UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Câmpus de Rio Claro

André Lara Temple de Antonio

# Combinação de características para recuperação e classificação de imagens médicas

Rio Claro - SP

2023

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

André Lara Temple de Antonio

# Combinação de características para recuperação e classificação de imagens médicas

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Orientador: Daniel Carlos Guimarães Pedronette

Rio Claro - SP
2023

Antonio, André Lara Temple de
A635c Combinação de características para recuperação e classificação de imagens médicas / André Lara Temple de Antonio. – Rio Claro, 2023
88 p. : il., tabs., fotos

Dissertação (mestrado) - Universidade Estadual Paulista (Unesp), Instituto de Geociências e Ciências Exatas, Rio Claro
Orientador: Daniel Carlos Guimarães Pedronette

1. Inteligencia artificial. 2. Aprendizado de máquinas. 3. Redes neurais (Computação). 4. Processamento de imagem assistida por computador. 5. Tumores cerebrais. I. Título.

UNIVERSIDADE ESTADUAL PAULISTA
"Júlio de Mesquita Filho"
Instituto de Geociências e Ciências Exatas
Câmpus de Rio Claro

André Lara Temple de Antonio

# Combinação de características para recuperação e classificação de imagens médicas

Dissertação de Mestrado apresentada ao Instituto de Geociências e Ciências Exatas do Câmpus de Rio Claro, da Universidade Estadual Paulista "Júlio de Mesquita Filho", como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Comissão Examinadora

Prof. Dr. Daniel Carlos Guimarães Pedronette – Orientador
IGCE/UNESP/Rio Claro (SP)

Prof. Dr. Marco Antonio Garcia de Carvalho
UNICAMP/Limeira (SP)

Prof. Dr. Denis Henrique Pinheiro Salvadeo
IGCE/UNESP/Rio Claro (SP)

Conceito: Aprovado.

Rio Claro (SP), 06 de Setembro de 2023.

# Agradecimentos

Agradeço a minha família, esposa e filhas, pela paciência e compreensão dos dias e noites que estive ausente para estudar, pesquisar e escrever este trabalho. Em um período atípico onde o mundo foi paralisado pela pandemia da COVID-19, tivemos a oportunidade de trabalhar e estudar remotamente, não deixando que o isolamento nos afastasse do conhecimento. Também agradeço aos professores e ao meu orientador, pelo direcionamento, apoio e esclarecimentos ao longo dessa jornada.

*"Pensar pequeno e pensar grande dá o mesmo trabalho,*
*mas pensar grande te liberta dos detalhes insignificantes."*
*(Jorge Paulo Leman)*

# Resumo

A evolução das tecnologias de aquisição e armazenamento de imagens tem sido fundamental em diversas áreas médicas, auxiliando na obtenção de diagnósticos mais precisos e, consequentemente, recomendarem tratamentos mais eficazes para seus pacientes. Recentemente, as técnicas de aprendizagem profunda têm desempenhado um papel fundamental na análise mais precisa de imagens médicas, principalmente devido à capacidade de representar efetivamente o conteúdo visual da imagem. No entanto, apesar dos enormes avanços, as técnicas de aprendizado profundo geralmente requerem grandes quantidades de dados para treinamento, que não estão disponíveis em muitos cenários, especialmente no domínio médico, além de não deixar transparente como o modelo chegou ao resultado. Por outro lado, várias técnicas de aprendizado de variedades foram aplicadas com sucesso em cenários não supervisionados e semi-supervisionados para codificação mais eficaz de relações de similaridade entre dados multimídia na ausência ou restrição de dados rotulados. Neste trabalho, propomos explorar conjuntamente o poder de representação de estratégias de aprendizado profundo com a capacidade de aprendizado não supervisionado de variedades na entrega de medidas de similaridade mais eficazes. Redes neurais convolucionais (CNNs) e modelos baseados em *Transformers* treinados por *transfer learning* são combinados por vários métodos de aprendizado não supervisionados, que definem uma similaridade mais efetiva entre as imagens. A saída pode ser usada para recuperação não supervisionada e classificação semi-supervisionada com base em uma estratégia kNN. Uma avaliação experimental foi realizada em diferentes conjuntos de dados de imagens de tumores cerebrais de ressonância magnética, considerando diferentes características. Resultados efetivos foram obtidos em tarefas de recuperação e classificação, com ganhos significativos obtidos por várias abordagens de aprendizado. Em cenários com dados de treinamento limitados, nossa abordagem alcança resultados competitivos ou superiores às abordagens de aprendizado profundo do estado-da-arte.

