturais no Cérebro, Entropia

1 Introdução

Esta pesquisa trata da detecção da Esquizofrenia em imagens estruturais do cérebro, obtidas por equipamento de imageamento por ressonância magnética. Esta pesquisa visa estudar o uso de mais de uma imagem de corte para avaliar estruturas anatômicas do cérebro.

Os desafios da esquizofrenia

A esquizofrenia é uma seria desordem psiquiátrica com graves implicações sociais e que afeta cerca de 1% da população adulta mundial [2]. Esta doença altera as relações do paciente com sua família , colegas de trabalho e amigos redundando em colapso das relações sociais. As esquizofrenia manifesta-se no inicio da vida adulta, possuindo uma combinação de sintomas positivos e negativos. Os sintomas positivos incluem alucinações, ilusões e pensamento desordenado; os sintomas negativos incluem falta de demonstração de emoções, baixa capacidade de se expressar em palavras, incapacidade de ter prazer em atividades que anteriormente gostava de realizar e incapacidade de iniciar e concluir atividades com metas [2]. Estes sintomas não são percebidos inteiramente pelos médicos e existem iniciativas como as descritas por G. Michalakis et al. [4], para projetar e desenvolver uma simulação baseadas em cenários de paciente com desordens mentais, visando a desenvolver uma percepção dos sintomas que um paciente de esquizofrenia enfrenta e desenvolver empatia com o paciente [4].

A esquizofrenia continua sendo essencialmente diagnosticada por um clinico [3]. O clinico precisa desenvolver um diagnostico subjetivo sobre o paciente a sua frente, principalmente pela observação dos sinais de pensamento desorganizado presentes na fala do paciente. Existe a possibilidade que este diagnostico subjetivo possa concluir que o paciente é tímido ou introvertido ao invés de corretamente diagnosticar a esquizofrenia. Este diagnostico impreciso pode acontecer mesmo quando o paciente é entrevistado por outros clínicos na busca por uma segunda ou terceira opinião sobre o paciente. Qualquer atraso na obtenção do diagnostico correto do paciente pode levar a eventos de conflito social, confusão mental e nos casos mais extremos pode levar a acidentes involuntários e até a tentativas de suicídio.

Diagnosticando a Esquizofrenia com o auxilio da Ressonância Magnética

O sistema de imageamento por ressonância magnética é uma ferramenta considerável para proceder exames não invasivos de estruturas anatômicas do cérebro. Estudos das estruturas internas do cérebro usando Ressonância Magnética, indicam a existência de uma redução significativa do lobo temporal e das estruturas temporais mediais [3]. Analise conduzidas em estudos por imagem de gêmeos monozigóticos, onde um deles foi afetado pela esquizofrenia e o outro não, mostraram que o gêmeo afetado apresentava ventrículos maores e menores tamanhos do córtex e do hipocampo. A redução do tamanho do cérebro é mais significativo no plano axial do que no plano sagital, sugerindo o envolvimento mais significativo de regiões tipicamente visíveis nos cortes axiais [3].

Limitações de estudos anteriores

A aprendizagem de maquina vem sendo aplicada a imagens de ressonância magnética para superar os limites de um exame clinico subjetivo e ajudar a descobrir mudanças na estrutura do cérebro. Esta abordagem visa dar suporte aos médicos na obtenção de um diagnostico objetivo o mais rápido possível. Fay(2019) propôs utilizar um modelo de aprendizagem de maquina baseado em redes convolucionais (CNN) para classificar imagens do cérebro obtidas por Ressonância Magnética e detectar esquizofrenia. Esta pesquisa teve resultados promissores mas a analise estava limitada a apenas dois cortes [8]. Ferreira(2016)propôs utilizar uma máquina de vetores de suporte (Support Vector Machine - SVM) como um classificador para detectar esquizofrenia. A pesquisa também obteve bons resultados a partir das analises das imagens de ressonância magnética mas teve o esforço adicional de selecionar manualmente os cortes medindo estruturas e avaliando quais características ofereceriam seriam as melhores para serem submetidas ao treinamento e predição pelo modelo [1]. Os resultados em acurácia, precisão e sensibilidade foram menores que os obtidos em [8]. Os resultados sugerem que a seleção manual e as características medias podem ter perdido informação importante no que concerne as imagem originais já que ambas as pesquisas usaram o mesmo conjunto de dados . Um artigo recente por JIhoon et al [7], também descreve o uso de uma rede neural profunda 3D para detectar esquizofrenia em conjuntos dados de imagens estruturais obtidas por Ressonância Magnética, com uma abordagem de converter cada imagem de corte em um frame e combinando todas elas para transforma em um video.

Os conjunto de dados de imagens obtidas por ressonância magnética contem poucas imagens de pessoas diagnosticadas com esquizofrenia e de pessoas saudáveis usadas como controle. Muitos aspectos de privacidade estão associadas com a disponibilidade dos dados de forma aberta a todos e os pesquisadores interessados nesses dados precisam concordar com os termos de contratos de manutenção da privacidade dos dados para que seja possível usá-los. Uma abordagem bastante usada é utilizar estrategias de aumento de dados (data augmentations) pela realização de transformações das imagens originais como modificações espaciais , filtragens por mudança de banda de passagem como proposto por Y. Nin et al [6]. Estudos anteriores apontaram que uma forma de obter diagnósticos mais acurados usando imagens de ressonância magnética associadas a aprendizagem de maquina mas foram notadas as seguinte limitações:

1. Uso de apenas um corte do imageamento por ressonância magnética, nos entendemos que usando mais de um corte para treinamento nos levará a identificar estruturas relevantes presentes em outros cortes;

2. A combinação de diferentes cortes do dataset usados no treinamento do modelo de inteligencia artificial poderá identificar informações sobre alterações anatômicas e proporções existentes de um corte para outros;

3. Comparação de diferentes combinações de cortes podem levar a uma melhor acurácia na classificação de imagens

4. Alguns sistemas usados para classificação de imagens não são transparentes em como atingem seus resultados e existe uma oportunidade para explicar as razoes para um resultado obtido;

5. Existem implicações de ordem pratica e legal relacionadas com o fornecimento de diagnostico automatizado sem uma explicação que possa ser entendida por pessoas , detalhando os passos realizados para atingir os resultados

Problema cientifico e proposta de pesquisa

Estudos baseados em imagens do cérebro de pacientes de esquizofrenia mostraram que existem mudanças estruturais características que não podem ser atribuídas a efeitos de drogas ou outros fatores. Imageamento usando projeção [5] identificaram diferenças geométricas ao nível de neurônios entre pacientes de esquizofrenia e indivíduos sadios: os pacientes de esquizofrenia apresentam uma rede entre neurônios muito tênue e tortuosa que não é visível nos indivíduos sadios, sugerindo estar associada a esta doença.

As limitações previamente descritas culminaram na seguinte pergunta de pesquisa:

É possível criar um conjunto cuidadosamente escolhido de imagens originadas exames por ressonância magnética, que quando submetidas a um modelo de aprendizado de maquina é capaz de dar suporte a um diagnostico de Esquizofrenia mais objetivo ?

Nossa pesquisa tem uma hipótese central de que algumas combina coes de diferentes imagens de corte oriundas de um exame de ressonância magnética , quando são submetidas a uma rede neural convolucional , CNN, pode ter uma performance melhor na detecção de esquizofrenia devido a possibilidade de avaliar relações entre estruturas anatômicas tridimensionais ao invés de avaliar uma única imagem de um corte bidimensional. Além disso, uma analise de aspectos relacionados as combinações de imagens de cortes , por exemplo, a entropia, pode revelar e identificar correlações entre esses valores e a performance de um classificador de imagens baseado em redes neurais convolucionais. É esta hipótese que iremos avaliar na nossa pesquisa.

Objetivos

Objetivos gerais

Nosso objetivo geral e avaliar de forma sistemática as varias combinações de imagens de cortes anatômicos axiais de exames de ressonância magnética, usando redes neurais convolucionais para detectar a presença de esquizofrenia, de acordo com rótulos previamente determinado por especialistas neste domínio de conhecimento e comparar a performance com classificadores que utilizam uma onica imagem de corte. Para atingir este objetivo m nos vamos analisar aspectos com a covariância e a entropia. Também vamos conduzir testes empíricos com modelos de redes neurais convolucionais para determinar a performance com diferentes combinações de imagens de cortes.

Objetivos específicos

Para atingir o objetivo geral desta pesquisa, nos propomos os seguintes objetivos específicos:

- Experimentar com métodos como entropia e correlação de Pearson para identificar as imagens de contes contendo mais informação que irão melhorar os resultados dos classificadores.
- Experimentar com diversas combinações de imagens de cortes de exames de ressonância magnética ao invés de usar a imagem de um único corte , visando obter as melhores métricas de performance de um modelo de classificação de imagens
- Avaliar a performance de uma pequena rede neural convolucional e comparar com a performance de uma rede neural utilizando uma arquitetura estado da arte , pre treinada;
- Avaliar técnicas de Inteligencia Artificial Explicável , XAI, aplicada ao problema de detecção de esquizofrenia.

2 Fundamentação teórica e estado da arte

Esquizofrenia

A Esquizofrenia é uma doença mental com muitas manifestações clinicas que tornam o diagnostico medico um desafio significativo. Pacientes são levados aos médicos apos eventos perturbadores envolvendo o paciente e sua família e outros grupos sociais. Esta doença afeta 1% da população adulta mundial e geralmente começa no inicio da idade adulta. O pico dos sintomas acontecem por volta da terceira de cada de idade ocorrendo alguns poucos anos antes em homens do que em mulheres. O curso do desenvolvimento da doença é variável. Apenas uma minoria dos pacientes irão apresentar deterioração da situação evoluindo para um estado cronico , enquanto muitos outros terão sintomas persistentes ou deficit funcionais.

Um diagnostico que possa ser obtido ainda nos estágios iniciais da doença é de grande importância já que irá guiar os médicos na prescrição de tratamento ainda nos estágios iniciais da doença, evitando o terrível sofrimento mental e as consequências sociais que um colapso mental do paciente podem ocasionar.. Além disso, pesquisas com o objetivo de identificar exames que possam detectar a esquizofrenia usando parâmetros objetivos e mensuráveis são de crucial importância. Imageamento por ressonância magnética associado a algoritmos de aprendizagem de maquina podem se tornar a melhor ferramenta para detectar mudanças características estruturais em regiões do cérebro de um paciente portador de esquizofrenia.

Imageamento por Ressonância Magnética

Imageamento por Ressonância Magnética é uma técnica para imagens de cortes anatômicos internos e da organização funcional de uma organismo sem precisar abri-lo. Da mesma forma que outros dispositivos de tomografia, o equipamento de imageamento por ressonância magnética é capaz de gerar matrizes de dados multidimensionais representando a distribuição espacial de medidas de estruturas físicas. Apesar de semelhante a outras tecnologias, o imageamento por ressonância magnética é superior a outras tecnologias já que consegue gerar imagens bidimensionais representandos cortes em qualquer orientação. imagens tridimensionais representando volumes e até imagens com 4 dimensões a partir de distribuições espectro espaciais sem qualquer modificação ou ajuste do equipamento para realizar as diferentes operações; adicionalmente, a operação do equipamento de imageamento por ressonância magnética é muito seguro devido a sua banda de transmissão de radiofrequência e o processo de imageamento não usa radiação ionizada, evitando efeitos danosos aos pacientes. O primeiro uso desta técnica foi para obter imagens anatômicas e morfológicas de cortes finos do corpo humano , mas novas aplicações estão sendo aperfeiçoadas continuamente, tais como imageamento funcional e fisiológico aplicado a todos os sistemas do corpo humano. Hoje em dia o imageamento por ressonância magnética evoluiu de exame de imagens de cortes bidimensional para uma técnica de analise de imagens volumétricas.

Engenharia de características

Enquanto analisávamos as imagens dos cortes obtidas por ressonância magnética visando construir os datasets para treinar o classificador baseado em aprendizagem de maquina , observamos que alguns cortes eram bem similar a outros cortes e alguns deles pareciam não conter informação relevante já que eram imagens da parte superior do cranio contento apenas o contorno dos ossos que formam a caixa craniana

Nossa abordagem para solucionar o problema foi utilizar a engenharia de características. Em essência o objetivo da engenharia de características é identificar as características que são mais relevantes em um conjunto de dados para um dado modelo de aprendizagem de maquina. Ao aplicar a engenharia de características a um problema é necessário aplicar transformações aos dados quantificar características e compará-la com outras, identificar redundâncias e correlações. A Engenharia de características é uma parte essencial de uma experimento de exploração em aprendizagem de máquina por tratar dados brutos e aparentemente não relacionados convertendo-os em dados relevantes com a ajuda de ferramentas estatísticas e de aprendizagem de maquina. Com o uso desta técnica o modelo pode convergir para ótimos parâmetros com a utilização de menos recursos computacionais.

Covariância

No estudo da teoria da probabilidade e estatística, a covariância mede o variabilidade conjunta de duas varáveis aleatórias. A covariância indica que os grande valores de duas variáveis estão relacionadas ou se seus menores valores estão relacionados. A covariância na situação exemplificada será positiva e se as variáveis se comportam de forma inversa então ela será negativa , indicando uma relação inversas entre as variáveis. Finalmente se o valor da covariância é zero então não existem relacionamento entre as variáveis.

Correlação de Pearson

A covariância busca mostrar se existe um comportamento de interdependência linear entre duas variáveis , mas a covariância é uma medida dimensional que é afetada pelas unidades das medias das series de valores sob analise. De forma a corrigir esta situação, a estatística dispõe da ferramenta correlação de Pearson. Esta media é uma normalização da covariância representada por um numero adimensional variando de -1 a +1. Quando o valor da correlação de Pearson é +1, ela indica que existe uma relação linear direta, ou seja quando uma serie de números aumenta o seu valor, a outra também aumenta. Quando o valor da correlação é -1 então existe uma perfeita relação linear inversa, quando uma serie aumenta os seus valores, na outra serie os valores diminuem. Se o valor da correlação é próximo de zero então o relacionamento entre as variáveis é mínimo.

Entropia

A entropia representa um conceito cientifico inicialmente introduzido como uma declaração da segunda lei da termodinâmica. É comumente associada com um estado de aleatoriedade, incerteza ou desordem. O termo é aplicado em diversos campos científicos, desde as suas raízes na termodinâmica até os princípios da teoria da informação. É usado principalmente na química, física, biologia, cosmologia, estudos climáticos, sociologia e sistemas de informação. De acordo com Cover, T. M. [5], Hartley, em 1930, introduziu a medida logarítmica da informação para comunicação que era, em essência, o logaritmo do tamanho do alfabeto.

A entropia é significativa no domínio da ciência de dados e na Inteligencia Artificial. Ele é usado para criar arvores de classificação, é a base da Informação Mutua que é usada para, por exemplo, medir o relacionamento entre dois conjuntos de dados. A Entropia também é a base para a Entropia Relativa (The Kullback Leibler Distance) e da Entropia Cruzada usada nos algoritmos de redução da dimensionalidade, tais como t-SNE e UMAP, pela quantificação das similaridades e diferenças. Simplificando , o resultado do calculo da entropia nos mostra quão randômicos são os valores em um conjunto de dados. Por exemplo, a entropia está associada com o quanto ficaremos surpresos ao escolher um valor de um conjunto de dados e prever qual valor será.

No domínio da nossa pesquisa, a Entropia nos fornece um valor numérico que indica a diversidade da informação contida na imagem de um corte do cérebro. Nossa intuição para este critério de seleção é que quanto maior o valor da entropia, mais estruturas estão contidas na imagem de um corte.

Redes Neurais Profundas e Redes Neurais Convolucionais

As Redes Neurais Profundas é uma das áreas de estudo da Inteligencia Artificial e lida com algoritmos de aprendizagem de maquina capazes de aprender em diversos níveis de representação, correspondendo a uma hierarquia de características ou fatores onde conceitos de alto nível são definidos a partir de conceitos de baixo nível. O aprendizado profundo Deep Learning consiste em métodos de representação do conhecimento com múltiplos níveis de representação como resultado da composição de módulos não lineares mais simples. Cada um deles transforma as representações em seu nível, começando com uma entrada bruta ate chegar em uma representação de alto nível, com um nível mais alto de abstração. Com a utilização de um numero suficiente de módulos, funções muito complexas podem ser aprendidas e realizadas como reconhecimento de imagens e da fala humana. As redes de aprendizado profundo receberam muita atenção da comunidade cientifica devido aos diversos benchmarks em que teve desempenho significativo. As redes neurais profundas podem superar as técnicas convencionais de aprendizagem de maquina por conta da sua habilidade de aprendem a partir de dados brutos através de filtragens encadeadas, operações lineares e não-lineares , levando a reconhecimento de padrões de grande complexidade em altos níveis de abstração.

As redes neurais profundas podem ser de de diversos tipos de arquitetura mas a rede Neural Convolucional é a primeira opção dos pesquisadores quando se trata de problemas envolvendo classificação e reconhecimento de imagens. A principal característica deste tipo de rede é a presença de uma camada Convolucional, onde é realizada a operação matemática de convolução entre os dados de entrada e um kernel contendo um padrão de valores capazes de reconhecer um determinado padrão de interesse se estiver presente nos dados de entrada.

XAI - Inteligencia Artificial Explicável

A sofisticação e adição de cadas vez mais camadas nas arquiteturas das redes neurais profundas trouxeram uma capacidade cada vez maior de aprendizado mas em contrapartida as Redes Neurais profundas tornaram-se verdadeiras caixas-pretas dificultando a compreensão das razões pelo qual um determinado resultado foi alcançado.

Esta situação motivou agencias governamentais , empresas e instituições acadêmicas a financiar pesquisas que estudem formas de garantir que os resultados obtidos como saída de modelos de aprendizagem de maquina possam ser confiáveis e também explicáveis. Uma das primeiras iniciativas no sentido de obter explicação de sistemas baseados em inteligencia artificial foi o programa Explainable Artificial Intelligence (XAI) patrocinado pelo Departamento de Defesa dos Estados Unidos, Defense Advanced Research Projects Agency (DARPA); Este programa cunhou o termo XAI, com X de Explainable, explicável, com a intenção explicita de criar sistemas de inteligencia artificial compreensíveis por seres humanos , através de explicações práticas. O principal objetivo do programa foi criar um coleção de técnicas de aprendizagem de maquina pra construir modelos explicáveis que combinados com técnicas de explicação pudesse permitir aos usuários entender com modelo de IA funciona corretamente,

A XAI tem duas principais abordagens usadas como base para suas técnicas e métodos. A primeira abordagem é interpretar e justificar uma predição de um modelo de Machine Learning ou Deep Learning com seus dados de entrada. Esta abordagem é conhecida como Post-Hoc , já que a explanação é obtida apos a predição. A segunda abordagem é construir um modelo de inteligencia artificial que seja naturalmente explicável, sendo desde o seu projeto concebido para explicar como obtêm seus resultados, sendo por isso conhecida como abordagem Ante-Hoc. a abordagem Post-hoc é a base de diversas ferramentas importantes para explicar modelos de redes neurais profundas com grande numero de camadas, pesos e parâmetros que de outra forma seriam caixas-pretas não interpretáveis. A abordagem Post-Hoc é agnóstica no que diz respeito ao modelo que está explicando. Ela não está interessada nas funções internas do modelo mas sim em criar um modelo equivalente ao original que a partir das saídas do modelo sob analise ira produzir um resultado explicável.

3 Materiais e Métodos

Os métodos utilizados na nossa pesquisa foram desenvolvidos para dar suporte ao objetivo geral de avaliar sistematicamente as diferentes combinações de imagens de cortes axiais obtidos a partir de imageamento por ressonância magnética, para o problema de detecção da presença ou não de modificações estruturais no cérebro causadas pela esquizofrenia , pela utilização de modelos de redes neurais convolucionais. Para atingir este objetivo geral, executamos métodos que nos permitiram cumprir os seguintes objetivos objetivos específicos: - Experimentar com a medida da entropia e com a correlação de Pearson para identificar as imagens de cortes com mais informações e que melhorassem os resultados do nosso classificador. - Experimentar com combinações de varias imagens de cortes por imageamento por ressonância magnética ao invés de usar uma única imagem de um corte central, visando obter as melhores métricas de performance de uma modelo classificador de imagens - Avaliar a performance de uma pequena rede neural convolucional comparada com uma arquitetura de rede estado-da-arte , pré-treinada para para classificação de imagens - avaliar técnicas de Inteligencia artificial Explicável aplicadas ao problema de detecção de esquizofrenia por um modelo classificador.

Nosso experimentos consistiram em

- Preprocessar imagens brutas originadas de exames por ressonância magnética de pacientes portadores de esquizofrenia e de pacientes sadios do grupo de controle. A partir dos dados de entrada gerar imagens bidimensionais de cortes do eixo axial do cérebro, usando-os para criar um conjunto de dados para a condução dos experimentos subsequentes.

- Usando o conceito da entropia de Shannon . avaliamos os cortes individuais e as combinações de cortes, buscando obter a combinação que poderia gerar os melhores resultados da maquina de aprendizagem usada como classificador.

- Usando a medida da covariância, nos avaliamos se imagens de cortes individuais e combinações de cortes poderiam ser correlacionados ou se haveria alguma dependência entre eles, indicando algum tipo de redundância de informação que poderia levar a uma otimização da quantidade de dados.

- Apresentar uma combinação de imagens selecionadas para uma rede neural convolucional para treinamento e avaliação . Os passos de treinamento e avaliação foram repetidos para diversas imagens de cortes e para diversas combinações.

- Baseado no experimento anterior, foram identificados as coleções de cortes que tinham as melhores métricas de performance.

- Submeter o conjunto de dados de cortes selecionados a uma arquitetura de rede neural convolucional o mais simples possível para avaliar os resultados

- Submeter o conjunto de dados de cortes selecionados a uma arquitetura de rede neural convolucional estado-da-arte e pre-treinada para classificação de imagens

- Comparar os resultados dos dois experimentos anteriores

- Aplicar um método de Inteligencia artificial Explicável para verificar o que esta sendo levado em consideração pelo modelo estado-da-arte de classificação de imagens, no momento em que classifica as imagens de cortes.

- Exibir os resultados dos experimentos em tabelas.

4 Resultados e Discussões

Experimentamos os valores de covariância/correlação de Pearson para investigar se havia indicação de redundância de informações que poderia levar à seleção das fatias mais relevantes para construir nosso conjunto de dados. Nossa intuição foi verificar se fatias adjacentes teriam uma grande correlação entre elas que poderia levar à eliminação de uma das fatias. Uma combinação com maior chance de estar nesta hipótese foi entre as fatias 1 e 2, mas com um pequeno valor de correlação de 0,234. Além dessa combinação especifica, a maioria das combinações estavam abaixo desse valor, por isso não consideramos essa abordagem para seleção de fatias.

A Avaliação de entropia de fatias individuais e de conjuntos de fatias apresentou resultados mais promissores. Os passos realizados na metodologia obtiveram uma lista de fatias, ordenadas pelo seu valor de entropia. Essa lista de fatias foi usada para construir o conjunto de dados para o classificador por aprendizado de máquina e obter métricas sobre seu desempenho. Esta informação permitiu explorar quais fatias melhoravam ou não os resultados do classificador por aprendizado de máquina.

Observamos que a fatia seis sozinha foi responsável por obter os melhores valores de métricas, não importa se usando as métricas perda ou precisão para comparação entre as coleções de fatias. Analisando os resultados ordenados por métricas de perda, observamos que as fatias 6, 7, 10 e 9 em combinações incrementais representam as principais combinações para precisão. As tabelas de resultados, ordenadas por precisão, mostram informações semelhantes, com as fatias 6, 7,10 e 9 presentes para as melhores combinações novamente, mas notamos que a segunda melhor precisão foi obtida por uma lista mais longa, composta pelas fatias 6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, um resultado um tanto inesperado.

Uma das tarefas finais desta pesquisa foi submeter os conjuntos de dados de coleções de fatias a uma arquitetura de rede neural de última geração, treiná-la, validá-la, coletar seus resultados e compará-los com os resultados de um modelo inicial usado para criar uma avaliação por linha de base. Usamos uma arquitetura VGG16 que embora não seja o mais recente, é ainda uma arquitetura muito utilizada para classificadores de imagens e detecção de objetos.

Com um modelo de arquitetura VGG16 adaptado ao nosso problema de classificação, submetemos o modelo para treinamento e avaliação do conjunto de dados de validação. Após aproximadamente 13 horas de treinamento para cada uma das 2 melhores coleções de fatias de precisão, obtivemos dados e montamos uma tabela com os resultados do desempenho do treinamento.

O experimento final da nossa pesquisa foi criar uma visualização de explicação sobre a predição feita pelo modelo. Os resultados foram obtidos usando um modelo customizado baseado em arquitetura VGG16 de última geração. O modelo foi treinado usando transferência de aprendizado como estratégia para reduzir o tempo de treinamento, usando os pesos anteriores da arquitetura treinado no conjunto de dados Imagenet.