**Palavras-chave**: aprendizado não supervisionado, classificação, classificação kNN, fusão, extração de características, tumor cerebral, imagens médicas, ressonância magnética.

# Abstract

The evolution of image acquisition and storage technologies has been fundamental in numerous medical fields, supporting doctors to deliver more precise diagnoses and, consequently, recommend more effective treatments for their patients. Recently, deep learning techniques have played a key role in more accurate medical image analysis, mainly due to the capacity to effectively represent the image visual content. However, in spite of tremendous advances, deep-learning techniques commonly require huge quantities of data for training, that are not available in many scenarios, especially in the medical domain. Conversely, manifold learning techniques have been successfully applied in unsupervised and semi-supervised scenarios for more effective encoding of similarity relationships between multimedia data in the absence or restriction of labeled data. In this work, we propose to exploit jointly the representation power of deep-learning strategies with the ability of unsupervised manifold learning in delivering more effective similarity measurement. Convolutional Neural Networks (CNNs) and Transformer-based models trained through transfer learning are combined by unsupervised manifold learning methods, which define a more effective similarity among images. The output can be used for unsupervised retrieval and semi-supervised classification based on a kNN strategy. An experimental evaluation was conducted on different datasets of MRI brain tumor images, considering different features. Effective results were obtained on both retrieval and classification tasks, with significant gains obtained by manifold learning approaches. In scenarios with limited training data, our approach achieves results that are competitive or superior to state-of-the-art deep learning approaches.

**Keywords**: unsupervised learning, ranking, knn classification, fusion, feature extraction, brain tumor, medical images, MRI.

# Lista de ilustrações

# Lista de tabelas

# Lista de abreviaturas e siglas

| | |
|---|---|
| ACC | *Color Autocorrelogram* |
| ANN | *Approximate Nearest Neighbor* |
| AP | *Average Precision* |
| BIC | *Border/Interior Classification* |
| BOVW | *Bag of Visual Words* |
| BRIEF | *Binary Robust Independent Elementary Features* |
| CBIR | *Content-Based Image Retrieval* |
| CNN | *Convolutional Neural Networks* |
| CPRR | *Cartesian Product of Ranking References* |
| CT | *Computerized Tomography* |
| DL | *Deep Learning* |
| FLAIR | *Fluid Attenuated Inversion Recovery* |
| GCH | *Global Color Histogram* |
| GLCM | *Gray Level Co-occurrence Matrix* |
| GLN | *Gray Level Non-uniformity* |
| GLRLM | *Gray Level Run Length Matrix* |
| GWF | *Gabor wavelet features* |
| HOG | *Histogram of Oriented Gradients* |
| HGLRE | *High Gray Level Run Emphasis* |
| ISOMAP | *Isometric Mapping* |
| JAC | *Joint Autocorrelogram* |
| kNN | *k-Nearest Neighbor* |
| LAS | *Local Activity Spectrum* |

| | |
|---|---|
| LBP | *Local Binary Pattern* |
| LBGLCM | *Local Binary Gray Level Co-occurrence Matrix* |
| LGLRE | *Low Gray Level Run Emphasis* |
| LE | *Laplacian Eigenmaps* |
| LHRR | *Log-based Hypergraph of Ranking References* |
| LLE | *Local Linear Embedding* |
| LRE | *Long Run Emphasis* |
| MAP | *Mean Average Precision* |
| MDS | *Multidimensional Scaling* |
| ML | *Machine Learning* |
| MLP | *Multi Layer Perceptron* |
| MRI | *Magnetic Resonance Image* |
| NLP | *Natural Language Processing* |
| ORB | *Oriented Fast and Rotated BRIEF* |
| PCA | *Principal Component Analysis* |
| PCANet | *Principal Component Analysis Network* |
| PET | *Positron Emission Tomography* |
| PHOG | *Pyramid Histogram of Oriented Gradients* |
| QCCH | *Quantized Compound Change Histogram* |
| RDPAC | *Rank-Based Diffusion Process with Assured Convergence* |
| ReLU | *Rectified Linear Unit* |
| ResNet | *Residual Network* |
| RF | *Random Forest* |
| RGB | *Red, Green, Blue* |
| RI | Recuperação de Informações |
| RLN | *Run Length Non-uniformity* |

| | |
|---|---|
| RF | Rádio Frequência |
| RP | *Run Percentage* |
| SFTA | *Segmentation-based Fractal Texture Analysis* |
| SIFT | *Scale-Invariant Feature Transform* |
| SNC | Sistema Nervoso Central |
| SRE | *Short Run Emphasis* |
| SURF | *Speeded-Up Robust Features* |
| SVM | *Support Vector Machines* |
| TR | *Transformer* |
| TTBD | *Two-Threshold Binary Decomposition* |
| t-SNE | *t-Distribution Stochastic Neighbor Embeddings* |
| UDL | *Unsupervisioned Distance Learning* |
| UDLF | *Unsupervised Distance Learning Framework* |
| UMAP | *Uniform Manifold Approximation and Projection* |
| ViT | *Vision Transformer* |