5 Conclusão

Neste trabalho, propusemos avaliar uma forma sistemática de determinar quais combinações de cortes axiais de ressonância magnética anatômica levariam a melhores resultados para o problema de classificação usando uma Rede Neural para detectar a presença de esquizofrenia ou não.

Primeiro, experimentamos com a Covariância/Correlação de Pearson para determinar se um fatia poderia ser correlacionada com outras fatias, mas os resultados não indicaram uma correlação entre as fatias.

Depois disso, experimentamos com Entropia. Desta vez, nossos resultados indicaram que o valor de entropia foi uma métrica significativa para indicar quais imagens são mais relevantes em um conjunto de dados para treinar o modelo de aprendizado de máquina, trazendo uma intuição de que quanto maior o valor da entropia da fatia, mais diversificado era o conteúdo da imagem, com mais estruturas representadas nele.

O próximo passo foi enviar a lista de fatias para uma avaliação do modelo de aprendizado de máquina e construir uma lista de coleções de fatias ordenadas por precisão. Esta lista nos mostrou essa fatia seis sozinha tinha a melhor precisão. A segunda melhor coleção em acurácia foi constituída das fatias 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 e 16.

Em todos os experimentos anteriores que precisavam de um modelo de aprendizado de máquina para avaliar o precisão de fatias e coleções, usamos uma pequena rede neural composta por 2 camadas convolucionais e camadas acessórias. Este modelo foi usado como linha de base para treinar e avaliar modelos e conjuntos de dados em menos tempo, pois os experimentos precisavam ser repetidos muitas vezes. No entanto, para uma avaliação final, optamos por experimentar com o estado da arte em arquitetura de última geração, não a mais recente, mas ainda relevante hoje em dia, a arquitetura VGG16. Com o auxílio da estratégia de transferência de aprendizagem, construímos um modelo customizado com esta arquitetura e submetemos a ela as duas melhores coleções fatias ordenadas por precisão. A curva de acurácia obtidas com essa arquitetura nos mostraram que ao utilizar as fatias de 0 a 16, este modelo pode atingir uma precisão superior a 80Nosso último experimento foi com Inteligência Artificial Explicável, a XAI, para desenvolver uma explicação sobre o que nosso modelo VGG16 personalizado estava procurando ao fazer sua classificação. Para o conjunto de dados apenas contendo a fatia seis, o modelo concentrou-se na parte central da imagem e para o conjunto de dados composto por fatias de 0 a 16, observou a parte central das fatias e o parte de trás do cérebro.

Assim, com base em nossa pesquisa, podemos resumir que:

1. A entropia é uma métrica de interesse para avaliar e selecionar as imagens de cortes de exames de ressonância magnética para compor o conjunto de dados de imagens.

2. A arquitetura VGG16 ainda é um bom modelo de arquitetura para ser usado como base para uma problema de classificação de imagens.

3. A imagem do corte seis nos confrontou com uma situação inesperada, pois por si só, tinha um desempenho melhor em precisão ao ser submetido ao modelo CNN de linha de base, embora aparente não ter nenhuma estrutura relacionada ao problema da esquizofrenia;

   Podemos sugerir três hipóteses a partir dessa observação:

   • Há um erro em nosso experimento. Alguma situação equivocada na metodologia utilizada para preparar o conjunto de dados ou a construção da arquitetura, determinação de parâmetros ou tamanhos de kernel de convolução, nos levaram a um erro.

- Existe uma relação oculta entre a fatia seis e a doença que não é conhecida ainda. Como notamos que temos tamanhos diferentes para crânios humanos, a fatia seis pode ser fatia sete ou oito no crânio de outra pessoa, desta forma precisamos desenvolver um método para normalizar as medidas e posições dos cortes no crânio e repetir os experimentos com estes novos dados.

- A IA pode nos trazer duas abordagens ao comparar o desempenho de um modelo de classificação por aprendizagem de maquina, a um especialista humano: - Treinado corretamente com uma grande quantidade de dados rotulados por um especialista, uma máquina de IA pode superar um humano para a mesma tarefa de classificação, abrindo a oportunidade para automatizar essa atividade; - Uma máquina de IA pode descobrir padrões e relações que nem suspeitávamos existir, percebendo detalhes que um ser humano não percebeu antes ou deliberadamente teria descartado como possivelmente irrelevante.

4. Embora a fatia seis sozinha tenha a melhor precisão em nossa avaliação do modelo CNN de linha de base, a coleção composta por fatias de 0 a 16 obteve a melhor precisão quando submetida ao modelo customizado baseado na arquitetura VGG16.

5. Estudos anteriores mostraram uma preferência por usar fatias de 9 a 16 ao analisar pesquisas aplicando ML ao problema de detecção de esquizofrenia, mas nossa pesquisa mostrou que fatias de 0 até 16 são relevantes para os resultados dos modelos de classificação.

Lista de Referências

[1] B. F. da Cruz. Classificação de esquizofrenia com base em máquinas de suporte vetorial aplicadas a características de imagens de ressonância magnética. Master's thesis, 7, 2016. University of Brasília at Gama.

[2] G. Flores, J. C. Morales-Medina, e A. Diaz. Neuronal and brain morphological changes in animal models of schizophrenia. *Behavioural Brain Research*, páginas 190–203, 2016.

[3] P. J. Harrison. The neuropathology of schizophrenia: A critical review of the data and their interpretation. *Brain*, 122(4):593–624, 04 1999.

[4] G. Michalakis, M. Pavlou, G. Gerogiannis, et al. Another day at the office: Visuohaptic schizophrenia vr simulation. In *2020 IEEE Conf. on Virtual Reality and 3D User Interfaces Abstracts and Ws. (VRW)*, páginas 515–516, 2020.

[5] R. Mizutani, R. Saiga, Y. Yamamoto, et al. Structural diverseness of neurons between brain areas and between cases, 2020.

[6] Y. Niu, Q. Lin, Y. Qiu, et al. Sample augmentation for classification of schizophrenia patients and healthy controls using ica of fmri data and convolutional neural networks. In *2019 Tenth International Conference on Intelligent Control and Information Processing (ICICIP)*, páginas 297–302, 2019.

[7] J. Oh, B. Oh, K. Lee, et al. Identifying schizophrenia using structural mri with a deep learning algorithm. *Frontiers in Psychiatry*, 11:16, 2020.

[8] R. F. Vergara. Detecção de alterações cerebrais anatômicas associadas à esquizofrenia com base em redes convolucionais aplicadas a imagens de ressonância magnética. Master's thesis, 3, 2019. University of Brasília at Gama.

Abstract

Schizophrenia is a mental disease with many clinical manifestations, making the diagnosis a significant challenge. Until a correct diagnosis is attained, the patient experiments with mental suffering that can lead to social conflicts, involuntary accidents, and suicides.

Despite the clinical complexity, early diagnosis is of utmost importance, and several recent studies focus on analyzing structural brain modifications that have been correlated to schizophrenia and can be detected in anatomical magnetic resonance images.

Previous research applying machine learning to such images presented promising results. However, the scope was limited to analyzing only one or few slices of the brain while not using recent algorithms at the core of the classifiers. Furthermore, using fewer slices and simple algorithms can lead to information loss due to sub-optimal feature extraction.

This study created machine learning models based on Convolutional Neural Networks (CNN) and evaluated the best training parameters. We used a Magnetic Resonance Images (MRI) dataset with scans from schizophrenia-diagnosed patients and a subjects control group. Also, we evaluated criteria to select the slices to be used to build a dataset and the various combinations of slices that could enhance the performance of an image classifier.

We obtained the slices by extracting them one by one from the 3D correspondent structure of a human skull for each image of the MRI scanning process. The slices were numbered based on the axial index of the mapped volume. We experimented with selecting the slices using metrics like covariance and entropy, and the best results were obtained when we used the entropy concept to evaluate the slice's images.

The slices were sorted by the greatest entropy. Using this criterion, we evaluated each slice individually, using a machine learning model and the collections of slices. Our approach was to create datasets with incremental addition of slices ordered by the entropy and use them as a training dataset for the machine learning model. First, we started training with a dataset containing only one individual slice from the scanned volumes. Then, in a second step, we used two slices to build the dataset, and so on, until we created a dataset with all the images extracted from the volume.

Each dataset created from the combinations of slices was used to train the ML model and evaluated to obtain the performance indicators.

Our results suggest that it is possible to obtain an accuracy near 80% when trained with a previously selected combination of slices.

In this study, we also explored the use of Explainable Artificial Intelligence (XAI) to comprehend the model output classification.

Keywords: Schizophrenia, Deep Learning, Machine Learning, Neuroscience, Magnetic Resonance Imaging, Structural Brain Changes, Entropy

Contents

List of Tables

List of Figures

1 Introduction

This research deals with detecting schizophrenia in brain structural images obtained by a Magnetic Resonance Imaging scanner. It specifically addresses using more than one brain image slice to evaluate brain anatomical structures.

1.1 Schizophrenia and its challenges

Schizophrenia is a serious psychiatric disorder with severe social implications, which affects about 1% of the adult world population [8]. It alters the relations between the patient, his family, co-workers, and friends, inducing a social breakdown. Schizophrenia starts in early adulthood and has a combination of positive and negative symptoms. The positive symptoms include hallucinations, delusions, and thought disorder; the negative symptoms include flat emotional expression, poor quality of speech, the inability to have pleasure with activities previously enjoyable, and the inability to start and complete goal-directed activities [8]. These symptoms are not entirely perceived by the clinician, and there are even initiatives, like those described by G. Michalakis et al. [18], to design and develop a scenario-based mental disorder simulation, aiming at empathizing the end-user with the symptoms that a person is living with schizophrenia faces [18].

Schizophrenia remains a clinical diagnosis [11]. A clinician has to make a subjective diagnosis about the person in front of her/him, mainly observing the signs of thought disorder present in the person's discourse. There exists a possibility that this subjective diagnosis can conclude that the patient is a little bit shy or less extroverted instead of correctly diagnosing schizophrenia. This imprecise diagnosis can even occur when submitting the patient to second and third opinions from other clinicians. Any delay in correctly diagnosing the disease can lead to social conflict events, mental confusion, and, in worst cases, involuntary accidents and even suicide attempts.

1.2 Diagnosing Schizophrenia with the assistance of structural MRI

Magnetic Resonance Imaging is a considerable tool for proceeding with non-invasive exams of brain anatomical structures. MRI studies of the brain's internal structures using MRI indicate a more significant reduction in the temporal lobe and medial temporal structures [11]. Analysis conducted on imaging studies of monozygotic twins, where one of them is affected by the disorder, showed the affected twin has larger ventricles and smaller cortical and hippocampal size. The brain size reduction is more significant in the axial plane than in the sagittal plane, suggesting more significant involvement of regions typically visible in axial slices [11].

1.3 Previous studies limitations

Machine learning is being applied to MRI brain scans to overcome the limits of a subjective clinical diagnosis and help discover changes in the brain structure. This approach aims to support physicians with a quicker and more objective diagnostic procedure. Fay (2019) proposed using a Machine Learning model based on Convolutional Neural Networks (CNN) to classify MRI scans of the brain and detect Schizophrenia. This research yielded promising results, but the analysis was limited to only two slices [32]. Ferreira (2016) proposed using a Support Vector Machine (SVM) as a classifier to detect Schizophrenia. He obtained good results from the analysis of the MRI scans, but with the added effort of manually selecting and measuring structures and evaluating which characteristics were the best to submit for training and prediction [6]. The results in accuracy, precision, and sensibility, were lower than those attained in [32]. The results suggest that the manually selected and measured features may have lost important information concerning the original images, as both studies used the same data sets. A recent article by Jihoon Oh et al. [22] also describes the use of a 3D deep CNN to detect Schizophrenia in structural MRI datasets, with an approach of converting each scan slice into a frame, combining all of them to transform into video.

The datasets of scanned images usually contain few images of persons diagnosed with schizophrenia and healthy control subjects. Many privacy concerns are associated with offering the data openly, and researchers must sign non-disclosure agreements to use datasets. One approach to data set augmentation is to create new images from transformations such as spatial smoothing and band-pass filtering, as proposed by Y. Niu et al. [21].

Previous studies had pointed the way for a more accurate diagnostics using MRI imaging associated with Machine Learning, but we noticed the following limitations:

1. Use of single slices of MRI imaging, we think that using more than one slice for training can lead to identifying relevant structures present in other slices;
2. The combination of different slices in the dataset used for model training can raise relevant information about anatomic alterations and proportions from one slice to another;
3. Comparison of different slices combinations can lead to better accuracy in classifying the images;
4. Some systems used for classification are opaque on how they achieve the results, and there is an opportunity to explain the reasons for a result;
5. There are practical and legal implications related to providing a diagnosis without a human-understandable explanation about the steps to achieve the results;

1.4 Scientific Problem and Research Proposal

Imaging studies of the brains of schizophrenia patients showed that there are characteristic structural changes that cannot be associated with drug effects or other factors. Imaging using projection [19] identified geometric differences at the neuron level between schizophrenia and control cases: schizophrenia cases showed a thin and tortuous neuronal network not visible in control cases, suggesting that it is associated with the disorder.

The limitations previously described sparked the following research question:

Is it possible to create a dataset of selected MRI scanned images that, when submitted to a machine learning model, support an early Schizophrenia diagnosis ?

Our research has a central hypothesis that some combinations of different MRI scan slices, when applied to a Convolutional Neural Network, CNN, can better perform schizophrenia classification due to the possibility of evaluating relations between anatomical tridimensional structures contrary to simply evaluating one single two-dimensional slice image. Furthermore, an analysis of aspects related to slice combinations, for example, the entropy, can reveal and identify correlations between these values and the performance of a CNN-based state-of-the-art image classifier. This hypothesis is what we will evaluate in this research.

1.5 Objectives

1.5.1 General Objective

Our general objective is to evaluate in a systematic way the various combinations of anatomical MRI scan axial slices for the problem of classification using CNN to detect the presence of schizophrenia or not, according to labels predetermined by specialists in the field and compare performance with single slices classification performance. To achieve this, we will analyze the slice's aspects like covariance and entropy. Also, we will conduct empirical tests with CNN models to determine the performance of different slice combinations.

1.5.2 Specific Objectives

In order to achieve the general objective of this research, we propose the following specific objectives:

- Experimenting with methods like entropy and Pearson's correlation to identify the most informative slices will enhance our classifier's results.
- Experimenting with various combinations of slices from MRI scans instead of a single central slice, aiming to obtain the best performance metrics from the classification model.
- Evaluate the performance of a small Convolutional Neural Network compared with a pre-trained state-of-the-art neural network architecture for image classification.
- Evaluation of techniques of Explainable Artificial Intelligence applied to schizophrenia classification.

1.6 Thesis Structure

We have structured this thesis in the following parts:

- Theoretical Foundation and State-of-the-Art: This part will present the theoretical basis of this research, describing the technologies and fundamental concepts we used to build this research.
- Materials and Methods: This item will present the material used, mainly the datasets containing the MRI scans. The methods will describe how we used the

obtained datasets, the treatments that we submitted the datasets, and how we processed these data in order to obtain the results

- Results and Discussions: Here, we will present the research results in tables comparing result metrics and discuss the results.

- Conclusion: The final part contains our conclusions from the experiments of the research, whether they align with our first hypothesis and conclude this work,

2 Theoretical Foundation and State-of-the-Art of Schizophrenia Analysis, Classification of Magnetic Resonance Images, and Explainable Classification Algorithms

2.1 Schizophrenia

Schizophrenia is a mental disease with many clinical manifestations that make diagnostics a significant medical challenge. Patients are conducted to clinicians mainly after disturbing events involving the patient, his family, and other social companions. This disease affects about 1% of the adult world population and generally starts in early adulthood. The peak age of onset is in the third decade, occurring a few years earlier in males than in females. The disease course of evolution is variable. Only a minority of the patients show chronic and deteriorating courses, while many others have enduring symptoms or functional deficits.

Researchers noted significant mortality from suicide and natural causes among persons affected by Schizophrenia. The clinician diagnoses Schizophrenia by evaluating medical history, investigating occurrences among family members, and interviewing the patient. The interviews aim to evaluate the patient's discourse, trying to identify the presence of delusions, hallucinations, and thought disorders. These are considered positive and complemented by negative symptoms, such as avolition (decrease in the motivation to initiate and perform self-directed purposeful activities), alogia (difficulty with speaking or the tendency to speak rarely), and affective flattening. *Thought Disorder* is described as confused speech, with terminologies such as *Formal Thought Disorder*, *Disorganized Thinking*, and *Disorganized Speech*, which refer to abnormalities in the amount and form of speech production associated with a disorganized thinking pattern [29]. The *Diagnostic and Statistical Manual of Mental Disorders* by the American Psychiatric Association [2] requires that these symptoms be present for at least six months to consider the possibility of Schizophrenia. Moreover, there must be impaired personal functioning, and the symptoms must not be secondary to another disorder such as depression or substance abuse. These criteria are very subjective and depend on the clinician's expertise and the patient's willingness to put feelings and thoughts in words.

An early diagnosis of schizophrenia is of great importance, as it will guide the prescription of treatment in the initial stages, avoiding the terrible mental suffering and social consequences of a patient's mental collapse. Therefore, research targeting exams that detect schizophrenia using measurable and objective parameters is crucial. MRI imaging, enhanced by Machine Learning algorithms, could be the ultimate tool for physicians, as it can help detect characteristic structural changes in brain regions of a patient with schizophrenia.

2.2 Magnetic resonance imaging – MRI

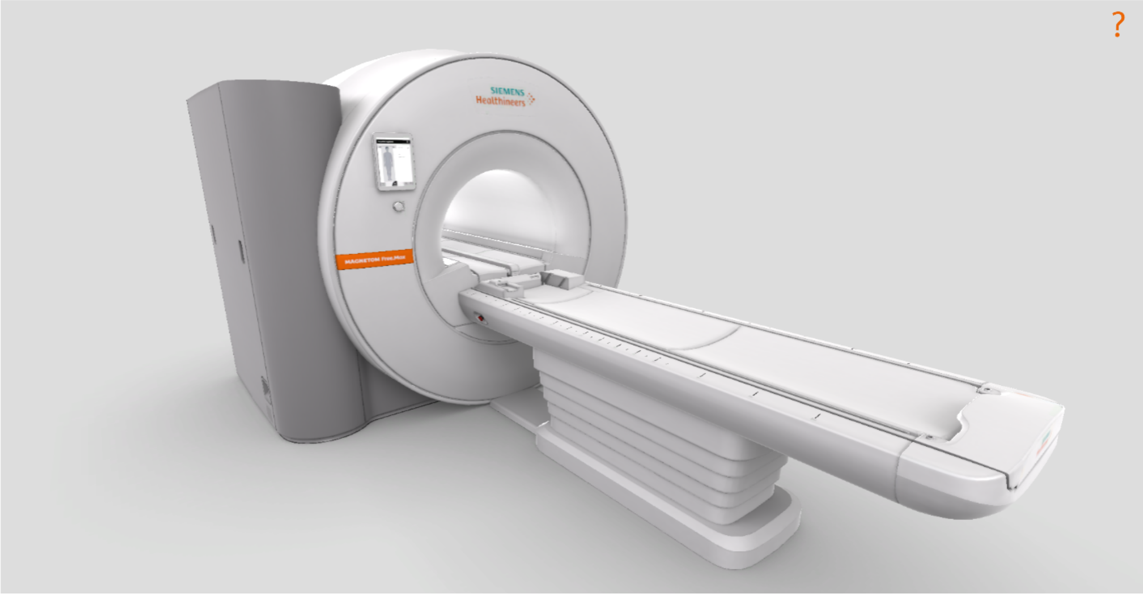


Figure 2.1. Siemens MRI Scanner MAGNETOM Free.Max. Source: [26]

Humans depend on their environment sensors (eyes, ears, noses, etc.), and the visual sensors, the eyes, are central for collecting and processing information relevant to our daily routines [16]. Based on this premise, medical images play an essential role in physicists' toolbox for evaluating patients' conditions and discovering diseases and anomalous health conditions. One of the best non-invasive medical imaging systems is the Magnetic Resonance Imaging (MRI), which can provide high-contrast images of structures, metabolism, and functioning of internal organs from biological systems, human or not. This innovative technology can provide images with excellent quality and ensure patient safety. It is continually being improved to support images with higher resolution and speed. Furthermore, with the lowers costs of computational resources, machine learning, and deep learning are used to understand the MRI image outputs better, building robust diagnosis systems.

MRI is based on the principles of nuclear magnetic resonance, a spectroscopic technique to obtain microscopic chemical and physical information about molecules. It is today the most versatile biomedical imaging technique [3]. Figure 2.2 illustrates the basic functioning of an MRI system. The invention of Magnetic Resonance Imaging was accomplished by many researchers that explored Nuclear Magnetic Resonance (NMR) and the physics of Magnetic Resonance Imaging in the early years of the 20th century. MR Imaging invention is credited to Paul C. Lauterbur as he developed a system to encode spatial information into an NMR Signal using magnetic field gradients in 1971 and published the theory sustaining it in 1973 [15].

MRI is a technique capable of creating tomographies (the Greek word ”tomos” meaning cuts), image cuts of the internal anatomical and functional organization of an object without opening it. Like other tomography devices, the MRI scanner can generate multidimensional data arrays representing the spatial distribution of measured physical quantities. However, MRI scanners exceed other devices as they can output 2D images representing sections at any orientation, 3D volumetric images, and even four dimensions images from spatial-spectral distributions without any modification or adjustment to the equipment to perform diverse operations. In addition, the operation of the MRI scanner is very secure as it operates in the radio-frequency range, and the imaging process does not use ionizing radiation, avoiding harmful effects on patients. The first use of the technique was to obtain anatomical and morphological images of thin slices through the human body, but new applications were discovered, such as functional and physiological imaging, which is applied to all systems of the human body. MRI has evolved from an initial tomographic imaging technique to a volume imaging technique.

2.2.1 Main components of a typical MRI scanner

A typical MRI scanner has the following main components:

- Main magnet: a resistive, permanent, or superconducting magnet whose primary function is to generate an intense uniform static field for polarizing nuclear spins in an object.
- Gradient system: usually consists of three orthogonal gradient coils designed to produce time-varying magnetic fields of controlled spatial non-uniformity. It is a crucial component as gradient fields are essential for signal localization.
- RF System: consists of a transmitter coil capable of generating a rotating magnetic field for excitation of a spin system and a receiver coil that converts a precessing magnetization into an electrical signal.

- Receiver/Digitizer: Consists of a very sensible antenna that detects the RF signals emitted by the patient's body under examinations, digitizes the signal, and feeds this information to the computer system.
- Computer system: a specialized computer to receive, record and analyze the images of the patient's body that have been scanned. Its primary role is to interpret the antenna's data via receiver and digitizer and produce an understandable image of the body part under examination.

Figure 2.2 shows how the main components of an MRI scanner are interconnected:

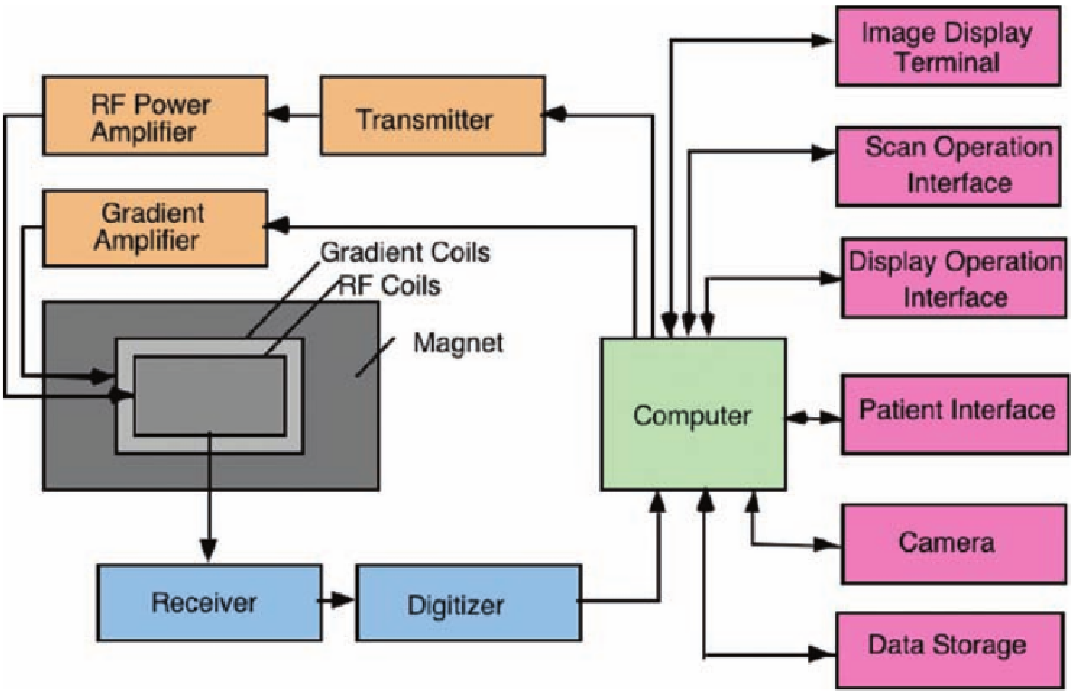


Figure 2.2. Simplified block diagram of typical MRI system. Functionally related subsystems such as transmit and receive chain, computer, and peripheral units such as patient and operator interface, are color-coded. Source: [3]

2.3 Feature Engineering

While analyzing the MRI Scans to build the datasets for training our machine learning engine, we had to treat the volume scans to obtain the image slices and analyze them to identify which scan-axis would present the most relevant structures. After obtaining the slices, we observed that some appeared to be quite similar, and some of them appeared to carry no significant information as they were images of the upper part of the skull with only bones of the cranial case.

These similarities and also the absence of relevant information, in our opinion, could indicate redundancies in the information contained in the images. We thought that could be some method for identifying what slices would be the best for building the dataset.

Our approach to solving this problem was to use Feature Engineering. In its essence, feature engineering is to identify the most relevant features in a dataset to a given supervised learning model and achieve it. One will need to apply transformations to the dataset, quantify features to compare against each other and identify redundancy and correlations. Feature Engineering is an essential part of a machine learning exploration experiment as it will help convert untreated, raw data into relevant features with the help of statistical and machine learning tools. With these techniques, the model can converge to the optimal parameters with fewer computational resources being used.

The reason to use feature engineering is that after applying transformations to the dataset, it will be more closely related to the target objective. Another possibility that feature engineering brings us is the possibility of bringing external data sources that could add information to the model. In our specific case, we thought of experimenting with covariance and entropy to evaluate if there was information redundancy between the slices by evaluating the covariance factor between them and which slices carried more information inside them by calculating their entropy.

2.4 Covariance

In the study of probability theory and statistics, covariance measure the joint variability of two random variables. It indicates if one of the variable's greater values corresponds to the other variable's greater values or if the lesser values of one variable correspond to the lesser values of the other. The covariance value will be positive in this situation as they vary accordingly. The other way, if the greater values of a variable correspond to the lesser values of the other variable or vice-versa, the covariance value will be negative, indicating they have an inverse relation. Finally, if the covariance value is zero, it indicates no relationship between the two variables.

For example, in our research, the covariance was used to verify if two different slices of an MRI scan had some correlation, indicating that they could carry similar information. If this was real, we could eliminate one of them to decrease the computation effort done during the training and evaluation of the neural network model used for classification.

Pandas function to calculate covariance computes the pairwise covariance among columns excluding NA/Null values. It returns a data frame containing the covariance matrix of the columns of the original data frame. A threshold can be set for the minimum number of observations for each value created. Comparisons below this threshold will be returned as NaN.This method is used for the analysis of series data in order to evaluate the relationship between different measures. Covariance is defined by the formula

$$cov\left(X,Y\right)=E\left[\left(X-E\left[X\right]\right)\right]\left(Y-E\left[Y\right]\right)$$

2.4.1 Pearson's Correlation

Covariance seeks to show if there is a linear interdependency behavior between two variables, but it is a dimensional measurement, being affected by the measurement units of the series under analysis. In order to correct this situation, the statistics term Pearson's correlation is used. It is a normalization of the covariance represented by an adimensional number ranging from -1 to $+1$. When the Pearson's correlation value is $+1$, it indicates a direct linear relationship, when one series increases its value, the other also increases; when one of them decreases, the other also decreases. For the -1 value for the correlation, we have a perfect inverse linear correlation: when one series increases its values, the other series decreases its values. As the values approach zero, there is less relationship between the variables. The correlation is defined by the formula

$$corr\left(X,Y\right)=\frac{cov\left(X,Y\right)}{\sigma_X\sigma_Y}$$

where σ_X and σ_Y are the standard deviations of X and Y.

2.5 Entropy

Entropy represents a scientific concept first introduced as a statement of the second law of thermodynamics. It is commonly associated with a state of randomness, uncertainty, or disorder. The term is applied in many scientific fields, from its roots in thermodynamics to the principles of information theory. It is used mainly in chemistry, physics, biology, cosmology, climate studies, sociology, and information systems. According to Cover, T. M. [5], Hartley, in 1930, introduced the logarithmic measurement of information for communication that was, in essence, the logarithm of the alphabet size.

2.5.1 Information Theory and Shannon Entropy

Claude Shannon was one of the most fantastic engineers, mathematicians, scientists, researchers, and inventors of the last century. His research initially focused on signal communication and cryptography but embraced early studies in artificial intelligence and machine learning. One of the most influential papers written by him was "A Mathematical Theory of Information" [25], developed during the war years, which is a founding layer for digital communications. Shannon has envisioned that communications could be seen as the same being radio, telephone, or television. All messages risk not being recovered at a destination because of noise. For him, a message was a sequence with statistical properties, and these statistics could be captured, coded, and minimized to allow effective transmission. The higher the entropy of the message, the more effort is needed to transmit it. Shannon stated that the information contained in a message be measured in "bits," digital bits. Shannon's calculations proved that the information content of a message, the number of bits, could not exceed the capacity of a given channel, the fundamental limit to that capacity. Shannon Limit is now the name of the term that names this capacity limit. Nowadays, the Information Theory went beyond the initial communications domain, and different areas such as speech recognition, artificial intelligence, information retrieval, and handwriting recognition make wide use of it.

Entropy in the communication domain is a calculation method based on Claude Shannon's work on mathematical models to abstract the transmitting and recovering of signals in a given media. The entropy of a variable indicates the level of randomness or uncertainty of the variable contents. Shannon Entropy provides an uncertainty measurement of a probability distribution. In his seminal work [25], Shannon lays down the foundations for the Mathematical Theory of Communication or simply the Information Theory, which proposes a model that represents the reproduction at a destination, the exact message initiated in its origin, or its most approximated representation. Its objective was to find a metric capable of characterizing the message without ambiguities.

Entropy is significant in the domain of data science and Artificial Intelligence. It is used to build classification trees; it is the basis of Mutual Information that is used to measure the relationship between two datasets, for example; Entropy also is the basis of Relative Entropy (The Kullback Leibler Distance) and Cross-Entropy used in dimension reduction algorithms like t-SNE and UMAP, by quantifying similarities and differences. In plain English, the result of the entropy calculation will show us how random the values in a dataset are. For example, it is associated with how surprised we will be when choosing any value from the dataset and predicting which value it will be.

We chose the Shannon Entropy calculation to evaluate the information in the MRI images used in our research. The Shannon Entropy is represented by the formula

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

, which provides us with a numerical value that indicates the diversity of the information contained in the slice image. Our intuition for this selection criterion is that the greater the entropy value, the more structures the slice image contain.

2.6 Deep Learning and Convolutional Neural Networks

Deep Learning is a sub-field of Artificial Intelligence (AI) that deals with machine learning algorithms capable of learning several levels of representations, corresponding to a hierarchy of features, factors, or concepts, where higher-level concepts are defined from lower-level ones. Deep Learning (DL) consists of representation-learning methods with multiple levels of representation, resulting from the composition of simple nonlinear modules. Each one transforms the representations at its level, starting with the raw input, into a representation at a higher level, a slightly more abstract level. With the composition of enough modules, very complex functions can be learned, such as image or audio recognition [7]. Deep Learning is also a subgroup of machine learning techniques that enables the construction of computational models composed of multiple layers capable of learning data representations with multiple levels of abstraction. Deep Learning has gained significant attention from the scientific community due to the benchmark records broken in areas such as speech and visual recognition. It can outperform conventional machine learning techniques because of its ability to learn from raw input data through consecutive filtering, linear operations, and nonlinear operations, leading to high complexity and abstraction levels.

The most common form of machine learning is supervised learning. In supervised learning, we ”teach” the model by presenting a large dataset with the concepts we want it to learn, along with the labels of the concepts, for example, images and the labels with their category. As the system is being trained, the machine receives an input image and produces an output in vectors of values, each representing a category. The training objective is that the correct category will present the highest value. If the right category does not have the highest value, an error value is produced by comparing the desired value and the value obtained; then, action will be performed in the model to adjust parameters and reduce the error. After that, the interaction will continue until the limit of interactions occurs or the desired metric of, for example, loss error, is achieved.

These adjustable parameters, often called weights, are real numbers that function as variable controls that define the input/output function of the machine. In a typical deep-learning system, there may be hundreds of millions of these adjustable weights and hundreds of millions of labeled examples to train the machine. To properly adjust the weight vector, the learning algorithm computes a gradient vector that, for each weight, indicates by what amount the error would increase or decrease if a tiny amount increased the weight value. The weight vector is then adjusted in the opposite direction.

Convolutional Neural Networks are the researcher's first option for many imaging applications dealing with classification and recognition. For a model to recognize image patterns, the following four stages [23] are involved:

- Acquisition: collect the image and adapt it to the input format needed in the following stages
- Preprocessing: this stage is responsible for tasks like noise reduction and geometric corrections
- Feature extraction: calculations named Convolutions are executed to compute and filter the fundamental attributes needed to differentiate one class of patterns from another
- Classification: this stage assigns an input pattern to one of the several pre-defined classes

The previous explanation can be visually understood in the figure 2.3, representing the process of classifying a manually written number digit. The neural network was trained with the MNIST dataset, and in each layer occurs the identification of more high-level details of the image until the final digit is classified and the model outputs a result indicating the identified digit.

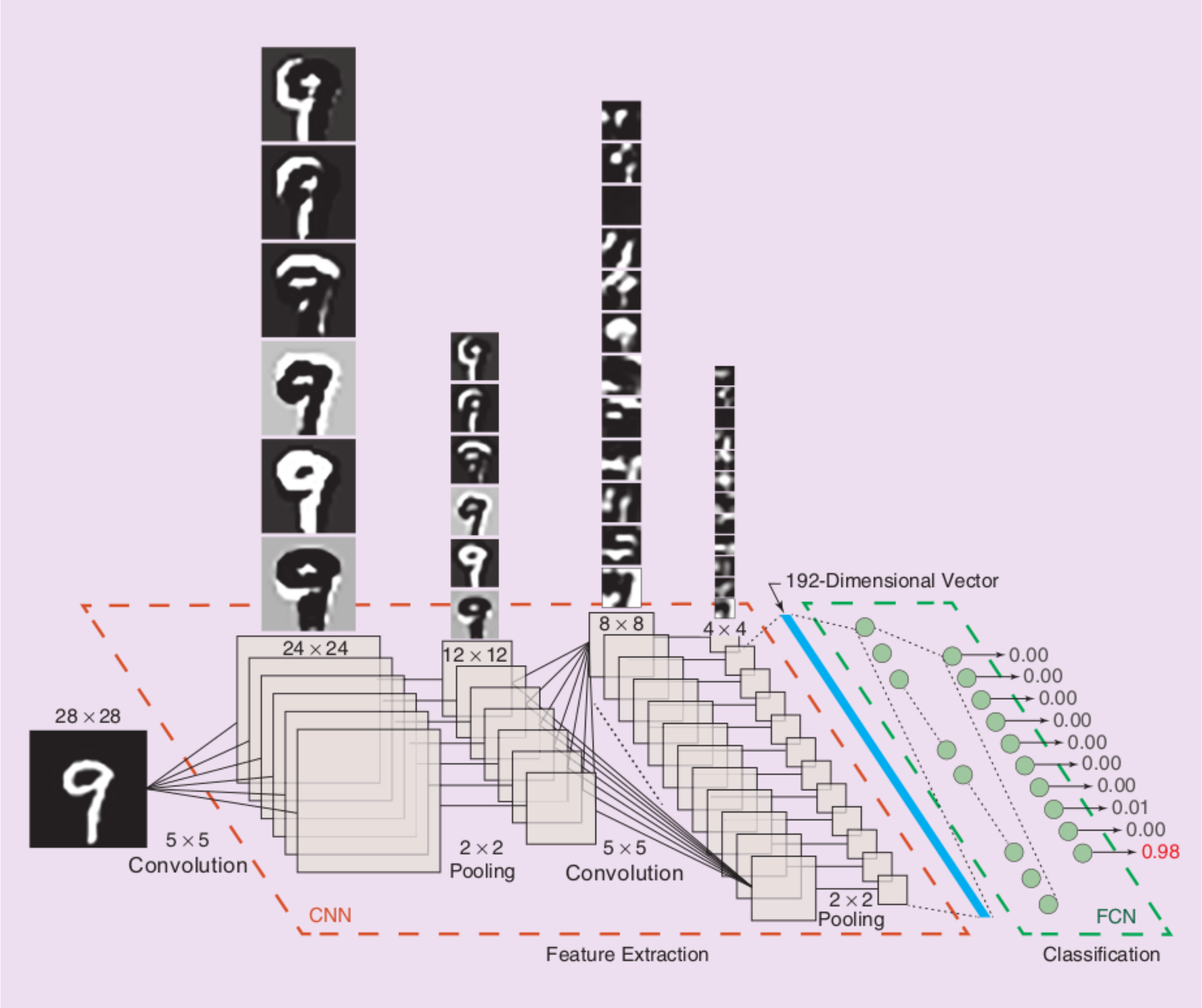


Figure 2.3. Figure shows a CNN trained to extract features that are then used by an fully connected neural network, FCN, to classify handwritten numerals. Source: [23]

For our experiments, we have used a sequential layers approach for our architecture, with one layer stacked over the predecessor, with its inputs connected to the previous layer's outputs. We have used the following types of layers and activation functions in our experiments:

2.6.1 Layers in Convolutional Neural Networks

The neural network model used in a significant part of the experiments was built using the following types of layers:

- *Convolution Layer – Conv2D:* This type of layer creates a convolution kernel that is convolved with the data applied to the inputs producing a transformed output. For example, a filter or a kernel in a conv2d layer is used as a slider over 2D input data, usually an image, performing an elementwise multiplication. The result will sum up all the results into a single output pixel. The kernel will perform this operation for every location it slides over, converting a 2D matrix into a different matrix of features.

  A discrete convolution basically is a mathematical operation on two functions h and x , that results in a third function $(h * x)$. It is also expressed as the amount of overlap of one function x when it is shifted over the function h "blending" one function with the other. The convolution of functions h and x over a finite range $\left[-k, k\right]$ is defined by the equation

  $$(h * x)[n] = \sum_{m=-k}^{k} h[m]x[n-m]$$

  A bidimensional discrete convolution is defined by the equation

  $$(H * I)[n_1, n_2] = \sum_{m_1=-k}^{k} \sum_{m_2=-k}^{k} H[m_1, m_2]I[n_1 - m_1, n_2 - m_2]$$

  In order to explain the convolution operation realized by this layer, let us suppose that we have a one-dimensional array composed of zeros and ones, and we want to detect where the values change from zero to one. We will use a kernel with two values, -1 and 1, that will be "slided" over the array, performing the convolution operation. Figure 2.4 show the steps to perform this operation:

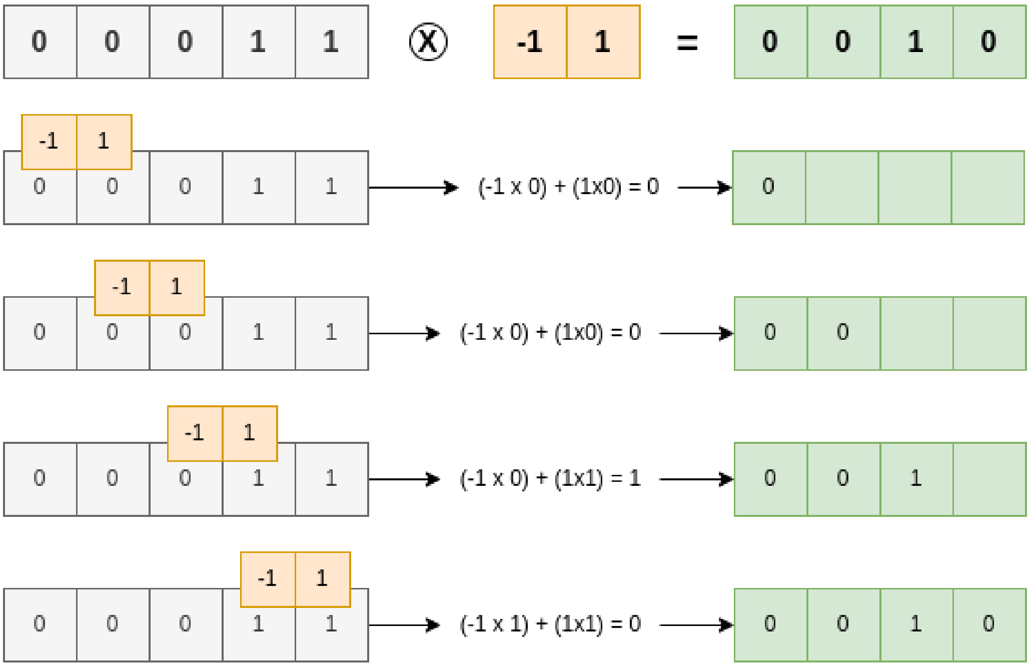


Figure 2.4. The figure shows the operation of sliding a kernel over a one-dimensional array to detect a transition from 0 to 1.

- *Max pooling Operation for 2D Spatial Data Layer – MaxPooling2D:*

  This type of layer downsamples the input, reducing its spatial dimensions, resulting in a lower resolution version of an input signal that still contains the significant or essential structural elements. A MaxPooling layer calculates the maximum value for each patch of the feature map, highlightning the most present feature and it is more informative than looking at the average presence. A pooling layer is generally applied to the output of a convolutional layer to reduce the size of each feature map.

- *Flattening Layer – Flatten:*

  This layer removes all the tensor dimensions except for one, reshaping it to have a shape equal to the number of elements contained in the tensor. This operation is equivalent to transforming into a one-dimension array. Finally, we flatten the output of the convolutional layers to create a single long feature vector output.

- *Dense Layer:* It is the regular, deeply connected neural network layer and is the most common and used Layer. The dense Layer implements the operation: *output = activation(dot(input, kernel) + bias)*, where,

  – activation is the element-wise function

  – input represents the input data

  – dot represents a dot product of all input and its corresponding weights

  – bias represent a biased value used in machine learning to optimize the model

  A dense Layer is a simple Layer of neurons in which each neuron receives input from all the neurons of the previous Layer, thus called as dense. A dense layer is used to classify images based on output from convolutional layers.

2.6.2 Activation Functions

2.6.2.1 Rectified Linear Unit (ReLU)

The REctfied Linear Unit activation function, or ReLU for short, is a function that returns the element-wise maximum value of 0 and the input tensor. It will output the input directly if it is positive; otherwise, it will output zero, as defined by $f(x) = max(0, x)$. Figure 2.5 shows the function behaviour for values of x from *-10* to *10.*

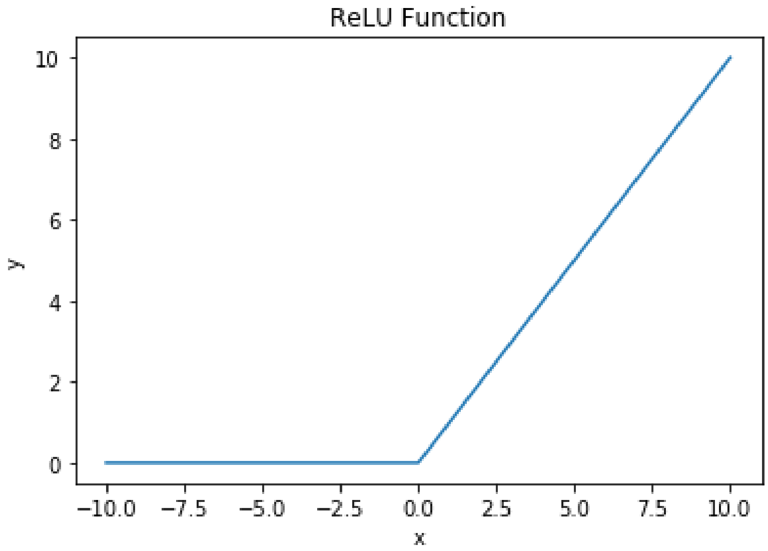


Figure 2.5. Line Plot of ReLU Activation Function for negative and positive inputs. ReLu is derivable, except on *0* value

2.6.2.2 Softmax

The Softmax function converts a vector of values to a probability distribution. The output vector elements range from 0 to 1 and sum to 1. This function often activates the last layer of a classification network because the result is interpreted as a probability distribution. The value for each class output is the probability associated with that class, and the class with the highest value indicates the probable correct answer for the classification.

2.6.2.3 Sigmoid

The sigmoid function is a mathematical function S define by

$$S(x) = \frac{1}{(1 + \mathrm{e}^{-x})}.$$

The Sigmoid Activation Function is also called a logistic function and is known for its characteristic S-shaped line-plot. It is equivalent to a Softmax function reduced to a 2 elements classification, where the second element is assumed to be zero. For small values, the function returns a value close to zero; for large values, the result of the function is close to 1. Because of this characteristics it is used for binary classifications. In addition, this function is continuous and derivable, making it a very useful function for classification. Figure 2.6 shows the function behaviour for values of x from *-10* to *7.5*.

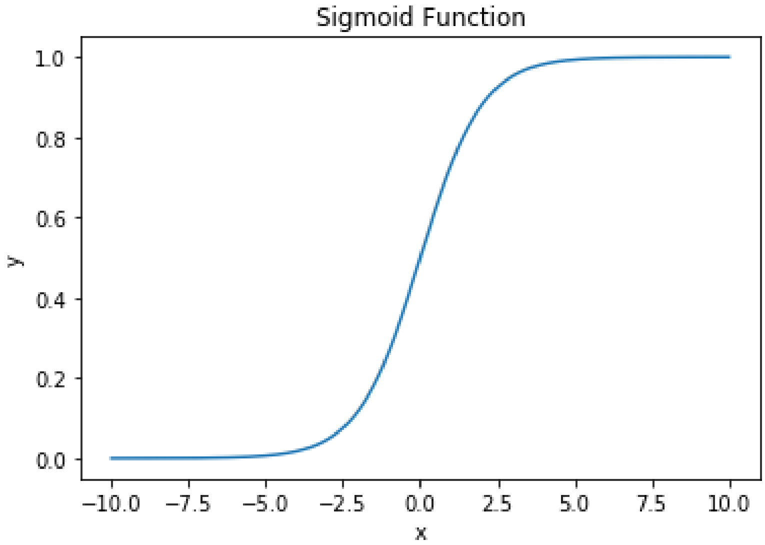


Figure 2.6. Line Plot of Sigmoid Activation Function for negative and positive inputs

2.6.3 VGG16

The VGG16 neural network model is a state-of-the-art architecture for image classification and object detection. Karen Simonyan and Andrew Zisserman developed the VGG architecture to demonstrate their research and submitted the paper as an entry to the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC).

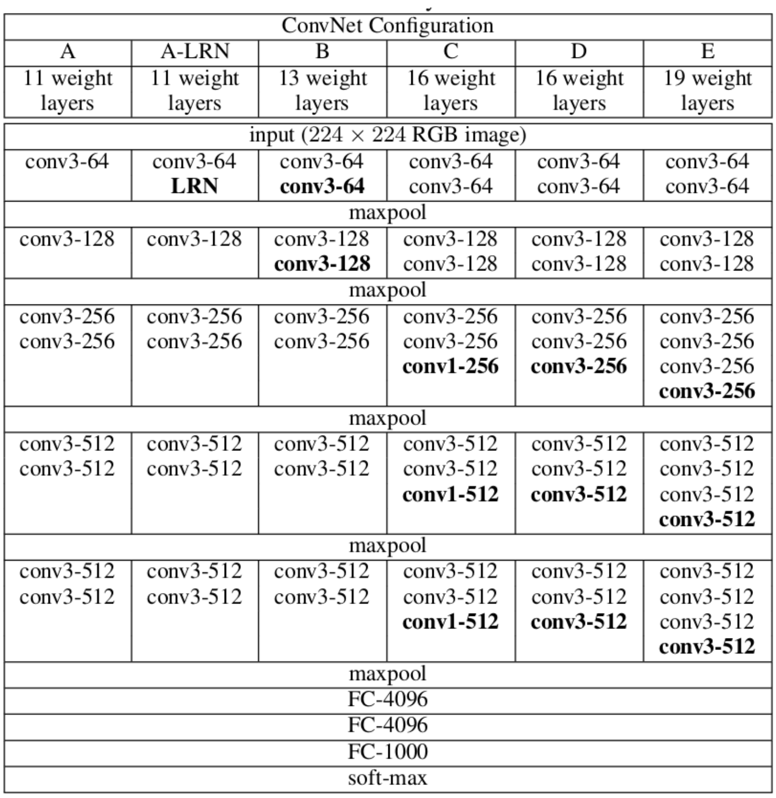
As cited in the paper [27], VGG is a significantly more accurate Convolutional Network Architecture that achieves state-of-the-art accuracy on ILSVRC classification and localization tasks. The architecture also applies to other image recognition datasets, achieving excellent performance even when used as part of simple pipelines like deep features classified by a linear SVM without fine-tuning. This architecture achieved 1st and second place in the 2014 ILSVRC challenge in detecting objects, object localization in an image coming from 200 classes, and in the task of image classification, each labeled with one of 1000 categories, an image classification task.