# Sumário

# 1 Introdução

Em vista do grande aumento de coleções de imagens médicas, especialmente com dados não rotulados, torna-se necessário o desenvolvimento de técnicas e estratégias capazes de organizar e recuperar imagens baseadas em conteúdo visual. Rotular manualmente todas as imagens de uma determinada coleção pode ser uma tarefa inviável, principalmente pela grande quantidade de imagens e pela complexidade inerente ao conteúdo presente nas imagens médicas. Dessa forma, torna-se premente a investigação de novos métodos automatizados para realizar a recuperação e classificação das imagens com base no seu conteúdo.

Dentre as tarefas na área de reconhecimento de padrões, podemos destacar a análise de dados com base em tarefas de predição, ou seja, dado um conjunto de dados de treinamento, pretende-se realizar predições com amostras desconhecidas pelo modelo. Esta tarefa é conhecida como aprendizado supervisionado, onde existem uma grande quantidade de dados rotulados que são utilizados para treinamento do modelo de aprendizado, que busca aprender informações e relações entre as instâncias do conjunto de dados (MURPHY, 2013; GOODFELLOW; BENGIO; COURVILLE, 2016). Um exemplo representativo da aplicação dessa técnica foi apresentada para o reconhecimento de dígitos manuscritos de código postal dos correios dos EUA (LECUN et al., 1989), onde a rede de aprendizagem foi alimentada diretamente com imagens.

Contudo, nem sempre é possível obter rótulos para treinamento de uma grande coleção de dados, visto que essa tarefa exige um grande esforço humano para identificar e rotular cada instância, muitas vezes incompletos ou inexistentes. Diante deste cenário, uma solução de grande potencial consiste em empregrar métodos de aprendizado semi-supervisionado (CHAPELLE BERNHARD SCHöLKOPF, 2006), que combina uma pequena quantidade de dados rotulados com um grande número de dados não rotulados. Problemas de aprendizado dessa natureza são desafiadores, e residem entre aprendizado supervisionado e não supervisionado, pois nenhum desses são capazes de fazer uso efetivo dos dados disponíveis para resolver o problema.

Em um cenário ainda mais desafiador, em que não existem informações rotuladas para um conjunto de dados, é necessário utilizar o aprendizado não supervisionado, em que são exploradas as relações entre os elementos para identificar padrões ocultos nos dados sem a necessidade de intervenção humana, aplicando diversas técnicas e análises de similaridade ou distância (XU; WUNSCH, 2005). O agrupamento é um dos métodos de aprendizado não supervisionado mais utilizados, onde um conjunto de dados é particionado em grupos de elementos distintos com base nas relações disponíveis entre esses elementos.

Essa técnica é uma abordagem útil para problemas que não possuem dados de saída suficientes para treinamento (MURPHY, 2013).

De maneira semelhante, as técnicas de *manifold learning* ou aprendizado de variedades podem ser utilizadas como métodos de aprendizado não supervisionado, buscando explorar as estruturas de dados para aperfeiçoar as relações entre eles. Tais técnicas têm sido utilizadas em diversas tarefas de aprendizado de máquina e recuperação de informação. Existem diversas abordagens que podem ser aplicadas (DONOSER; BISCHOF, 2013), sendo que técnicas baseadas em ranqueamento visam explorar a relação de similaridade entre os elementos codificados em formato de listas ranqueadas, com o objetivo de calcular novas medidas de similaridade baseadas em relações mais globais do conjunto (PEDRONETTE et al., 2019).