The VGG16 neural network model was named after the Visual Geometry Group from Oxford University, England. It is a research group inside the engineering department of Oxford University that focuses on researching the sense of vision and artificial intelligence and its impact on robotics, searching large image and video collections, production and quality control on industrial processes, and computer vision. The number 16 refers to the depth that is 16 layers deep.

The paper was submitted as an entry to the 2014 ImageNet Large Scale Visual Recognition Challenge. It is a challenge that evaluates algorithms for object detection and image classification on a large scale. In 2014 there were two competitions: a detection challenge on fully labeled data for 200 categories of objects and an image classification plus an object localization challenge with 1000 categories.

Convolutional Networks had at the time great success in the large-scale image and video recognition due to sizeable public image repositories and higher performance computing systems with GPUs or large-scale distributed clusters. As hardware was becoming more cost achievable and ConvNets became a commodity in the computer vision field, various groups were researching improvements to the architecture. The ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) played a significant role in advance of deep visual recognition architectures, serving as a testbed for large-scale image classification systems.

The VGG is an important development as it is focused on standardizing deep convolutional networks design to support deeper and better performing models This architecture achieves 92.7% test accuracy on the ImageNet dataset, which contains 14 million images belonging to 1000 classes. The first significant advance in the architecture is using many small filters, specifically 3x3 and 1x1, with a stride of one, contrary to the large filters used in other architectures like AlexNet. Max pooling layers are used after most of the convolutional layers. VGG networks use two, three, or four convolutional layers stacked together before a max pooling layer is used. This arrangement intends that stacked convolutional layers using small filters approximate the effect of one convolutional with a filter of a large size. Another critical implementation is the use of a very large number of filters. The number of filters increases according to the model depth. It starts with a significant number of 64 filters, increases to 128, 256, and 512 filters at the end of the feature extraction of the model. Figure 2.7 shows the architecture of VGG.



ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 2.7. Architecture of the VGG Convolutional Neural Network. Source: [27]

Figure 2.8 below shows a graphical representation of VGG16 architecture.

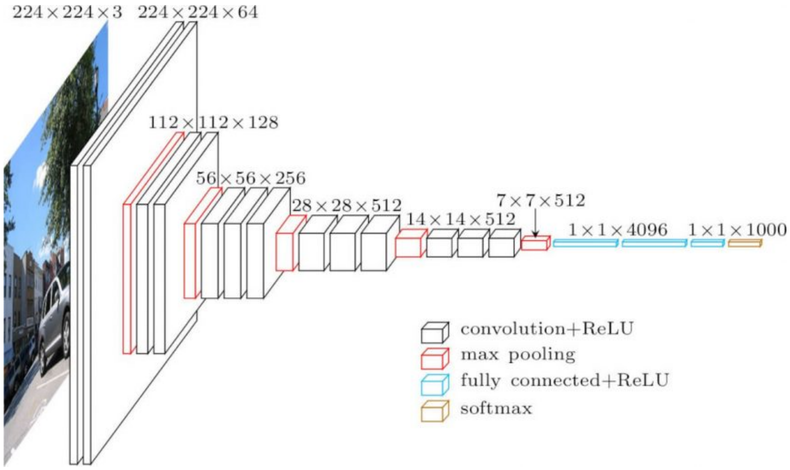


Figure 2.8. Graphical representation of VGG16 architecture. Source: [20]

2.6.4 Transfer Learning

Transfer learning is a machine learning strategy where a model developed and trained for a task has its knowledge reused as a starting point for solving a different but related problem. For example, we can use a model previously trained to classify vehicles and repurpose it to differentiate buses from trucks. A significant advantage of using pre-trained models is to diminish the time for training for our problem as the weights of the initial training, which have lasted for many hours, are reused, and only a tiny part of the model is trained with the new data.

The most common use of transfer learning is to use state-of-the-art deep-learning models pre-trained for the ImageNet classification and object detection competition like VGG, Inception, and ResNet and repurpose them to new image classification and detection tasks. These models took days to be trained and tuned, and thanks to permissive licenses from the researcher's institutions, these models and their weights can be downloaded and used freely. We took this approach in our research, where we repurposed a VGG16 neural network to classify collections of slices images in order to detect the presence of Schizophrenia.

2.7 XAI – Explainable Artificial Intelligence

Artificial Intelligence and the sub-field of Machine Learning are revolutionizing the area of Decision Support Systems. All software systems, from financial services to medical support and precision agriculture, can benefit from Artificial Intelligence, becoming more accurate, autonomous, and "intelligent" with the addition of machine learning algorithms. The addition of these technologies enables the automation of several repetitive tasks reducing of needed time to process and reproduce decision-making processes done by human operators. For example, Healthcare Area is experimenting with a revolution in machine learning supported diagnosis, with an amplified visualization of characteristics that were hidden, beyond human perception, leading to more precise diagnosis with accuracy similar to human decisions or superior. Although these advances are astonishing, more sophisticated, and accurate, machine learning models compare to black boxes, where little is known of its internal process of classification or prediction of results.

This situation has motivated government agencies, corporations, and universities to fund researchers to study ways to guarantee that machine learning algorithm's output results are reliable and also explainable. As a result, several libraries and methodologies were developed to explain and justify the outputs obtained from a machine learning engine.

One of the first initiatives towards obtaining explanations from Artificial Intelligence based systems was the Explainable Artificial Intelligence (XAI) program, launched by the Defense Advanced Research Projects Agency (DARPA) of the United States Department of Defense in May 2017 [10]. The program coined the acronym XAI, with X for Explainable, with the explicit intention to create human-understandable AI systems through practical explanations rather than interpretable, comprehensible, or transparent AI. The program's main objective was to create a collection of ML techniques to produce explainable models that, combined with explanation techniques, would enable users to understand when a model works correctly, when it fails, trust its results and how to improve its performance.

XAI uses two main approaches to base its techniques and methods. The first approach is to interpret and justify a prediction from a non-interpretable ML or DL model with input data. This approach is known as a Post-Hoc one, as the explanation is obtained after the prediction. The second approach is to build a naturally explainable model capable of explaining its results since its conception, known as the Ante-Hoc approach. Post-hoc approaches are valuable tools when accessing DL or ML models that can be viewed as black boxes as the number of layers, weights, and parameters increases. The Post-hoc approach is agnostic concerning the model it is explaining. It is not interested in its inner functions but in creating a proxy model with the same outputs with an explainable result.

XAI brings benefits to three types of DL users [9] as follows:

- Model Developers and Builders: These individuals' primary job is to develop, experiment with, and deploy deep neural networks. They strongly understand DL techniques and have a well-developed intuition surrounding model building. They can decide on key issues, such as identifying what models perform best on which types of data.

- Model Users: This group of users have some technical background but are neural network novices. They use well-known neural network architectures to develop domain-specific applications, small-scale training of models, and downloading pretrained model weights to use as a starting point.

- Model Novices: The third group typically has no prior knowledge about DL and may not have a technical background. They simply use AI-powered devices and applications.

- Physicians: We added this fourth group based on the domain of this work, as they will use the DL model results, associated with the explanation provided by the XAI results, to support their diagnosis.

XAI also gains importance as more laws and regulations are created to mediate conflicts that model results can cause when applied to daily aspects of human life. The Brazilian Chamber of Deputies approved the 13.709 Law on August 14th, 2018, known as *Lei Geral de Proteção de Dados Pessoais (LGPD)*, General Law for Personal Data Protection contains items that deal directly with the *Right of Explaining.* It's 20th article states that the owner of some personal data has the right to review decisions exclusively generated from an automated process that uses his data exclusively as input to make professional profiles, financial profiles, and consumption profiles and affects his interests. The European GPDR inspired Brazilian law, the European Union General Data Protection Regulation, effective from May 25th, 2018. GPDR regulates personal data protection and privacy in the European Union region and states at the 71st item that an explanation must be provided to an individual for decisions and results obtained by an automated process.

2.7.1 Algorithms and Libraries for XAI

The most used strategy to obtain an explanation for a model behavior is to use an explainable model like decision trees, rules, additive models, attention-based networks, or sparse linear models. This kind of model offers the possibility of inspecting the paths followed to achieve a result, but with DL models, this is not so easy, as they are composed of several layers with lots of weights and parameters to adjust. In order to explain the ML and DL models, several approaches and algorithms have been created. We will describe two of the more used ones.

2.7.1.1 LIME

With the increasing adoption of Deep Neural Networks in real-world applications, DL users achieve greater accuracy as they perfect the DL networks for the task at hand. Consequently, they obtain results that are far better than those that a naturally interpretable model could provide, leading to a situation where explainability is underrated. In this case, providing explainability would mean damaging the model, lowering the number of feature data to be analyzed, and degrading the model results. LIME comes to aid in providing this explainability without modifying the model under analysis.

Local Interpretable Model-agnostic Explanation [31] LIME proposes to obtain an explanation by treating the model under inspection as a black box. This explanation technique explains the predictions of any AI classifier by learning an interpretable model locally around the prediction.

LIME's objective is to identify an interpretable model over the interpretable representation that is locally faithful to the classifier. Although an interpretable model may not approximate the black-box model globally, it can approximate it for an individual instance. The black-box model under the explaining approach will receive data at its inputs, and the model responses will be learned. Also, data with perturbations will be submitted to see how the black-box model reacts to this noise, and the set of responses will be used to infer/approximate the model behavior. This strategy is model-agnostic as it does not need to know the model's internal components or how the information flows inside it.

An example of an output that can be obtained with the usage of LIME is shown below in figure 2.9. *(a)* shows the original image of a husky dog that a model misclassified as a wolf. In *(b)*, LIME helped explain what was wrong with the classification job as none of the dog's characteristics was recognized. It was the background that the model took into consideration to predict a wolf.

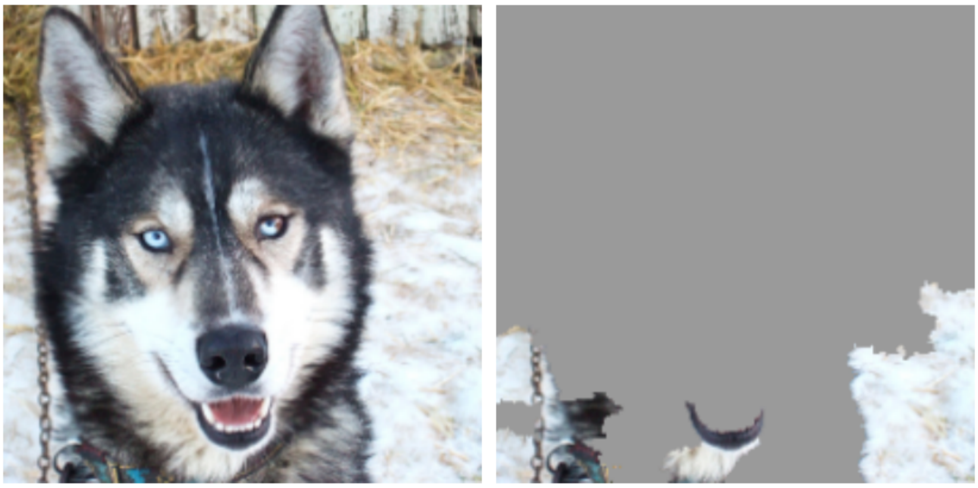


Figure 2.9. (a) Original image of a husky dog (b) Areas that explain misclassification as a wolf. Source: [31]

2.7.1.2 TF-Explain GradCAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [24] is a class-discriminative localization technique for making CNN-based models transparent by producing visual explanations for the model predictions. This technique aims to improve the interpretability of CNN models, explaining why they predicted what they predicted. The Grad-CAM technique uses the gradients of any target concept, flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. This technique can be applied for tasks like image classification, image captioning, and visual question answering (VQA)and help identify biases in datasets.

To explore the interpretability of the models with Grad-CAM, we have used the *tf-explain* library that offers the implementation of models constructed with Tensorflow. It offers an explanation method that outputs the explanation, as a heat-map, for example. This explanation can be obtained from a trained model loaded with the core API of Tensorflow and saved to a disk or at training time with callback functions integrated with Tensorboard. Tensorboard is a Tensorflow module that provides the visualization and tooling for analyzing a machine learning model experimentation. With it, we can validate what our network model looks at when making the predictions and if the relevant input patterns activate it. An example of the Grad-CAM model output visual explanation can be seen in figure 2.10, where *(a)* shows the original image is explained. *(b)* Moreover, *(c)* shows the relevant areas colored in red that identify respectively a *Cat* and a *Dog*.

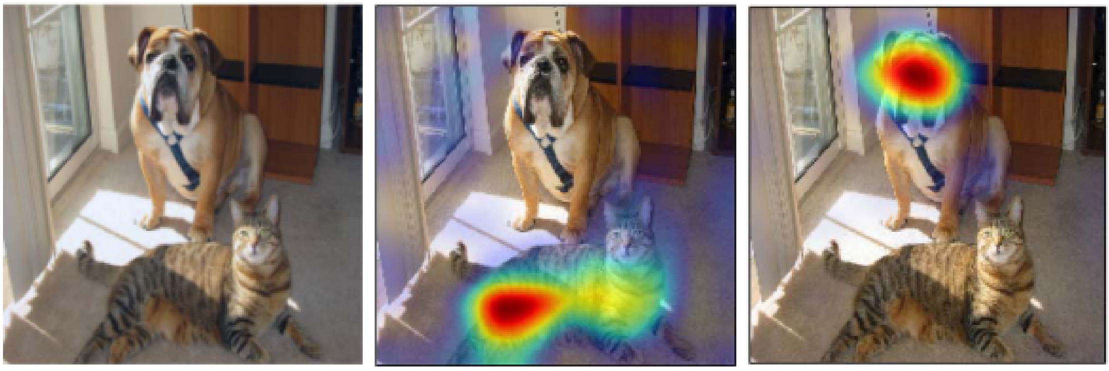


Figure 2.10. (a) Original image of a cat and a dog. (b) Area in red shows what identifies a cat. (c) Area in red shows what identifies a dog. Source: [31]

3 Proposed Methods for Schizophrenia Detection Based on Multi-Slice Anatomical Images

The methods described in this chapter were designed to support the general objective of systematically evaluating the different combinations of anatomical MRI scan axial slices for the problem of classification using Convolutional Neural Networks to detect the presence of schizophrenia structural modifications or not.

In order to achieve this objective, we executed methods to achieve the following specific objectives:

- Experimenting with entropy and Pearson's correlation to identify the most informative slices that would enhance our classifier's results.
- Experimenting with various combinations of slices from MRI scans instead of a single central slice, aiming to obtain the best performance metrics from the classification model.
- Evaluate the performance of a small Convolutional Neural Network compared with a pre-trained state-of-the-art neural network architecture for image classification.
- Evaluate techniques of Explainable Artificial Intelligence applied to schizophrenia classification.

Our experiments consisted of:

- Preprocessing the MRI images raw data of schizophrenia (SZ) patients and healthy controls (HCs) subjects, extracting two-dimensional images, like slices, along the axial direction, used to compose an image dataset to conduct the subsequent experiments.
- Using the Shannon Entropy, we evaluated the individual slices and combination of slices, searching for the combination which could bring the best results for the classification machine learning engine

- Using the Covariance Estimation, we evaluated if the individual slices and combination of slices could be correlated or have any dependence on each other, indicating some information redundancy that would be better to exclude.
- Present a selected mix of 2D images of slices to a Convolutional Neural Network for training and evaluation. The training and evaluation were repeated for several individual MRI scan slices and a combination of slices, each combination or individual slice training being submitted through a k-fold evaluation strategy.
- Based on the previous experiment, identify the collection of slices that have the better performance metrics.
- Submit the dataset of selected slices to the simple CNN architecture model for evaluating the results
- Submit the dataset of selected slices to a pre-trained state-of-the-art neural network architecture for image classification.
- Compare the results for the two previous experiments.
- Apply XAI methods for verifying what is being taken into consideration by the machine learning engine to classify the slices images.

Tables showing the result data from these evaluations are in the results section.

3.1 Data Description

In order to evaluate the proposed methods, we used the MRI dataset from SZ patients and HC subjects from BIRN [30]. This dataset initially appeared in [13]. Each patient or subject file downloaded from the BIRN website contains functional and structural MRI scans.

The used dataset is a collection of scanned images resulting from the process of MRI scanning. Their type is NIfTI, which stands for Neuroimaging Informatics Technology Initiative [12], a data format created for storing MRI, Functional Magnetic Resonance Imaging (fMRI), and other medical images. NIfTI format is an adaptation of the Analyze™ 7.5 format, also used for medical images, developed by Biomedical Imaging Resource (BIR) at Mayo Clinic. NIfTI and Analyze 7.5 are compatible as NIfTI only adds more fields to clarify the orientation in the space of the images, such as which side is left or right.

The NIfTI data format comprises three files:

- A header file for storing meta-data with extension *.hdr*.
- The actual data file with extension *.img*.
- A file with extension *.nii* with the contents of the two previous files for easy data processing.

In order to explore MRI Scanning files, we used the python library *nilearn*, which specializes in the treatment of NIfTI files, including the plotting of images. For extracting the image slices, we used the python library *NiBabel* [17]. This library provides access to the most common neuroimaging file formats, such as ANALYZE (plain, SPM99, SPM2, and later), GIFTI, NIfTI1, NIfTI2, CIFTI-2, MINC1, MINC2, AFNI BRIK/HEAD, MGH, and ECAT as well as Philips PAR/REC and limited access to DICOM file format. In addition, it provides access to meta-data in the header file and the image data, viewed as NumPy arrays.

Our experiment used the structural scan file that contains axial, coronal, and sagittal slices from the patient/subject of examination. Each participant file represents the brain as a 3D image with 252 x 252 x 27 voxels. From this file containing the scanned brain volume, we extracted the 27 axial slices with 252 x 252 pixels as input to the neural network. We used 79 scans of control subjects and 79 scans of diagnosed schizophrenic patients, adding up to 158 scans.

Each MRI scan is a volume composed of images that can be viewed and indexed in three different types of orientations of the human head:

- Axial plane: it is an X-Y from top to down
- Coronal plane: it is an X-Z plane, with images from front to back
- Sagittal plane: it is a Y-Z plane from the left side to the right side

Figure 3.1 shows the three axis of a MRI scanned brain.

The *NiBabel* library enables us to access the scanned volume meta-data information present in the header.

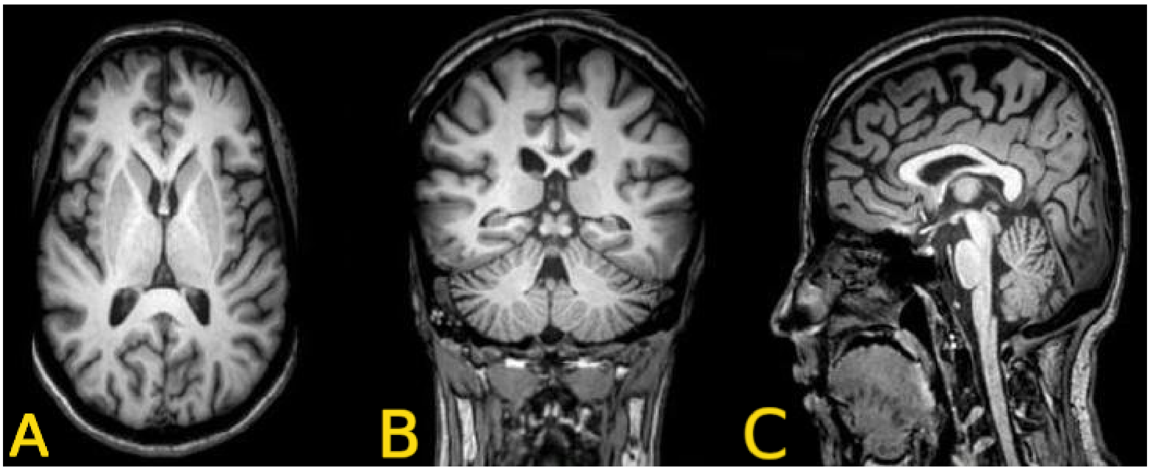


Figure 3.1. Figure showing the axis planes for a MRI scanned brain volume (A) Axial, (B) Coronal, (C) Sagittal. Source: [28]

3.2 Neural Network Architecture

The network was built using the Tensorflow library [1] with Keras API [4] as a high-level abstraction. The coding was done in Python 3.7, using Jupyter Notebooks as the running environment and Anaconda for Python environment management. The Neural Network model was defined as Sequential, as the layers were stacked over each other, data flows through the inputs, is processed in the hidden layers, and the output layer presents the results.

Figure 3.2, shows a simplified view of the architecture:

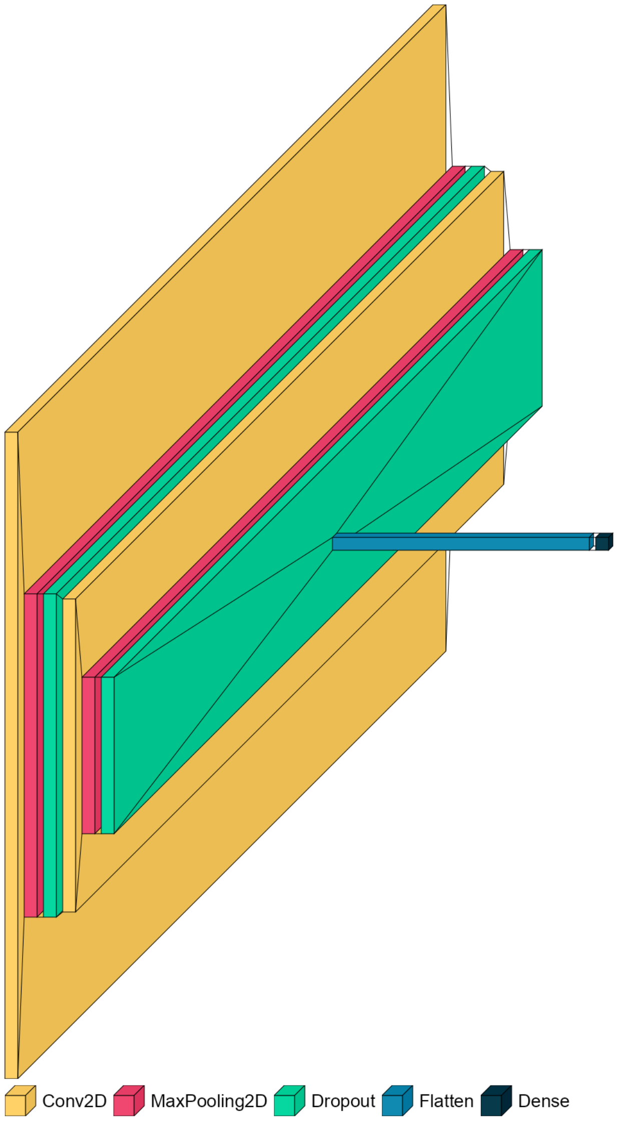


Figure 3.2. Simplified architectural view of the Neural Network Model used in the experiments.

- The first layer (FL) is defined as a Conv2D type, which creates a convolutional kernel convolved with the inputs to produce a tensor of outputs. The FL has 32 neurons and takes inputs representing 252 x 252 x 1 tensor, according to the slice dimensions. Moreover, the FL uses 5x5 kernels and a rectified linear unit (RELU) activation function.

- The second layer is the MaxPooling2D type that downsamples the input by taking the maximum value over a window of 2 x 2 cells.

- The third layer is another Conv2D type, applying another convolution with a kernel of 5 x 5 cells

- The fourth layer is another MaxPooling2D 2 x 2 cells.

- The fifth layer is a Flatten type. It reshapes the tensor to have the shape equal to the number of elements needed for the last layer.

- The last layer is Dense, with 2 neurons and "softmax" activation function, responsible for presenting the probabilities of each possible classification

Table 3.1 shows the network architecture summary.

Table 3.1. Used Neural Network Architecture

Model: *sequential*

Layer (type)	Output Shape	Parameter value
Bidimendional convolutional	$252 \times 252 \times 32$	832
Max Pooling	$126 \times 126 \times 32$	0
Bidimendional convolutional	$122 \times 122 \times 64$	51,264
Max Pooling	$61 \times 61 \times 64$	0
Flatten	238,144	0
Dense	2	476,290

Total number of parameters: 528,386
Trainable parameters: 528,386
Non-trainable parameters: 0

3.3 Extraction of Image Slices

In order to create a dataset for training the model, we had to obtain the images of individual slices of the scanned brain volume. The library *nibabel* provides a function to obtain a *NumPy* array from the volume and extract the images. The extraction of an individual slice was executed as follows:

1. After downloading the collection of scanned volumes from the BIRN site, two directories will separate the files of control subjects from Schizophrenia patients. Two

files represents each MRI scan , they have the same filename, one with the extension *hdr* containing meta-data information and the other with the extension *img* containing the scanned data.

2. The volume image is read from the disk using the *nibabel* library. The volume is read into memory as a *NumPy* array with dimensions 256 x 256 x 27.

3. Our primary interest is in the images in the Axial plane, as this axis is where the anatomical modifications of brain structures are most noticed in schizophrenia patients. In order to access these images, we access each slice from the volume by accessing the third index of the *NumPy* array. For example, to access the second slice image, we use the following pseudo-code:

   ```
   read a file containing MRI Scan 3D volume to a numeric array
   extract two dimensions sub-array, a slice,  from 3D volume using index 2
   save the slice to disk as an image with gray shades color map
   ```

   Figure 3.3 shows slice 13, slice 14 and *(c)* shows slice 15 from a subject extracted from MRI scans. Figure 3.4 shows the same index slices from a second subject MRI scan.

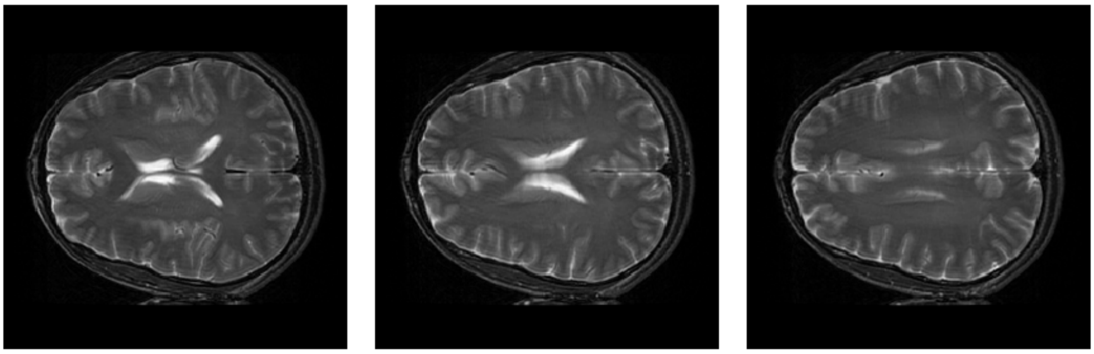


Figure 3.3. MRI Scan slices of a subject (a) Slice 13 (b) Slice 14 (c) Slice 15

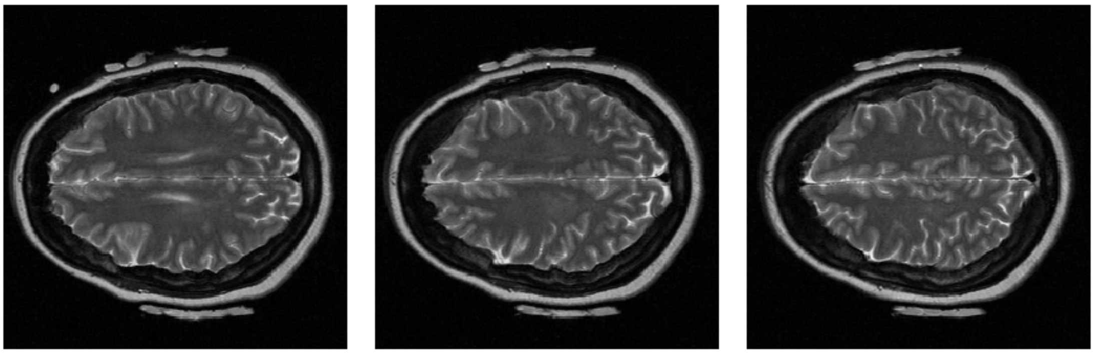


Figure 3.4. MRI Scan slices of a second subject (a) Slice 13 (b) Slice 14 (c) Slice 15

3.3.1 Data Augmentation

Our dataset is composed of slices extracted from 158 MRI scans, and this is a small number of images. A small dataset can lead to poor metrics values associated with the validation dataset, inducing evaluation errors. Data augmentation techniques are the commonly used solution to this situation. Data augmentation increases the dataset size by creating new images from existing images and introducing noise or transformations like rotations, color inversions, and position shifts. The intuition is to create new images similar to real ones that can be presented to the model when deployed in production. In our dataset, we observed that some images have rotations with angles varying from -10 to 10 degrees, as shown in figure 3.5, probably caused by variations of subject positioning in the MRI Scanner. Therefore, we programmed a function to create 20 new images from each image extracted from the scanned volume by rotating 1 degree from -10 degrees to 10 degrees for each original image. Before executing the data augmentation function, we put part of the dataset aside to create the validation dataset.

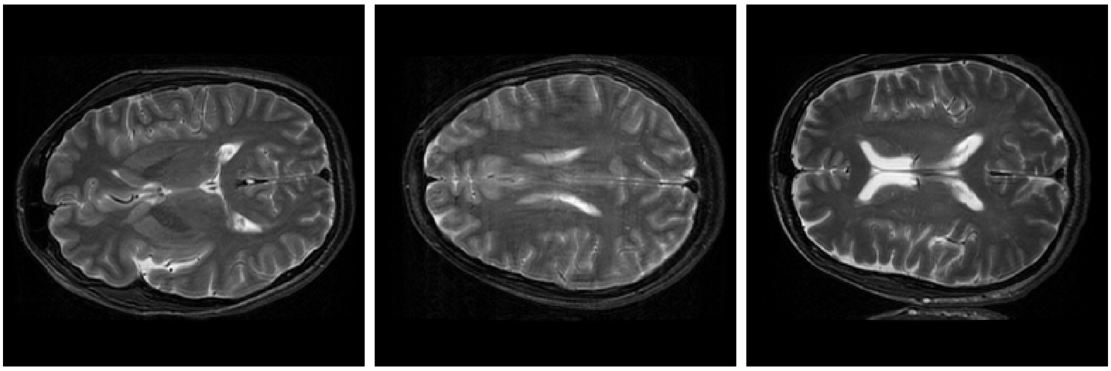


Figure 3.5. Images of slices showing rotations due to positioning in MRI Scanner. (a) Rotated down image. (b) Slightly rotated up the image. (c) Rotated up the image.

3.3.2 Analysing a 3D volume of images as a photo film reel

Since the initial stages of our research, we had the vision that we needed to create a dataset composed of various slice images, but each sample would be composed of different slice images of the same subject/patient. Investigating possible input layers to use in our neural network, we identified two main possibilities: use a first 3D dimensional layer with the complexities and extended training and validation time, or convert the sequence of images of the same subject/patient to an analogy of a photo film roll. We choose the second option as it would provide the simplified possibility of treating the ”photo film reel” as a two-dimensional image composed of several slice images collated.

The following algorithm executed the collation of images:

1. Obtain an MRI scan from a subject/patient.
2. Creates a zeroed numeric array with the size of 252 pixels of height and 27 slices x 252 pixels of width.
3. Interacting extracting the slices of the MRI volume from index zero to 26, in the axial direction.
4. For each slice, make a copy of the slice to a zeroed array, inserting it at the width position 252 multiplied by the index.
5. Write the numeric array as an image file to disk.

Figure 3.6 shows the example image resulting from the collation of slice images.

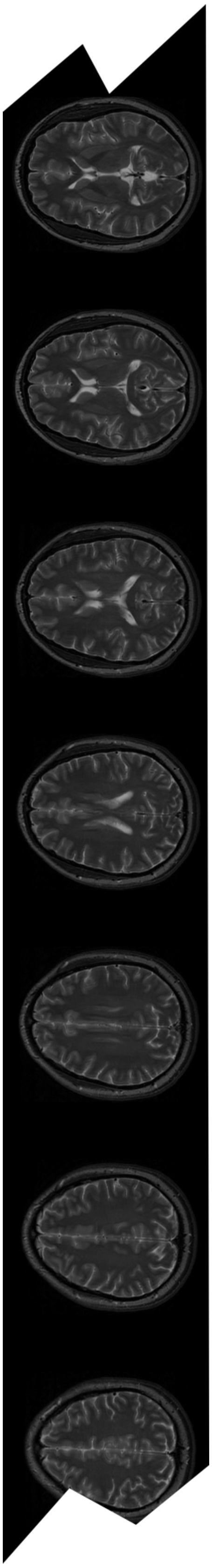


Figure 3.6. Example of slices extracted from an MRI Scanned 3D brain converted to 2D *photo film reel* of slices.

3.4 Folder organization for storing the images

The slices' evaluation starts with preparing the datasets for the training and validation of the machine learning models. We prepared the datasets by extracting the slices images, transforming them into images like photo film reels, and organizing them on the hard disk in folders as shown in table 3.2

Table 3.2. Table showing the folders organization for training and validation of the machine learning model. For each slice collection, we have an organization like this one.

Folder	Contents
/data/train/slicescollection/schizo	Images of one collection of slices from the Esquizophrenia patient group to be used for training
/data/train/slicescollection/control	Images of one collection of slices from the healthy control group to be used for training
/data/val/slicescollection/schizo	Images of one collection of slices from the Esquizophrenia patient group to be used for validation
/data/val/slicescollection/control	Images of one collection of slices from the healthy control group to be used for validation

The organization of the collections was executed as follows:

- First, we have created folders following the organization of the table 3.2: one folder for the training images and another for the validation images. Below each one, we have a subfolder named with a slices collection identifier, where we created the two folders named *control* and *schizo* that will contain the images for each of the classes. The images of each subject are organized as photo film reels.

- The division of the dataset in two parts were 70% for *training* and 30% for *validation.*

- We used a python library named *split-folder* and a custom shell script to execute the distribution of the images in the correct folders according to the group, slice collection, training, or validation dataset.

3.5 Evaluation of Network Models

We evaluated the neural network models with the following steps:

1. Evaluate the best combination of batch size and epochs to train the model. For batch size, we used the values 10,15,20, and 25. For epochs, the values were 5,10,20, and 30.

2. Evaluate the best optimizer for the model using the algorithms "Stochastic Gradient Descent (SGD)," RMSprop," Adagrad," Adadelta," Adam," Adamax," and "Nadam."

3. Use the best combination of batch size, epochs, and optimizer, determined in previous steps, to create the model.

4. Evaluate model mean performance using a k-fold strategy. This strategy divides the data set into k parts, trains the model with k-1 parts, and tests with the remaining part. In this k-fold strategy, we used $k = 10$, as it is one of the recommendations in [14], and using larger values of k did not change our results in the preliminary tests.

We repeated the steps for the selected individual slices and slice combinations.

3.5.1 Evaluation of Model's Performance Metrics

The evaluation of the models training and predicted performance uses four terms:

- *True Positive (TP):* Observation and prediction are positive.
- *False Positive (FP):* Observation is negative, but the prediction is positive.
- *True Negative (TN):* Observation and prediction are negative.
- *False Negative (FP):* Observation is positive but is predicted negative.

The four terms are used to formulate the following metrics:

- Loss: The purpose of loss functions is to compute the quantity a model should seek to minimize during training. As we classify an MRI scan as probable SZ patients or HC subjects, it is a binary classifier, and we use the binary cross-entropy loss function to compute cross-entropy between the labels and predictions. This metric is defined by

$$Loss = \frac{1}{N}\sum_{i=1}^{N} -(y_i.\log(p_i) + (1 - y_i).\log(1 - p_i)).$$

- Accuracy: Calculates how often predictions equal labels. This metric creates two local variables, total and counts, that compute the frequency with which predicted labels match true labels. This frequency is ultimately returned as binary accuracy: an idempotent operation that simply divides the total by count. Function A defines this metric as
$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)}.$$

- Precision: Computes the precision of the predictions concerning the labels. The metric creates two local variables, true positives and false positives , to compute the precision. This value is ultimately returned as precision, an idempotent operation that divides true positives by the sum of true positives and false positives. This metric is defined by
$$P = \frac{TP}{(TP + FP)}.$$

- Sensitivity: This metric measures the proportion of actual positives that are correctly identified as such. This metric is defined by
$$Sy = \frac{TP}{(TP + FN)}.$$

- Specificity measures the proportion of actual negatives that are correctly identified as such. It is defined by
$$Sp = \frac{TN}{(TN + FP)}.$$

- AUC: A Riemann sum computes the approximate AUC (Area under the curve). This metric creates four local variables, true positives, true negatives, false positives, and false negatives, that are used to compute the AUC. A linearly spaced set of thresholds is used to discretize the AUC curve to compute pairs of recall and precision values. For example, the area under the ROC curve is computed using the height of the recall values by the false positive rate, while the area under the PR curve is computed using the height of the precision values by the recall.

- Recall: Computes the recall of the predictions with respect to the labels. This metric creates two local variables, true positives and false negatives used to compute the recall. This value is ultimately returned as recall, an idempotent operation that divides true positives by the sum of true positives and false negatives. The metric is defined by
$$Rcl = \frac{TP}{(TP + FN)}.$$

3.6 Evaluation of the correlation of information between slices using Covariance Estimation

Observing the slices images, we noted that some slices seemed to have a repetition of structures with the same format or pixel levels or with small variations, making us think of the possibility of existing repeated information from one slice to another or some correlation. If this hypothesis were valid, that would be a possibility of excluding some slices as they only would add computational effort to the machine learning engine and not present characteristics relevant to the machine learning model, capable of helping to discriminate healthy control subjects from schizophrenic patients. As a method for evaluating this possibility, we used the Covariance Estimation for the slice images, comparing the images between each other.

This procedure was executed for the two groups of subjects, with individual results for each group. It was executed as follows:

1. From previously extracted slice images from different subjects, we extract the central portion of the slice image, size of 48 pixels, same slice of every participant of the group and combined all of them in an array
2. The previous procedure was repeated for every slice of the scanned MRI volumes in the axial direction
3. All the values were combined in a matrix to be submitted to the covariance evaluation algorithm
4. After calculating the covariance of the slices, a graph is built to provide a visual representation of the results

3.7 Evaluation of best slices for dataset creation using Shannon Entropy

In order to evaluate which 2D slice images would be relevant to use in our machine learning classifier, we choose to use the Shannon Entropy value to provide an insight into the overall information contained in the image. The Shannon entropy, represented by the formula

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

provides a numerical value that indicates the diversity of the information contained in the slice image.

Our intuition to use this selection criterion was that the greater the entropy value, the more diverse information was present, indicating that more structure representations would be present in the slice image. This would indicate that a specific slice is a better representative to differentiate the control group from schizophrenia subjects. These slices, in our intuition, would be the ones with better results when submitted to the CNN model training and evaluation.

The Shannon Entropy value of the image slice was obtained using the function *shannon-entropy* from the library *skimage.measure*. It defines the equation of Shannon entropy internally as

$$S = -\sum(pk.\log(pk))$$

, where *pk* are frequency/probabilities of pixel of value *k*. Its inputs are the grayscale image of the slice and the logarithmic base value to use, which usually equals 2. The output value of the function is measured in bits or Shannon values for base=2.

Our method for this experiment was as follows:

First part: Calculate the entropy for the individual slices datasets

1. Extract the individual slices of the scanned MRI volume, separated into folders for schizo and control, and under these folders, the sub-folders with the individual slices
2. Calculate the average entropy for each slice dataset in the control folder
3. Calculate the average entropy for each slice dataset in the schizo folder
4. Calculate the average entropy for each slice dataset for the aggregated groups control and schizo
5. Create a table with the results and generate the graphs for the results
6. Create a sorted list of the slices ordered by the calculated average entropy

Second part: Identify the best collection of slices to use for classifying the subjects based on the average entropy of the individual slices

1. Using the classified order of individual slices by their entropy, create a dataset with slices of the first slice whit the greatest entropy from control and schizophrenia patients with respective labels
2. Split this dataset into training test and validation
3. Submit the split dataset to model training and test

4. Using the trained model, submit the validation split dataset to the model for evaluation and collect the metrics of accuracy and loss error

5. Add the next slice from the ordered list of slices by accuracy and repeat the steps from step 1 until all the ordered list has been iterated.

6. Collect the data for tabulation and analysis

Third part: Present the collected data

1. Using the classified order of individual slices by their entropy, create a dataset with slices of the first slice whit the greatest entropy from control and schizophrenia patients with respective labels

2. Split this dataset into training test and validation

3. Submit the split dataset to model training and test

4. With the trained model, submit the validation split dataset to the model for evaluation and collect the metrics of accuracy and loss error

5. Add the next slice from the ordered list of slices by accuracy and repeat the steps from step 1

6. If the new slice does not enhance the previous results, then exclude these slices from the list of best slices to use for classification

7. Repeat until all the ordered list has been iterated.

8. Collect the data for tabulation and analysis

3.8 Evaluation of top accuracy slices collection by a state-of-the-art Deep Learning architecture

After evaluating the entropy of the slices, collection of slices, and submitting them to a small-scale neural network used as a baseline, we determined a rank of collections ordered by accuracy. With this list in hand, we submitted the top 2 collections to a state-of-the-art architecture, the VGG16, to evaluate the results obtained with a special collection of slices. We used the transfer learning technique; with this approach, we took advantage of the previous training in the architecture in the ImageNet dataset. The procedures for this evaluation are as follows:

1. Create a dataset with the collection of the slices that obtained the best accuracy, separated in training and validation

2. Create a VGG16 Neural Network model using a template available in the repository for the Tensorflow Keras library, modifying its input to adapt to our images and adding a last Dense layer with its output adapted to our number of classes. The weights used to come from the ImageNet dataset training.

3. Configure the neural network to enable the training of only the last layer of the architecture

4. Do the training, with a configuration to save the models that have the smaller loss metric

5. Collect the history of the training

6. With the resulted data from the training create graphs of loss and accuracy versus training epoch for evaluation of training performance

7. Repeat all the steps for the second best accuracy slice collection

3.9 Applying XAI method GradCam

Our last experiment was to apply an Explainable Artificial Intelligence method to the output of our state-of-the-art model, based on VGG16 architecture. The procedure objective was to have an insight into what the model is looking at when it makes a prediction. This visualization will help verify if the prediction makes sense and could lead to new conclusions about what structures are relevant when diagnosing Schizophrenia. We experimented with the best models for a dataset composed of only slice six and for a dataset composed of slices 0 until 16. we wrote custom scripts to obtain the heatmap and the superimposed image to visualize the relevant areas. To execute this experiment, we took the following steps:

1. Load a random image from the dataset with only slice 6 for evaluation

2. Preprocess the image to adapt for the model input layer

3. Load trained model, based on custom VGG16 architecture

4. Generate a heatmap for the image based on the model's gradients for the last convolutional layer

5. Generate a superimposed image of the original image and the heatmap

6. Display the result

7. Repeat the procedures for an image from the dataset with images from 0 to 16

4 Results and Discussions

In order to help locate the slices of MRI brain scan referenced in experiments results, we will show below the images with respective indexes in an MRI volume axial scan:

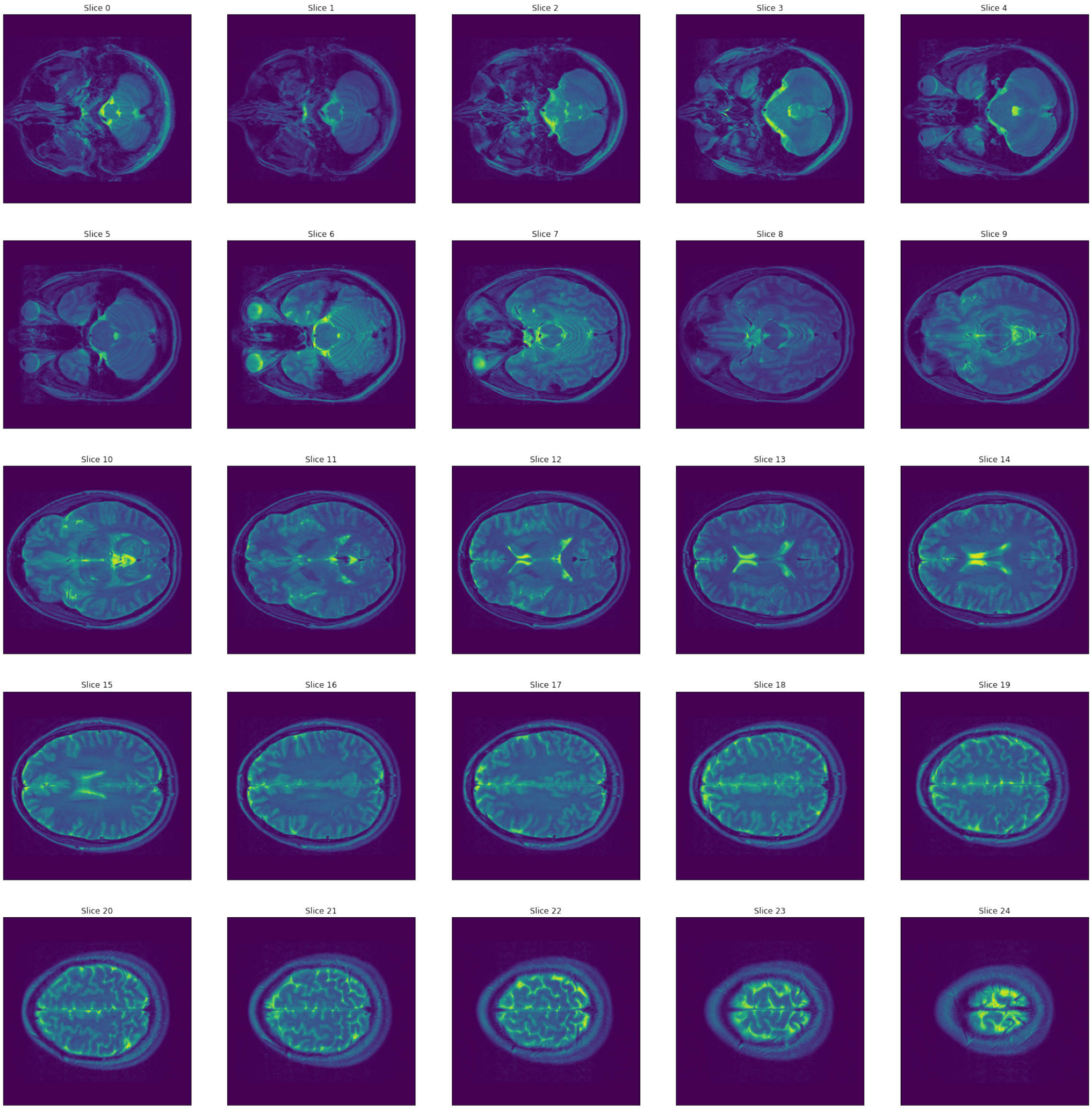


Figure 4.1. Examples extracted slices from MRI Scan, from index 0 to 24 from a total of 27 slices.

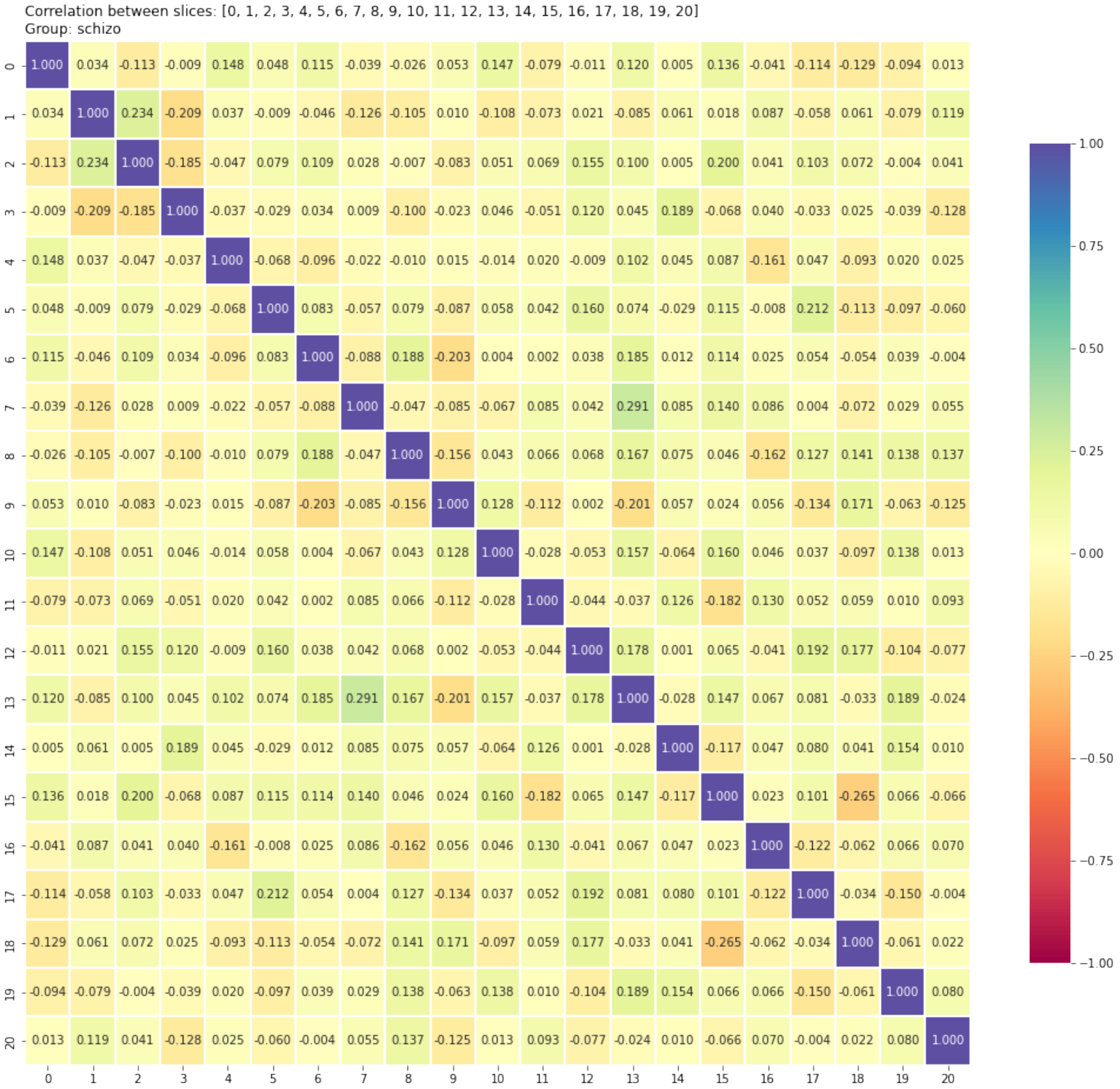
4.1 Evaluating anatomical MRI scan brain slices to obtain the most representative dataset for training

In previous papers we observed that the researchers had a trend of analyzing only central axial slices of the brain, from 9 to 16, in the majority of the cases selecting only one of these slices to analyze. Our intuition was that other relevant slices containing useful information would be used in a classifier. In the other direction, there was a possibility that different slices could have duplicated information or correlated information, irrelevant if duplicated in the dataset. In this section, we report the results obtained from the experiments to obtain the best mix of slices representing the most informational relevant brain slices for the machine learning training dataset.

4.1.1 Covariance/Pearson's Correlation evaluation

We experimented with the covariance/Pearson's correlation values to investigate if there was an indication of information redundancy that could lead to selecting the more relevant slices to build our dataset of slices.

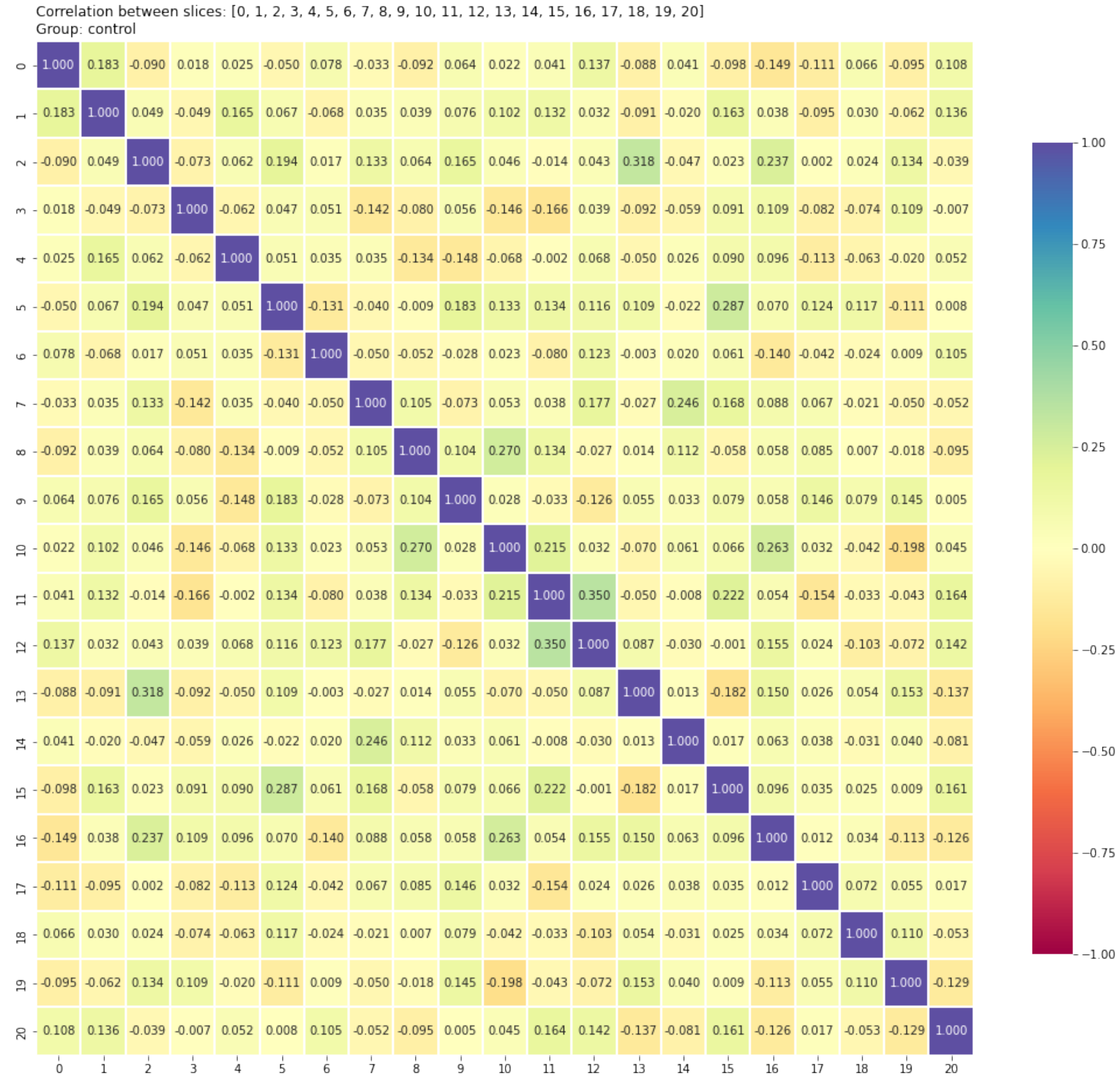
In the following graphics, we present the results from evaluating the covariance between slices of the Schizophrenia subjects group. Our intuition was to verify if adjacent slices would have a great correlation between them that could lead to eliminating one of them. A combination having the greatest chance to be in this hypothesis was between slices 1 and 2 but with a small correlation value of 0.234. However, most combinations were under this value, so we did not consider this approach.



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	1.000	0.034	-0.113	-0.009	0.148	0.048	0.115	-0.039	-0.026	0.053	0.147	-0.079	-0.011	0.120	0.005	0.136	-0.041	-0.114	-0.129	-0.094	0.013
1	0.034	1.000	0.234	-0.209	0.037	-0.009	-0.046	-0.126	-0.105	0.010	-0.108	-0.073	0.021	-0.085	0.061	0.018	0.087	-0.058	0.061	-0.079	0.119
2	-0.113	0.234	1.000	-0.185	-0.047	0.079	0.109	0.028	-0.007	-0.083	0.051	0.069	0.155	0.100	0.005	0.200	0.041	0.103	0.072	-0.004	0.041
3	-0.009	-0.209	-0.185	1.000	-0.037	-0.029	0.034	0.009	-0.100	-0.023	0.046	-0.051	0.120	0.045	0.189	-0.068	0.040	-0.033	0.025	-0.039	-0.128
4	0.148	0.037	-0.047	-0.037	1.000	-0.068	-0.096	-0.022	-0.010	0.015	-0.014	0.020	-0.009	0.102	0.045	0.087	-0.161	0.047	-0.093	0.020	0.025
5	0.048	-0.009	0.079	-0.029	-0.068	1.000	0.083	-0.057	0.079	-0.087	0.058	0.042	0.160	0.074	-0.029	0.115	-0.008	0.212	-0.113	-0.097	-0.060
6	0.115	-0.046	0.109	0.034	-0.096	0.083	1.000	-0.088	0.188	-0.203	0.004	0.002	0.038	0.185	0.012	0.114	0.025	0.054	-0.054	0.039	-0.004
7	-0.039	-0.126	0.028	0.009	-0.022	-0.057	-0.088	1.000	-0.047	-0.085	-0.067	0.085	0.042	0.291	0.085	0.140	0.086	0.004	-0.072	0.029	0.055
8	-0.026	-0.105	-0.007	-0.100	-0.010	0.079	0.188	-0.047	1.000	-0.156	0.043	0.066	0.068	0.167	0.075	0.046	-0.162	0.127	0.141	0.138	0.137
9	0.053	0.010	-0.083	-0.023	0.015	-0.087	-0.203	-0.085	-0.156	1.000	0.128	-0.112	0.002	-0.201	0.057	0.024	0.056	-0.134	0.171	-0.063	-0.125
10	0.147	-0.108	0.051	0.046	-0.014	0.058	0.004	-0.067	0.043	0.128	1.000	-0.028	-0.053	0.157	-0.064	0.160	0.046	0.037	-0.097	0.138	0.013
11	-0.079	-0.073	0.069	-0.051	0.020	0.042	0.002	0.085	0.066	-0.112	-0.028	1.000	-0.044	-0.037	0.126	-0.182	0.130	0.052	0.059	0.010	0.093
12	-0.011	0.021	0.155	0.120	-0.009	0.160	0.038	0.042	0.068	0.002	-0.053	-0.044	1.000	0.178	0.001	0.065	-0.041	0.192	0.177	-0.104	-0.077
13	0.120	-0.085	0.100	0.045	0.102	0.074	0.185	0.291	0.167	-0.201	0.157	-0.037	0.178	1.000	-0.028	0.147	0.067	0.081	-0.033	0.189	-0.024
14	0.005	0.061	0.005	0.189	0.045	-0.029	0.012	0.085	0.075	0.057	-0.064	0.126	0.001	-0.028	1.000	-0.117	0.047	0.080	0.041	0.154	0.010
15	0.136	0.018	0.200	-0.068	0.087	0.115	0.114	0.140	0.046	0.024	0.160	-0.182	0.065	0.147	-0.117	1.000	0.023	0.101	-0.265	0.066	-0.066
16	-0.041	0.087	0.041	0.040	-0.161	-0.008	0.025	0.086	-0.162	0.056	0.046	0.130	-0.041	0.067	0.047	0.023	1.000	-0.122	-0.062	0.066	0.070
17	-0.114	-0.058	0.103	-0.033	0.047	0.212	0.054	0.004	0.127	-0.134	0.037	0.052	0.192	0.081	0.080	0.101	-0.122	1.000	-0.034	-0.150	-0.004
18	-0.129	0.061	0.072	0.025	-0.093	-0.113	-0.054	-0.072	0.141	0.171	-0.097	0.059	0.177	-0.033	0.041	-0.265	-0.062	-0.034	1.000	-0.061	0.022
19	-0.094	-0.079	-0.004	-0.039	0.020	-0.097	0.039	0.029	0.138	-0.063	0.138	0.010	-0.104	0.189	0.154	0.066	0.066	-0.150	-0.061	1.000	0.080
20	0.013	0.119	0.041	-0.128	0.025	-0.060	-0.004	0.055	0.137	-0.125	0.013	0.093	-0.077	-0.024	0.010	-0.066	0.070	-0.004	0.022	0.080	1.000

Figure 4.2. Graphics showing correlation between slices of the Schizophrenia patients group

The following graphics show the results of evaluating the correlation between healthy control subjects group slices. As in the previous graphics, we made this experiment to verify if there was a correlation between two adjacent slices, indicating that we could eliminate one of them. We observed that slices 12 and 11 had a small correlation of 0.350, but we considered it to not be a relevant case for eliminating one of them based on this criterion.

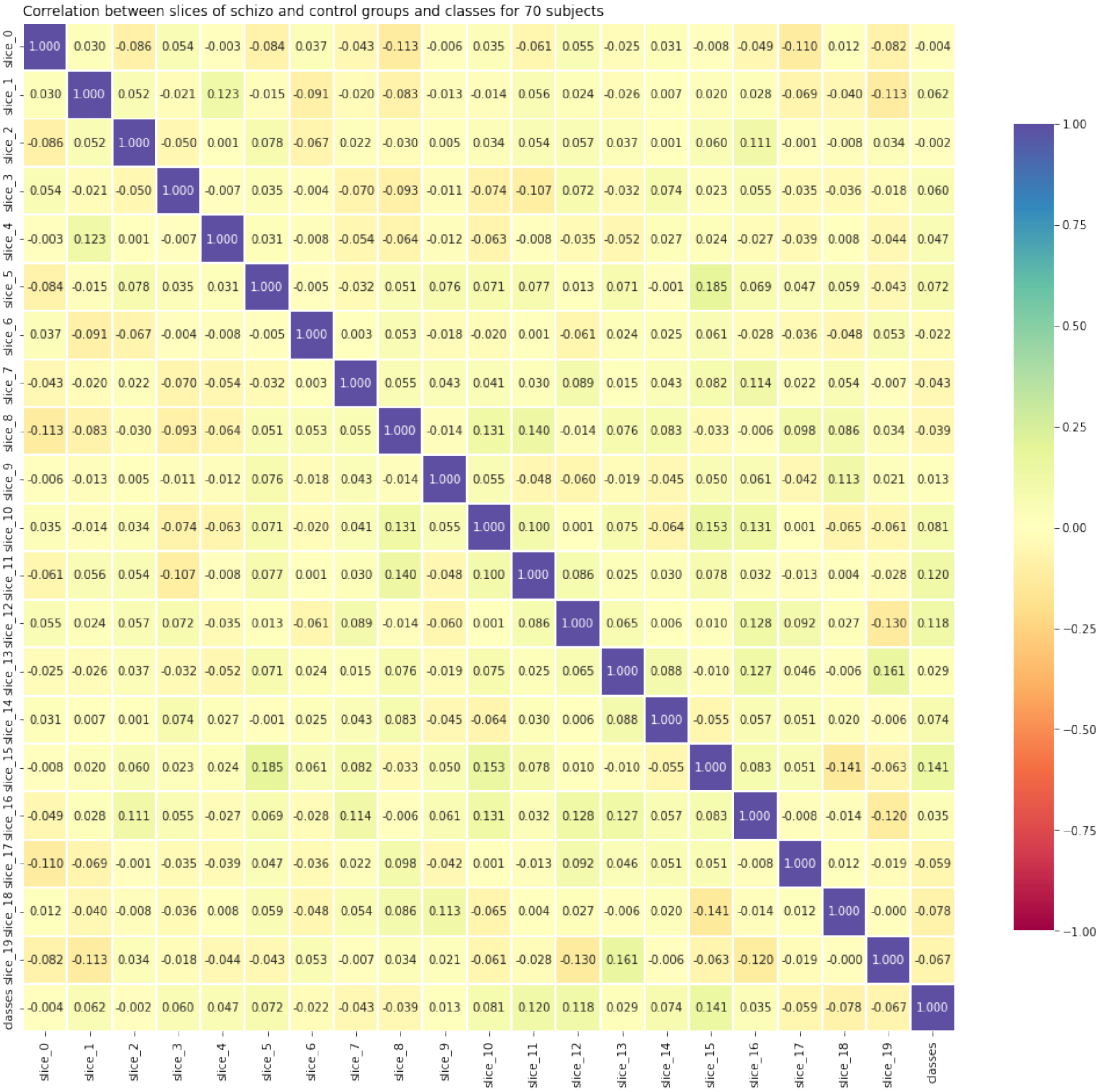


Correlation between slices: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]
Group: control

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	1.000	0.183	-0.090	0.018	0.025	-0.050	0.078	-0.033	-0.092	0.064	0.022	0.041	0.137	-0.088	0.041	-0.098	-0.149	-0.111	0.066	-0.095	0.108
1	0.183	1.000	0.049	-0.049	0.165	0.067	-0.068	0.035	0.039	0.076	0.102	0.132	0.032	-0.091	-0.020	0.163	0.038	-0.095	0.030	-0.062	0.136
2	-0.090	0.049	1.000	-0.073	0.062	0.194	0.017	0.133	0.064	0.165	0.046	-0.014	0.043	0.318	-0.047	0.023	0.237	0.002	0.024	0.134	-0.039
3	0.018	-0.049	-0.073	1.000	-0.062	0.047	0.051	-0.142	-0.080	0.056	-0.146	-0.166	0.039	-0.092	-0.059	0.091	0.109	-0.082	-0.074	0.109	-0.007
4	0.025	0.165	0.062	-0.062	1.000	0.051	0.035	0.035	-0.134	-0.148	-0.068	-0.002	0.068	-0.050	0.026	0.090	0.096	-0.113	-0.063	-0.020	0.052
5	-0.050	0.067	0.194	0.047	0.051	1.000	-0.131	-0.040	-0.009	0.183	0.133	0.134	0.116	0.109	-0.022	0.287	0.070	0.124	0.117	-0.111	0.008
6	0.078	-0.068	0.017	0.051	0.035	-0.131	1.000	-0.050	-0.052	-0.028	0.023	-0.080	0.123	-0.003	0.020	0.061	-0.140	-0.042	-0.024	0.009	0.105
7	-0.033	0.035	0.133	-0.142	0.035	-0.040	-0.050	1.000	0.105	-0.073	0.053	0.038	0.177	-0.027	0.246	0.168	0.088	0.067	-0.021	-0.050	-0.052
8	-0.092	0.039	0.064	-0.080	-0.134	-0.009	-0.052	0.105	1.000	0.104	0.270	0.134	-0.027	0.014	0.112	-0.058	0.058	0.085	0.007	-0.018	-0.095
9	0.064	0.076	0.165	0.056	-0.148	0.183	-0.028	-0.073	0.104	1.000	0.028	-0.033	-0.126	0.055	0.033	0.079	0.058	0.146	0.079	0.145	0.005
10	0.022	0.102	0.046	-0.146	-0.068	0.133	0.023	0.053	0.270	0.028	1.000	0.215	0.032	-0.070	0.061	0.066	0.263	0.032	-0.042	-0.198	0.045
11	0.041	0.132	-0.014	-0.166	-0.002	0.134	-0.080	0.038	0.134	-0.033	0.215	1.000	0.350	-0.050	-0.008	0.222	0.054	-0.154	-0.033	-0.043	0.164
12	0.137	0.032	0.043	0.039	0.068	0.116	0.123	0.177	-0.027	-0.126	0.032	0.350	1.000	0.087	-0.030	-0.001	0.155	0.024	-0.103	-0.072	0.142
13	-0.088	-0.091	0.318	-0.092	-0.050	0.109	-0.003	-0.027	0.014	0.055	-0.070	-0.050	0.087	1.000	0.013	-0.182	0.150	0.026	0.054	0.153	-0.137
14	0.041	-0.020	-0.047	-0.059	0.026	-0.022	0.020	0.246	0.112	0.033	0.061	-0.008	-0.030	0.013	1.000	0.017	0.063	0.038	-0.031	0.040	-0.081
15	-0.098	0.163	0.023	0.091	0.090	0.287	0.061	0.168	-0.058	0.079	0.066	0.222	-0.001	-0.182	0.017	1.000	0.096	0.035	0.025	0.009	0.161
16	-0.149	0.038	0.237	0.109	0.096	0.070	-0.140	0.088	0.058	0.058	0.263	0.054	0.155	0.150	0.063	0.096	1.000	0.012	0.034	-0.113	-0.126
17	-0.111	-0.095	0.002	-0.082	-0.113	0.124	-0.042	0.067	0.085	0.146	0.032	-0.154	0.024	0.026	0.038	0.035	0.012	1.000	0.072	0.055	0.017
18	0.066	0.030	0.024	-0.074	-0.063	0.117	-0.024	-0.021	0.007	0.079	-0.042	-0.033	-0.103	0.054	-0.031	0.025	0.034	0.072	1.000	0.110	-0.053
19	-0.095	-0.062	0.134	0.109	-0.020	-0.111	0.009	-0.050	-0.018	0.145	-0.198	-0.043	-0.072	0.153	0.040	0.009	-0.113	0.055	0.110	1.000	-0.129
20	0.108	0.136	-0.039	-0.007	0.052	0.008	0.105	-0.052	-0.095	0.005	0.045	0.164	0.142	-0.137	-0.081	0.161	-0.126	0.017	-0.053	-0.129	1.000

Figure 4.3. Graphics showing correlation between slices of the healthy control subjects group

The last experiment we made was to use slices of both groups and search for correlation between adjacent slices and between slices and classes. The following graphics show minimal values for the results, leading us to conclude that covariance is not a good metric for our present problem: select the best slices to build our dataset.



	slice_0	slice_1	slice_2	slice_3	slice_4	slice_5	slice_6	slice_7	slice_8	slice_9	slice_10	slice_11	slice_12	slice_13	slice_14	slice_15	slice_16	slice_17	slice_18	slice_19	classes
slice_0	1.000	0.030	-0.086	0.054	-0.003	-0.084	0.037	-0.043	-0.113	-0.006	0.035	-0.061	0.055	-0.025	0.031	-0.008	-0.049	-0.110	0.012	-0.082	-0.004
slice_1	0.030	1.000	0.052	-0.021	0.123	-0.015	-0.091	-0.020	-0.083	-0.013	-0.014	0.056	0.024	-0.026	0.007	0.020	0.028	-0.069	-0.040	-0.113	0.062
slice_2	-0.086	0.052	1.000	-0.050	0.001	0.078	-0.067	0.022	-0.030	0.005	0.034	0.054	0.057	0.037	0.001	0.060	0.111	-0.001	-0.008	0.034	-0.002
slice_3	0.054	-0.021	-0.050	1.000	-0.007	0.035	-0.004	-0.070	-0.093	-0.011	-0.074	-0.107	0.072	-0.032	0.074	0.023	0.055	-0.035	-0.036	-0.018	0.060
slice_4	-0.003	0.123	0.001	-0.007	1.000	0.031	-0.008	-0.054	-0.064	-0.012	-0.063	-0.008	-0.035	-0.052	0.027	0.024	-0.027	-0.039	0.008	-0.044	0.047
slice_5	-0.084	-0.015	0.078	0.035	0.031	1.000	-0.005	-0.032	0.051	0.076	0.071	0.077	0.013	0.071	-0.001	0.185	0.069	0.047	0.059	-0.043	0.072
slice_6	0.037	-0.091	-0.067	-0.004	-0.008	-0.005	1.000	0.003	0.053	-0.018	-0.020	0.001	-0.061	0.024	0.025	0.061	-0.028	-0.036	-0.048	0.053	-0.022
slice_7	-0.043	-0.020	0.022	-0.070	-0.054	-0.032	0.003	1.000	0.055	0.043	0.041	0.030	0.089	0.015	0.043	0.082	0.114	0.022	0.054	-0.007	-0.043
slice_8	-0.113	-0.083	-0.030	-0.093	-0.064	0.051	0.053	0.055	1.000	-0.014	0.131	0.140	-0.014	0.076	0.083	-0.033	-0.006	0.098	0.086	0.034	-0.039
slice_9	-0.006	-0.013	0.005	-0.011	-0.012	0.076	-0.018	0.043	-0.014	1.000	0.055	-0.048	-0.060	-0.019	-0.045	0.050	0.061	-0.042	0.113	0.021	0.013
slice_10	0.035	-0.014	0.034	-0.074	-0.063	0.071	-0.020	0.041	0.131	0.055	1.000	0.100	0.001	0.075	-0.064	0.153	0.131	0.001	-0.065	-0.061	0.081
slice_11	-0.061	0.056	0.054	-0.107	-0.008	0.077	0.001	0.030	0.140	-0.048	0.100	1.000	0.086	0.025	0.030	0.078	0.032	-0.013	0.004	-0.028	0.120
slice_12	0.055	0.024	0.057	0.072	-0.035	0.013	-0.061	0.089	-0.014	-0.060	0.001	0.086	1.000	0.065	0.006	0.010	0.128	0.092	0.027	-0.130	0.118
slice_13	-0.025	-0.026	0.037	-0.032	-0.052	0.071	0.024	0.015	0.076	-0.019	0.075	0.025	0.065	1.000	0.088	-0.010	0.127	0.046	-0.006	0.161	0.029
slice_14	0.031	0.007	0.001	0.074	0.027	-0.001	0.025	0.043	0.083	-0.045	-0.064	0.030	0.006	0.088	1.000	-0.055	0.057	0.051	0.020	-0.006	0.074
slice_15	-0.008	0.020	0.060	0.023	0.024	0.185	0.061	0.082	-0.033	0.050	0.153	0.078	0.010	-0.010	-0.055	1.000	0.083	0.051	-0.141	-0.063	0.141
slice_16	-0.049	0.028	0.111	0.055	-0.027	0.069	-0.028	0.114	-0.006	0.061	0.131	0.032	0.128	0.127	0.057	0.083	1.000	-0.008	-0.014	-0.120	0.035
slice_17	-0.110	-0.069	-0.001	-0.035	-0.039	0.047	-0.036	0.022	0.098	-0.042	0.001	-0.013	0.092	0.046	0.051	0.051	-0.008	1.000	0.012	-0.019	-0.059
slice_18	0.012	-0.040	-0.008	-0.036	0.008	0.059	-0.048	0.054	0.086	0.113	-0.065	0.004	0.027	-0.006	0.020	-0.141	-0.014	0.012	1.000	-0.000	-0.078
slice_19	-0.082	-0.113	0.034	-0.018	-0.044	-0.043	0.053	-0.007	0.034	0.021	-0.061	-0.028	-0.130	0.161	-0.006	-0.063	-0.120	-0.019	-0.000	1.000	-0.067
classes	-0.004	0.062	-0.002	0.060	0.047	0.072	-0.022	-0.043	-0.039	0.013	0.081	0.120	0.118	0.029	0.074	0.141	0.035	-0.059	-0.078	-0.067	1.000

Figure 4.4. Graphics showing correlation between slices of both subjects group

The results presented by the covariance evaluation did not indicate a high correlation between the two classes and the slices and the slices among them. Therefore, covariance does not seem to be the best indicator. We think that this is an intrinsic limitation of this measurement as it is a metric that analyzes only one kind of dependency, and also we had a small dataset. The figures in this section showed that the correlation values were near zero, indicating an insignificant correlation between the slices. It would be more significant if it were near the +1 or the -1 value. Because of this, we experimented with the Entropy value to have a metric for evaluating the best slices to build the dataset.

4.1.2 Entropy evaluation

4.1.2.1 Entropy evaluation of individual slices of a random control subject

As stated in the methodology, one of the experiments was to evaluate the entropy as a criterion to select the slices that would provide the best results when added to the dataset used for training a machine learning classifier. Therefore, we choose the Shannon Entropy calculation for the image file. The Shannon entropy, represented by the formula

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

, will give us an estimation of the diverseness of the information contained in the image. The intuition is that the greater this metric, the more information the image contains more relevant it will be for the machine learning classifier. Therefore, we submitted the images of the slices individually and calculated the entropy for all the control subjects and the schizophrenia patients. Table 4.1 shows the value of Shannon entropy for image slices of a random control subject.

Slice	Entropy	Slice	Entropy
6	5.459513	4	5.100624
7	5.454676	19	5.075619
3	5.387418	13	5.068898
10	5.376120	20	4.964131
9	5.277592	8	4.937223
15	5.261407	5	4.924651
0	5.248929	21	4.850773
2	5.244146	1	4.840004
12	5.238758	22	4.715308
11	5.229948	23	4.383436
18	5.189273	24	4.191262
17	5.163017	25	3.736641
14	5.160322	26	3.619608
16	5.145192		

Table 4.1. Shannon Entropy values of MRI scan slices from a random subject. It is in descent sorted order of entropy:

As we axially traverse the slices from bottom to top of the skull, we can observe that the Shannon entropy value of the image decreases as it reaches the top of the skull. As we can see in 4.1, the images have fewer internal structures and randomness. Except for slice 1, all the greater values of Shannon entropy values are obtained until slice 20, and after we notice a significant decrease of the Shannon entropy value.

4.1.2.2 Entropy evaluation of individual slices of a group of subjects and comparison between the groups

After evaluating the Shannon entropy for a unique subject for establishing a baseline reference, we calculated the average Shannon entropy for the two groups of subjects: the control group and the group of schizophrenia patients. The objective was to observe the variation of the average values between the two groups and if there was any significant variation in the sorted order of the slices based on their Shannon entropy values.

The results are shown as follows: table 4.2 shows the obtained average Shannon entropy values for the control group per slice, and table 4.3 shows the average Shannon entropy obtained values for the schizophrenic patient's group per slice:

Slice	Entropy
6	5.667120
7	5.660915
10	5.651297
11	5.632566
9	5.625398
8	5.618431
5	5.595073
12	5.581329
4	5.565319
13	5.540825
14	5.518174
3	5.510389
1	5.478136
15	5.471145
2	5.456008
16	5.442481
17	5.415122
0	5.403902
18	5.389171
19	5.294878
20	5.201705
21	5.080780
22	4.939645
23	4.728316
24	4.506750
25	4.322952
26	4.178690

Table 4.2. Shannon Entropy average values of MRI scan slices for the control group. It is in descent sorted order of entropy

Slice	Entropy
10	5.741600
7	5.740798
6	5.735013
5	5.726238
9	5.721837
11	5.712964
12	5.699338
8	5.693178
3	5.682558
4	5.678587
2	5.655890
1	5.646118
13	5.638080
14	5.634989
0	5.600223
15	5.597278
16	5.535400
17	5.471225
18	5.437461
19	5.369407
20	5.284730
21	5.167518
22	5.000726
23	4.814832
24	4.586878
25	4.343048
26	4.182541

Table 4.3. Shannon Entropy average values of MRI scan slices for schizophrenia patients group. It is sorted in descent order of entropy

Values presented in both tables were obtained by calculating Shannon entropy. First, it was calculated for every slice of the MRI scan for all subjects in both groups; finally, the average value was obtained. The same situation that we have observed in the results of one random control individual is repeated for the average value of the group, the five last slices near the top part of the skull have a smaller Shannon entropy value than the others leading us to the same intuition that they have fewer varieties for the contained information. Another observation is that for the two groups, the slices with greater entropy are 5,6,7,8,9,10,11. They only exchange positions between the two groups. In general, the control group's average entropy values are a little smaller than the group of schizophrenia patients.

Table 4.4 shows us again the average value of Shannon entropy for each slice of the dataset, but this time we have calculated the average for the whole dataset, control group, and schizophrenia patients.

Slice	Entropy	Slice	Entropy
6	5.700776	15	5.533673
7	5.700515	0	5.501223
10	5.696063	16	5.488543
9	5.673205	17	5.442934
11	5.672421	18	5.413110
5	5.660095	19	5.331824
8	5.655485	20	5.242862
12	5.639829	21	5.123778
4	5.621469	22	4.969925
3	5.595738	23	4.771205
13	5.589036	24	4.546472
14	5.576082	25	4.332914
1	5.561409	26	4.180599
2	5.555095		

Table 4.4. Shannon Entropy average values of MRI scan slices for the whole dataset. It is in descent sorted order of entropy

The same information is shown in graph form in figure 4.5. Again, it shows little variation in entropy values at initial slices with an accentuated descent after slice 19.

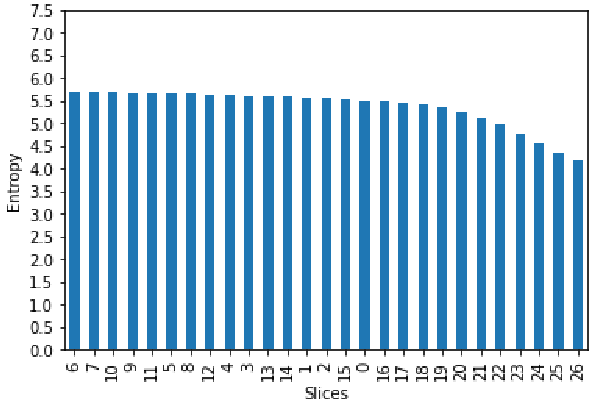


Figure 4.5. Graphics showing average entropy for individual slices of all subjects.

4.1.2.3 Slice selection based on images entropy

The steps made in the previous subsections aimed to obtain a list of slices, ordered by its entropy value. This list was used for the next step: submit the dataset formed by the slices to a machine learning classifier and obtain metrics about their performance as a classifier of images. This information will permit us to explore which slices enhance not the results of the machine learning classifier.

In order to evaluate which slices would be the best to build a dataset for the classifier, we prepared datasets with an incremental list of slices, using the average Shannon Entropy value ordered list referenced in 4.4 as input and presented this dataset for training in the inputs of the machine learning model, training and evaluation. The result list of slices for the interaction was the following: *6 , 7 , 10, 9 , 11, 5 , 8 , 12, 4 , 3 , 13, 14, 1 , 2 , 15, 0 , 16, 17, 18, 19, 20, 21, 22, 23, 24, 25 and 26.* In each interaction, we added a new slice, built a novel dataset, trained the ML model, evaluated it, and collected the metrics of loss and accuracy for the best model evaluating the validation dataset. These collected data is presented in the table 4.5:

Slices	Accuracy	Loss
6	0.765258	0.834440
6, 7	0.629108	1.798356
6, 7, 10	0.690141	1.900789
6, 7, 10, 9	0.671362	2.221895
6, 7, 10, 9, 11	0.586854	5.479521
6, 7, 10, 9, 11, 5	0.591549	4.305631
6, 7, 10, 9, 11, 5, 8	0.638498	3.156163
6, 7, 10, 9, 11, 5, 8, 12	0.610329	2.968923
6, 7, 10, 9, 11, 5, 8, 12, 4	0.544601	3.753664
6, 7, 10, 9, 11, 5, 8, 12, 4, 3	0.596244	3.409455
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13	0.624413	3.031253
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14	0.633803	6.443608
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1	0.507042	5.394758
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2	0.516432	2.899813
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15	0.441315	4.633669
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0	0.596244	2.920489
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16	0.708920	3.967879
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17	0.403756	6.129658
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18	0.422535	5.542782
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19	0.525822	6.297647
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20	0.666667	5.515784
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21	0.530516	6.183190
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22	0.638498	5.120807
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23	0.516432	3.087831
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23, 24	0.535211	4.538959
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	0.549296	6.915903
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26	0.572770	5.577716

Table 4.5. Result metrics were obtained from incrementally adding slices to the training dataset

In order to permit a better view of the results, we have shown the same table 4.5 in two versions:

- Table 4.6 shows the collected data, sorted by loss metrics of the model, ascending ordered.
- Table 4.7 shows the data sorted by accuracy metrics of the model in descending order.

Slices	Accuracy	Loss
6	0.765258	0.834440
6, 7	0.629108	1.798356
6, 7, 10	0.690141	1.900789
6, 7, 10, 9	0.671362	2.221895
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2	0.516432	2.899813
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0	0.596244	2.920489
6, 7, 10, 9, 11, 5, 8, 12	0.610329	2.968923
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13	0.624413	3.031253
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23	0.516432	3.087831
6, 7, 10, 9, 11, 5, 8	0.638498	3.156163
6, 7, 10, 9, 11, 5, 8, 12, 4, 3	0.596244	3.409455
6, 7, 10, 9, 11, 5, 8, 12, 4	0.544601	3.753664
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16	0.708920	3.967879
6, 7, 10, 9, 11, 5	0.591549	4.305631
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23, 24	0.535211	4.538959
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15	0.441315	4.633669
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22	0.638498	5.120807
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1	0.507042	5.394758
6, 7, 10, 9, 11	0.586854	5.479521
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20	0.666667	5.515784
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18	0.422535	5.542782
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26	0.572770	5.577716
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17	0.403756	6.129658
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21	0.530516	6.183190
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19	0.525822	6.297647
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14	0.633803	6.443608
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	0.549296	6.915903

Table 4.6. Result metrics ascending ordered by loss

Slices	Accuracy	Loss
6	0.765	0.834
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16	0.709	3.968
6, 7, 10	0.690	1.901
6, 7, 10, 9	0.671	2.222
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20	0.667	5.516
6, 7, 10, 9, 11, 5, 8	0.638	3.156
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22	0.638	5.121
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14	0.633803	6.443608
6, 7	0.629108	1.798356
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13	0.624413	3.031253
6, 7, 10, 9, 11, 5, 8, 12	0.610329	2.968923
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0	0.596244	2.920489
6, 7, 10, 9, 11, 5, 8, 12, 4, 3	0.596244	3.409455
6, 7, 10, 9, 11, 5	0.591549	4.305631
6, 7, 10, 9, 11	0.586854	5.479521
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26	0.572770	5.577716
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	0.549296	6.915903
6, 7, 10, 9, 11, 5, 8, 12, 4	0.544601	3.753664
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23, 24	0.535211	4.538959
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21	0.530516	6.183190
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19	0.525822	6.297647
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18, 19, 20, 21, 22, 23	0.516432	3.087831
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2	0.516432	2.899813
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1	0.507042	5.394758
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15	0.441315	4.633669
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17, 18	0.422535	5.542782
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, 17	0.403756	6.129658

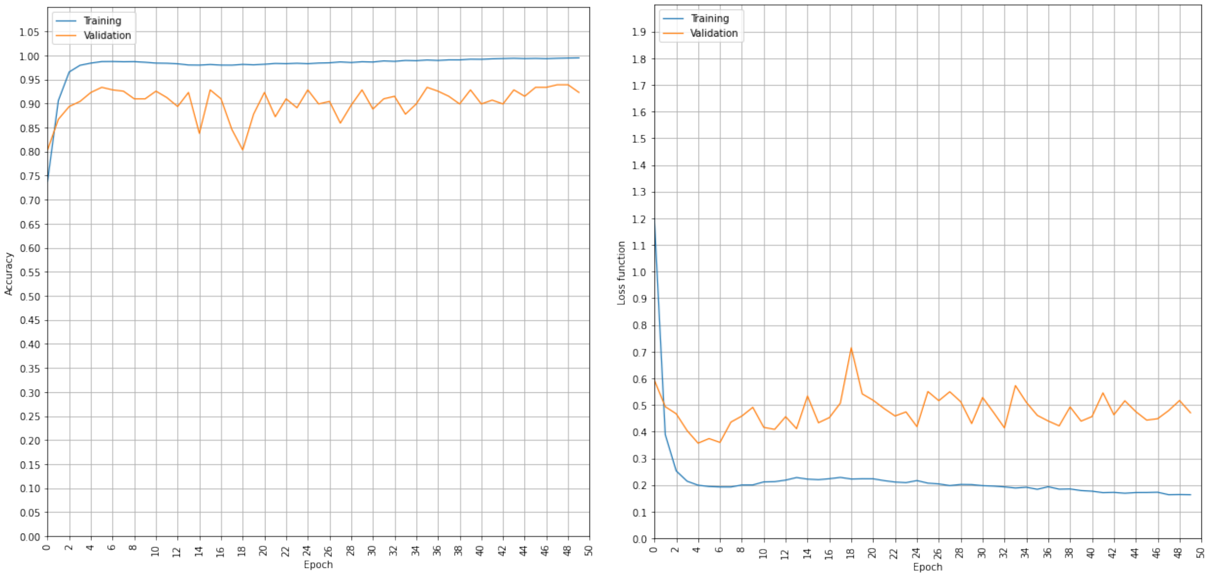
Table 4.7. Result metrics are sorted by accuracy descending.

We observed that slice six alone was responsible for obtaining the best metrics values, no matter whether using loss or accuracy for comparison between the collection of slices. Analyzing the results sorted by *loss metrics*, we observe that the slices *6, 7, 10, and 9* incremental combinations represent the top c4 combinations for accuracy. The table 4.7 sorted by accuracy shows similar information, with slices *6, 7,10 and 9* being present in the top combinations again, but we noticed that the second-best accuracy was obtained by a longer list, composed of the slices *6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16*, a somewhat unexpected result.

4.2 Model Training and Classification Output

4.2.1 Model training metrics

The model training was configured to collect metrics during its execution. We choose to collect *Accuracy* and *Loss* for the training dataset and the validation dataset originating from a 20% split of the training dataset. Due to the small number of images, we worked on increasing the number of images by using data augmentation. Also, we worked on a new CNN architecture, yet to be described in a future version of this document. Figure 4.6 shows the graphs obtained from the collected data. Figure *(a)* is the graph obtained from *Accuracy* data and shows that the validation dataset is following the training dataset. Figure *(b)* shows the *Loss function* with the validation dataset oscillating around 0.4, indicating a good response to images not previously seen by the model.



(a)

(b)

Figure 4.6. Model Training Metrics Graphs. (a) Epochs x Accuracy. (b) Epochs x Loss function.

4.2.2 Performance metrics for various slices combinations

We used slices 9 to 16 for model evaluation, first as individual slices and after doing incremental combinations of slices, creating the dataset used for training and evaluation. Figure 4.1 shows examples of the slices used in the experiment.

The performance metrics collected from the experiment are displayed in the following tables:

Table 4.8. Metrics for Individual Slices and Combinations

Slice	Accuracy	AUC
9	56.73% (+/- 10.41%)	0.58 (+/- 0.16)
10	56.73% (+/- 10.41%)	0.58 (+/- 0.16)
11	52.05% (+/- 19.25%)	0.52 (+/- 0.17)
12	55.00% (+/- 14.63%)	0.57 (+/- 0.15)
13	50.14% (+/- 13.46%)	0.52 (+/- 0.18)
14	58.95% (+/- 18.03%)	0.63 (+/- 0.15)
15	46.68% (+/- 15.69%)	0.49 (+/- 0.16)
16	63.14% (+/- 11.84%)	0.65 (+/- 0.15)
9-10	78.10% (+/- 9.69%)	0.84 (+/- 0.12)
9-11	88.97% (+/- 5.77%)	0.94 (+/- 0.04)
9-12	87.22% (+/- 3.15%)	0.93 (+/- 0.03)
9-13	88.98% (+/- 4.49%)	0.93 (+/- 0.04)
9-14	88.73% (+/- 4.96%)	0.94 (+/- 0.04)
9-15	90.13% (+/- 4.20%)	0.94 (+/- 0.04)
9-16	87.77% (+/- 3.04%)	0.93 (+/- 0.03)

Table 4.9. Metrics for Individual Slices and Combinations

Slice	Precision	Recall
9	56.73% (+/- 10.41%)	56.73% (+/- 10.41%)
10	56.73% (+/- 10.41%)	56.73% (+/- 10.41%)
11	51.99% (+/- 19.27%)	51.09% (+/- 19.57%)
12	54.90% (+/- 14.67%)	55.00% (+/- 15.14%)
13	50.30% (+/- 13.65%)	49.18% (+/- 13.04%)
14	58.79% (+/- 18.26%)	58.45% (+/- 18.75%)
15	47.22% (+/- 15.55%)	48.18% (+/- 15.38%)
16	63.08% (+/- 11.93%)	62.64% (+/- 12.67%)
9-10	78.10% (+/- 9.69%)	78.10% (+/- 9.69%)
9-11	88.94% (+/- 5.60%)	88.97% (+/- 6.08%)
9-12	87.31% (+/- 3.08%)	87.10% (+/- 3.29%)
9-13	89.24% (+/- 4.66%)	88.49% (+/- 5.00%)
9-14	88.93% (+/- 4.96%)	89.02% (+/- 4.09%)
9-15	91.07% (+/- 4.45%)	87.59% (+/- 3.12%)
9-16	87.91% (+/- 3.04%)	88.70% (+/- 4.41%)

We have chosen to add slices to the dataset by their index at each interaction, as our intuition suggested that we would be able to identify the slices that would enhance the results metrics or turn it worse. The results table above shows that each slice added increased the results until slice *16*, which decreased the metrics. When using only one slice for training, the worst metrics were obtained with slice *15.* Slices *15* and *16* are the slices where brain structures are less visible.

4.2.3 Results using a State-of-the-art Model Architecture

One of the final tasks of this research was to submit the datasets of collections of slices to a state-of-the-art Neural Network architecture, train, validate and collect its results, and compare them with the results of a shallow model first used to create an evaluation baseline. We used a VGG16 architecture; although it is not the most recent one, it is still an architecture with great accuracy.

In order to obtain the results we choose to use the top 2 accuracy slices combinations by accuracy , in table 4.10 that is part of 4.7.

Slices	Accuracy	Loss
6	0.765258	0.834440
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16	0.708920	3.967879

Table 4.10. Top 2 result metrics sorted by accuracy descending.

Slices	Accuracy	Loss
6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16	0.791	2.593
6	0.677	2.032

Table 4.11. Top 2 result metrics sorted by accuracy descending.

With a VGG16 model adapted to our classification problem, we submitted the model to training and evaluation of the validation dataset. After approximately 13 hours of training for each of the top 2 accuracy collection of slices, we obtained history data with the results of the training performance.

Although in all our work, we pursue the possibility that a dataset with more than one scan slice image would be the perfect collection to submit for training and prediction in a machine learning engine, the results obtained by evaluating the entropy of the images revealed that the slice 6 submitted alone, was responsible for achieving the maximum accuracy. Therefore, we submitted training a dataset built only with slice six images to the state-of-the-art engine VGG16 and obtained the performance graph Accuracy vs. Epoch, shown in 4.7 and the graph Loss vs. Epoch showed in 4.8.

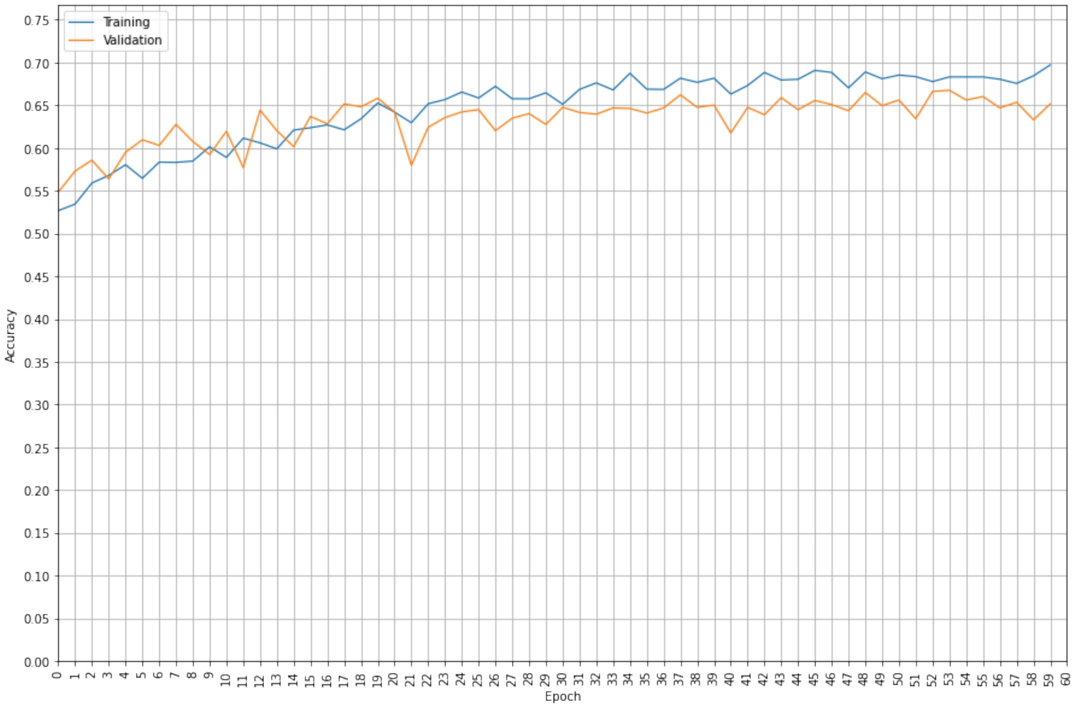


Figure 4.7. Graph showing curve Accuracy vs Epochs for first place slice collection.

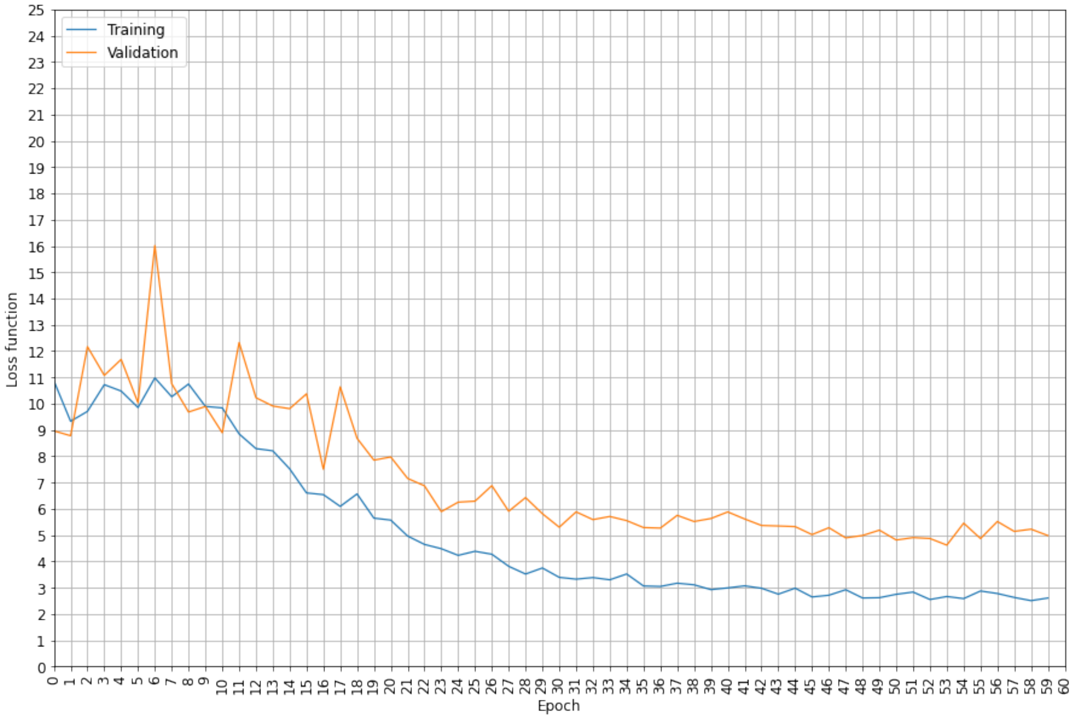


Figure 4.8. Graph showing curve Loss vs Epochs for first place slice collection.

The second place slice image collection, composed of slices 6, 7, 10, 9, 11, 5, 8, 12, 4, 3, 13, 14, 1, 2, 15, 0, 16, has the performance graph for accuracy represented in figure 4.9 and the performance graph for loss metric represented in figure 4.10

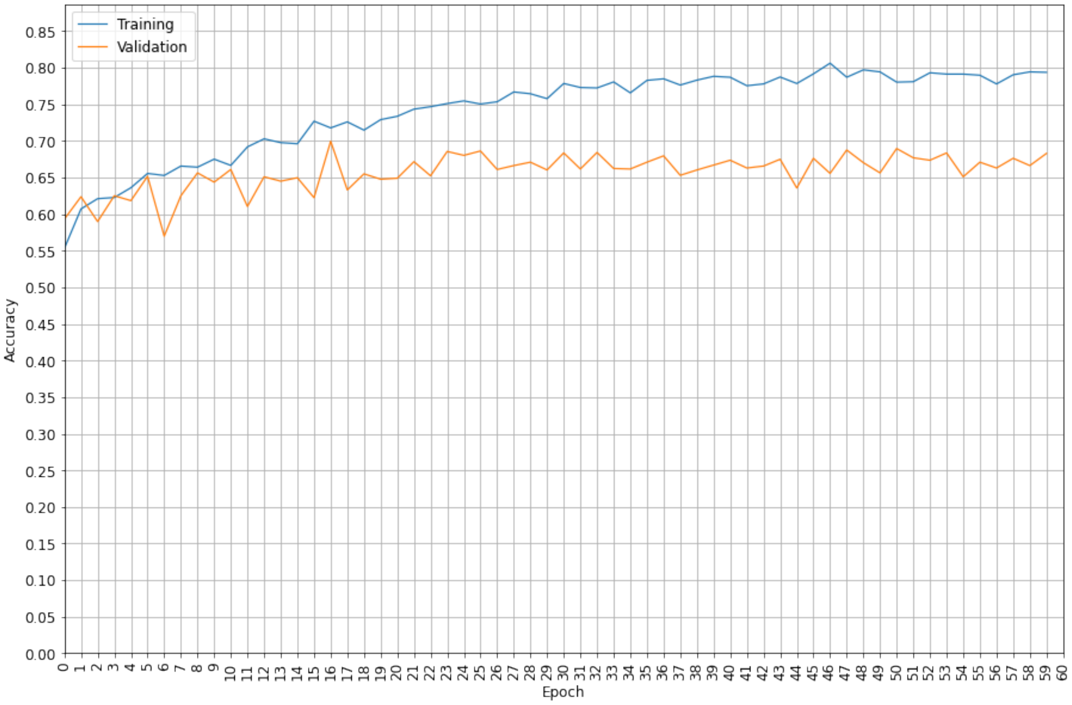


Figure 4.9. Graph showing curve Accuracy vs Epochs for second place slice collection.

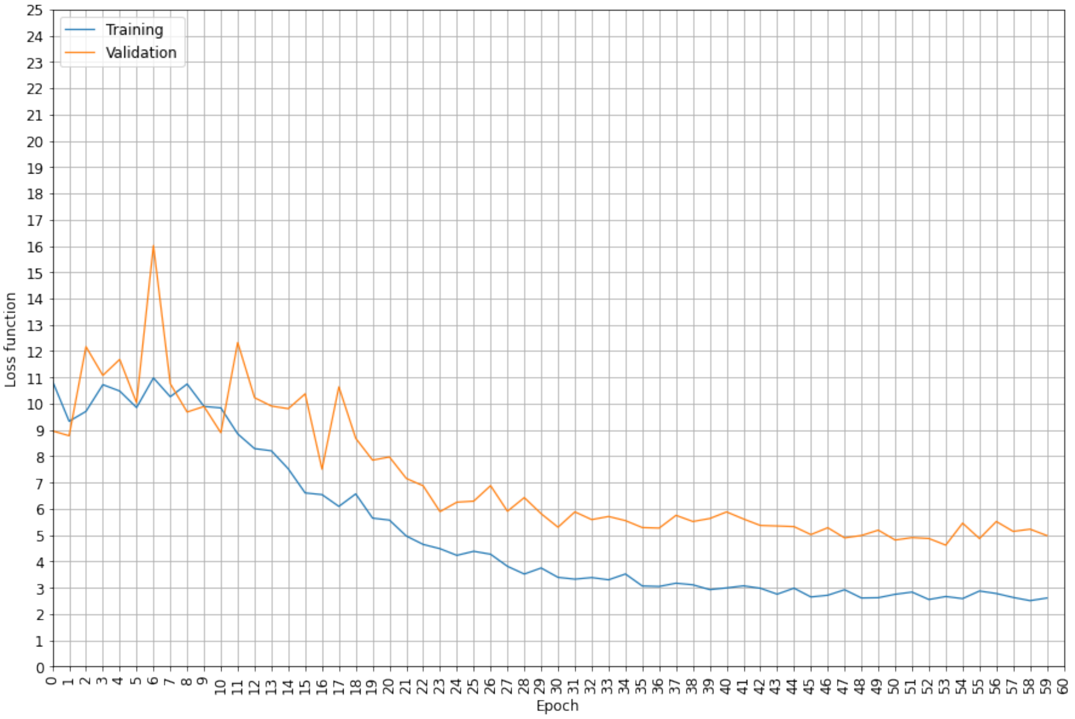


Figure 4.10. Graph showing curve Loss vs Epochs for second place slice collection.

During the presentation of the thesis to the university post-graduation evaluation committee, one of the members Dr. Fabricio Ataides Braz, noted that slice 16 improved the results significantly when added to the slices collection for evaluation. Based on this observation, we made an experiment using only slice 16, which resulted in the graphs shown in 4.11 and 4.12.

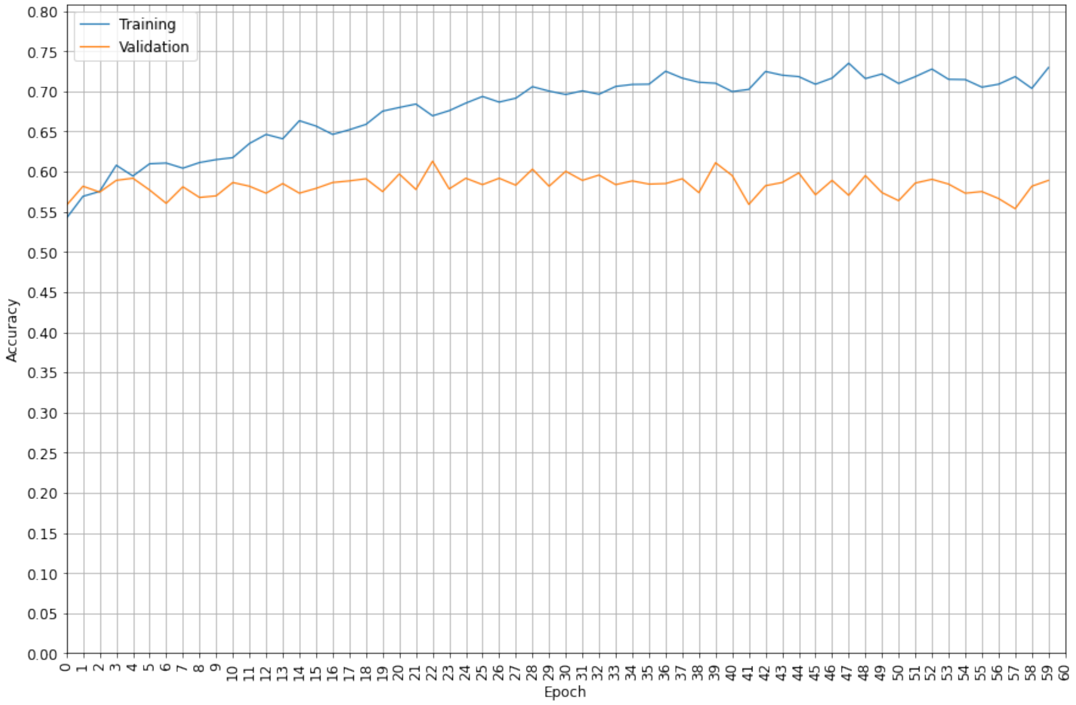


Figure 4.11. Graph showing curve Accuracy vs Epochs for slice 16.

4.3 Explaining the prediction with XAI

The final experiment of this research was to visualize an explanation about the prediction made by the model. The results were obtained using a custom model based on state-of-the-art VGG16 architecture. The model was trained using a transfer learning strategy to reduce the training time, by using the previous weights of the architecture trained on the Imagenet dataset

First we evaluated a random image from the dataset built from only images of slice six of the Schizophrenic patients group and the model trained only with slice 6 images dataset . The results are represented in figure 4.13. *(a)* shows the original image of subject and figure 4.13. *(b)* shows the superimposed heatmap image indicating the relevant areas for the classification.

We can notice that the heatmap is indicating that the relevant areas are locate at the midle of the slice. It includes the

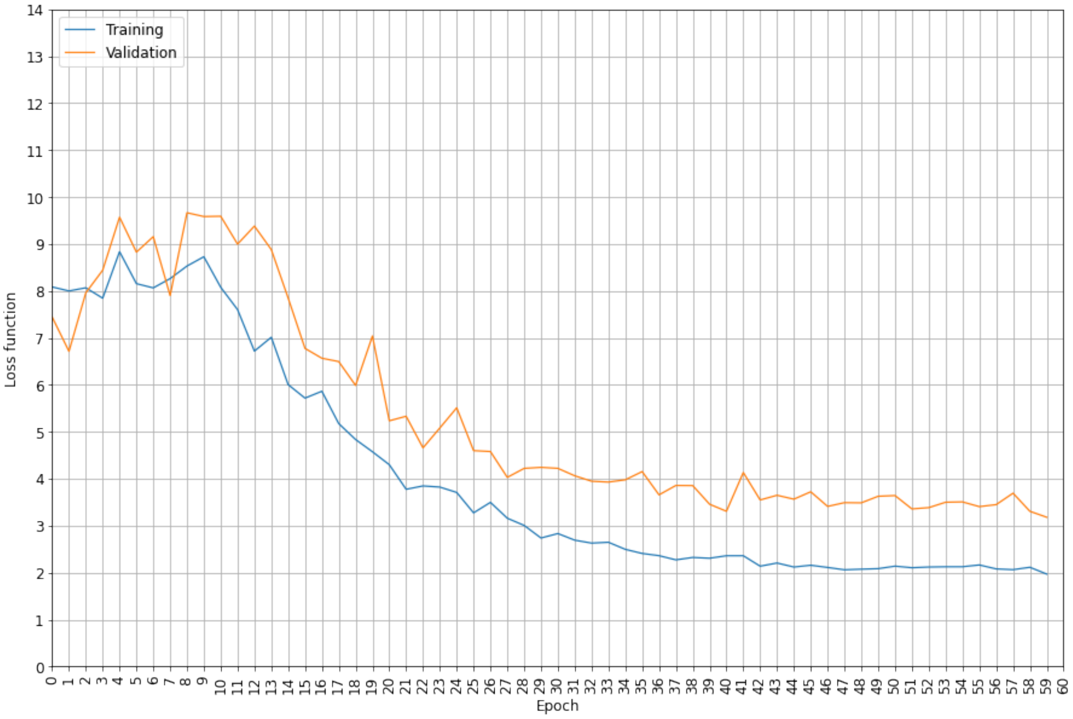


Figure 4.12. Graph showing curve Loss vs Epochs for slice 16.

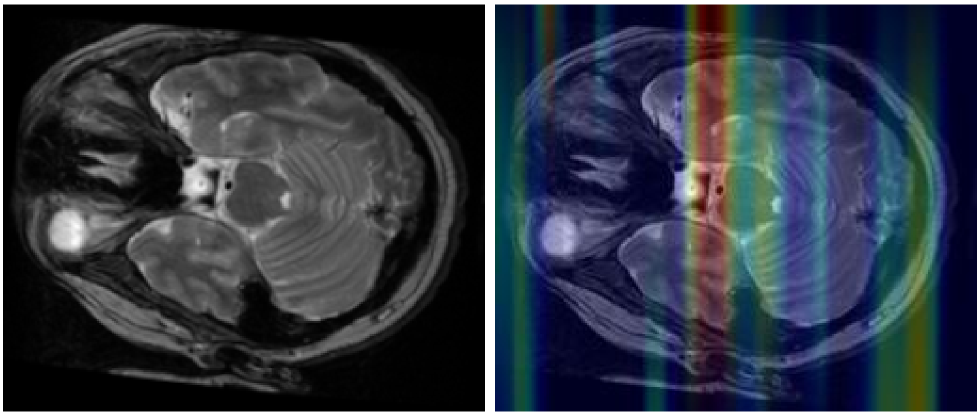


Figure 4.13. (a) The figure shows the original image of slice 6 for the random Schizophrenia subject (b) Visualization of heatmap showing relevant areas for the classification result

After that, we evaluated another random image but this time from the dataset built with slices from 0 to 16 of the Schizophrenic patients group and trained with the dataset of images from 0 to 16. The results are represented in figure 4.14 that shows the original image of subject and figure 4.15, shows the superimposed heatmap image indicating the relevant areas for the classification.

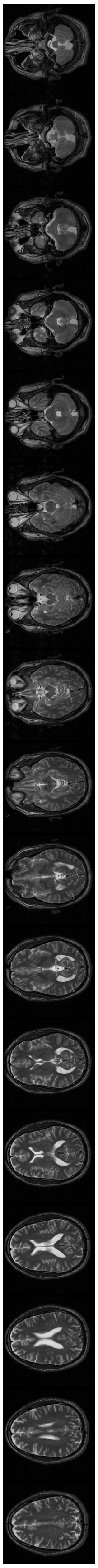


Figure 4.14. The figure shows the original image of slices 0 until 16 for the random Schizophrenia subject.

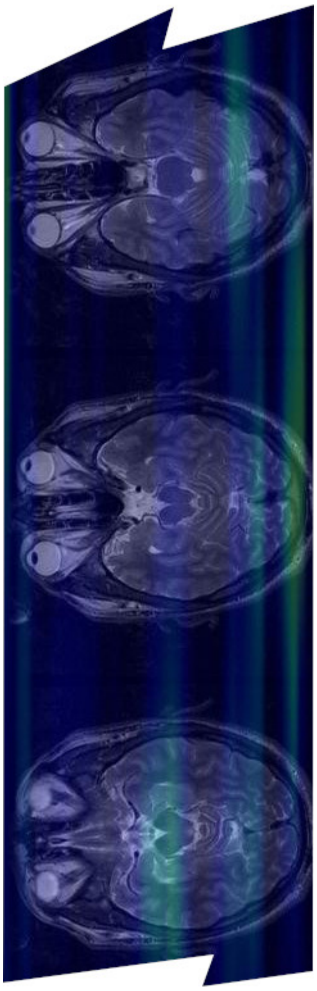


Figure 4.15. Visualization of heatmap showing relevant areas for the classification result

5 Conclusions

In this work, we proposed to evaluate a systematic way to determine which combination of anatomical MRI scan axial slices would lead to better results for the problem of classification using a Neural Network to detect the presence of schizophrenia or not.

First, we experimented with Covariance/Pearson's Correlation to determine if one slice could be correlated with other slices, but the results did not indicate a relevant correlation between the slices.

After that, we experimented with Entropy. This time our results indicated that the Entropy value was a significant metric to indicate which images are more relevant in a dataset for training the machine learning model, bringing an intuition that the greater the value of the Entropy of the slice, the more diverse was the content of the image, with more structures represented in it.

The next step was to submit the list of slices to a machine learning model evaluation and build a list of slices collections ordered by accuracy. This list showed us that slice six alone had the best accuracy. The second best accuracy collection was constituted of the slices 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16.

In all previous experiments that needed a machine learning model to evaluate the accuracy of slices and collections, we used a small neural network composed of 2 convolutional layers, each convolutional layer followed by a max pooling layer, having its flattening and dense layer as the last layers. This model was used as a baseline to train and evaluate models and datasets in less time as the experiments needed to be repeated many times. However, for a final evaluation, we choose to experiment with state-of-the-art architecture, not the most recent but one still relevant nowadays, the VGG16 architecture. With the aid of the transfer-learning strategy, we built a custom model with this architecture and submitted the top 2 accuracy collections to it. The accuracy curves obtained with this architecture showed us that when using the slices from 0 to 16, this model could achieve an accuracy superior to 80%, given enough samples of images and epochs to train it.

Our last experiment was with Explainable Artificial Intelligence, XAI, to develop an insight into what our custom VGG16 model was looking for when doing its classification.

For the slice six only dataset, it focused on the center part of the image, and for the dataset composed of slices from 0 to 16, it observed the central part of slices and the back part of the brain.

So, based on our research, we can resume that:

1. Entropy is a metric of interest to evaluate and select the MRI slices to compose the images dataset.

2. VGG16 is still a good architecture model to be used as a basis for an image classification problem.

3. Slice 6 confronted us with an unexpected situation as it alone had a more excellent value of accuracy when experimenting with the baseline CNN model. Although it appears not to have any structure related to the schizophrenia problem, some pattern or relation has been detected by the CNN classifier, resulting in correctly labeling the predictions with reasonable accuracy even by the VGG16 classifier.

   We can hypothesize three situations from this observation.

   - There is an error in our experiment, some situation with the methodology used to prepare the dataset or the architecture, parameters, or kernel sizes that led to a construction error.
   - There is a hidden relation between slice six and the disease that is not known yet. As we noted that we have different sizes for human skulls, slice six can be slice 7 or 8 in another person's skull, and we need to develop a method to normalize the measurements and repeat the experiments with this normalized data.
   - AI can bring us two approaches when comparing the performance of an ML classifier to a human specialist: - Correctly trained with a large amount of data labeled by a specialist, an AI machine can outperform a human doing the same classification, opening the opportunity to automate that activity - An AI machine can discover patterns and relations that we never suspected existed, noticing details that a human being did not notice before or deliberately discarded as possibly irrelevant.

4. Although slice six only had the best accuracy in our baseline CNN model evaluation, the collection composed of slices from 0 to 16 achieved the best accuracy when submitted to the VGG16 custom model.

5. Previous studies have shown a preference to use slices from 9 to 16 when analyzing Schizophrenia studies, but our research showed that slices 0 to 16

5.1 Future Work

Suggestions for future work are:

- Submit the MRI Scan slice images to a Radiologist or Neurology specialist and discuss how to index the images and the initial reference to count them. This analysis by a specialist will lead to creating references for comparing slices from different subjects.
- Develop a method to normalize slices sizes and position for different skull sizes
- Expand the research using image slices from the sagittal and coronal axis
- Experiment with more recent ML architectures and strategies as Transformers.

References

[1] M. Abadi, A. Agarwal, P. Barham, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] American Psychiatric Association. and American Psychiatric Association. *Diagnostic and statistical manual of mental disorders : DSM-5.* American Psychiatric Association Arlington, VA, 5th ed. edition, 2013.

[3] R. N. Bryan. *Introduction to the science of medical imaging.* 01 2009.

[4] F. Chollet. Keras, python deep learning api running on top of machine learning platform tensorflow. `https://keras.io`, 2015. Accessed: 2021-03-23.

[5] T. M. Cover. *Elements of information theory / Thomas M. Cover, Joy A. Thomas.* Wiley-Interscience, Hoboken, N.J, 2nd ed. edition, 2006.

[6] B. F. da Cruz. Classificação de esquizofrenia com base em máquinas de suporte vetorial aplicadas a características de imagens de ressonância magnética. Master's thesis, 7, 2016. University of Brasília at Gama.

[7] L. Deng and D. Yu. Deep learning: Methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[8] G. Flores, J. C. Morales-Medina, and A. Diaz. Neuronal and brain morphological changes in animal models of schizophrenia. *Behavioural Brain Research*, pages 190–203, 2016.

[9] H. Fred, K. Minsuk, P. Robert, and C. D. Horng. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2674–2693, 2019.

[10] D. Gunning and D. W. Aha. Darpa's explainable artificial intelligence (XAI) program. *AI Mag.*, 40(2):44–58, 2019.

[11] P. J. Harrison. The neuropathology of schizophrenia: A critical review of the data and their interpretation. *Brain*, 122(4):593–624, 04 1999.

[12] Neuroimaging Informatics Technology Initiative. Nifti, analyze-style data format, proposed by the nifti dfwg to facilitate inter-operation of functional mri data analysis software packages. https://nifti.nimh.nih.gov/. Accessed: 2021-06-11.

[13] D. B. Keator, J. S. Grethe, D. Marcus, et al. A national human neuroimaging collaboratory enabled by the biomedical informatics research network (birn). *IEEE Trans. Inf. Technol. Biomed.*, 12(2):162–172, 2008.

[14] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. *Intl. Conf. Emerg. Trends Comput. Electron. Eng. (ICETCEE 2012)*, 1995.

[15] P. Lauterbur. Image formation by induced local interactions: Examples employing nuclear magnetic resonance. *Nature*, 242:190–191, 1973.

[16] Z.P. Liang, P.C. Lauterbur, IEEE Engineering in Medicine, and Biology Society. *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*. IEEE Press series in biomedical engineering. SPIE Optical Engineering Press, 2000.

[17] B. Matthew, M. Christopher, H. Michael, et al. nipy/nibabel: 3.2.1, November 2020.

[18] G. Michalakis, M. Pavlou, G. Gerogiannis, et al. Another day at the office: Visuo-haptic schizophrenia vr simulation. In *2020 IEEE Conf. on Virtual Reality and 3D User Interfaces Abstracts and Ws. (VRW)*, pages 515–516, 2020.

[19] R. Mizutani, R. Saiga, Y. Yamamoto, et al. Structural diverseness of neurons between brain areas and between cases, 2020.

[20] neurohive.io. Vgg16 – convolutional network for classification and detection. https://https://neurohive.io/en/popular-networks/vgg16/. Accessed: 2022-07-16.

[21] Y. Niu, Q. Lin, Y. Qiu, et al. Sample augmentation for classification of schizophrenia patients and healthy controls using ica of fmri data and convolutional neural networks. In *2019 Tenth International Conference on Intelligent Control and Information Processing (ICICIP)*, pages 297–302, 2019.

[22] J. Oh, B. Oh, K. Lee, et al. Identifying schizophrenia using structural mri with a deep learning algorithm. *Frontiers in Psychiatry*, 11:16, 2020.

[23] G. Rafael. Deep convolutional neural networks [lecture notes]. *IEEE Signal Processing Magazine*, 35(6):79–87, 2018.

[24] S. Ramprasaath, C. Michael, D. Abhishek, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[25] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.

[26] Siemens. Siemens MRI scanner magnetom free.max. https://www.siemens-healthineers.com/magnetic-resonance-imaging/high-v-mri/magnetom-free-max. Accessed: 2021-03-23.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2014.

[28] P. Sriramakrishnan, T. Kalaiselvi, T. Padmapriya, et al. A role of medical imaging techniques in human brain tumor treatment. 8:565–568, 01 2020.

[29] P. J. Sumner, I. H. Bell, and S. L. Rossell. A systematic review of the structural neuroimaging correlates of thought disorder. *Neuroscience and Biobehavioral Reviews*, pages 299–315, 2018.

[30] The Biomedical Informatics Research Network (BIRN) Neuroimaging tools & resources collaboratory (nitrc) website. Available at https://www.nitrc.org/projects/birn/.

[31] R. M. Tulio, S. Sameer, and G. Carlos. Model-agnostic interpretability of machine learning. 2016.

[32] R. F. Vergara. Detecção de alterações cerebrais anatômicas associadas à esquizofrenia com base em redes convolucionais aplicadas a imagens de ressonância magnética. Master's thesis, 3, 2019. University of Brasília at Gama.