Diante dos avanços significativos em técnicas de aprendizado de máquina nos últimos anos, abordagens baseadas em aprendizado profundo tem assumido um papel central. O aprendizado profundo é um conjunto relativamente novo de abordagens que transformaram radicalmente o aprendizado de máquina. O aprendizado profundo não é um algoritmo em si, mas uma série de algoritmos que implementam redes neurais com camadas profundas. A triagem de doenças assistida por aprendizado profundo e a previsão de resultados clínicos que não eram viáveis usando métodos anteriores, está sendo utilizada com bastante sucesso atualmente. De fato, quanto mais profunda e com mais dados de treinamento é uma rede neural, maior a precisão que ela pode produzir (YANG; YE; XIA, 2022).

Atualmente, as investigações radiológicas, independentemente da modalidade, requerem a interpretação de um médico experiente para se obter um diagnóstico em tempo hábil. Com o aumento das demandas sobre os médicos, há uma necessidade crescente de automação do diagnóstico. Este é um problema que o aprendizado profundo pode apresentar-se como uma ferramenta fundamental para suporte ao diagnóstico. (AGGARWAL et al., 2021).

Podemos observar que as técnicas e arquiteturas *Convolutional Neural Networks* (CNN) vem sendo amplamente aplicadas em diversas áreas de medicina, para apoiar os médicos no diagnóstico preciso através de classificação de imagens de raio-X, tomografia computadorizada (CT), imagens de ressonância magnética (MRI) entre outros. Em *Transformers* ainda observamos trabalhos iniciais na área médica, visto que se trata de técnicas mais recentes, mas ainda assim com trabalhos relevantes.

A aplicação das técnicas de aprendizado profundo tem sido utilizadas no apoio ao diagnóstico em diversas doenças e anormalidades, mas a falta de ferramentas para inspecionar o comportamento dos modelos caixa-preta impacta a sua aplicação na área médica, onde a confiabilidade são elementos-chave para o seu uso pelos médicos.

Neste trabalho, consideramos desafios comumente encontrados em aplicações reais em domínios médicos, onde os médicos têm como entrada imagens cerebrais de MRI e desejam analisá-las em comparação com imagens anteriores, na maioria sem rótulo, com o objetivo de obter suporte de ferramentas baseadas em aprendizado de máquina para um diagnóstico mais preciso da doença do tumor cerebral.

Dessa forma, as principais hipóteses que vamos analisar e discutir nesse trabalho são:

- Os métodos de aprendizado não supervisionado podem ser explorados para compor uma solução de suporte ao diagnóstico, modelados sob um arcabouço comum com foco em cenários de ausência ou restrição de dados rotulados;
- Características baseadas em aprendizado profundo podem ser utilizadas, mesmo em um cenário de transferência de aprendizado;
- A combinação de características distintas utilizando métodos de aprendizado não supervisionado contextuais pode incorporar informações complementares, aumentando a eficácia dos resultados.

Modelamos esse desafio como uma recuperação não supervisionada ou uma tarefa de classificação semi-supervisionada. Ambas as tarefas são abordadas com base em uma estrutura comum. Em primeiro lugar, extraímos as características da imagem do conjunto de dados, usando diferentes modelos de aprendizado profundo no estado-da-arte, treinados por meio de estratégias de *transfer learning*. Posteriormente, as listas ranqueadas obtidas para diferentes características são reclassificadas e combinadas por vários métodos de aprendizado não supervisionados. Essa estratégia permite combinar informações complementares de diferentes modelos de aprendizado profundo, considerando as informações de similaridade contextual codificadas no conjunto de dados de variedades. Os resultados de ranqueamento obtidos podem ser explorados para recuperação não supervisionada ou para classificação usando o algoritmo kNN.

As principais contribuições deste trabalho estão resumidas a seguir:

- Uma formulação baseada em similaridade é usada para derivar uma abordagem baseada em ranqueamento flexível para endereçar o problema desafiador da escassez de dados rotulados. A abordagem proposta pode ser usada tanto para recuperação de imagens não supervisionada quanto para classificação semi-supervisionada;
- O método proposto emprega técnicas de aprendizado não supervisionado de variedades, usadas para levar em consideração informações contextuais codificadas no conjunto de dados do *manifold*, a fim de melhorar os resultados de recuperação e classificação;

- Modelos recentes e distintos de aprendizado profundo treinados por *transfer learning* são usados para extração de características. Foram considerados modelos de CNN e baseados em *Transformers*.

Uma avaliação experimental foi realizada em dois conjuntos de dados de tumores cerebrais de MRI, que considerou várias características e aprendizado não supervisionado de variedades. Os resultados experimentais alcançaram resultados efetivos em tarefas de recuperação e classificação, indicando os ganhos significativos obtidos pelo aprendizado de variedades e o potencial para combinação de características.

O trabalho está organizado da seguinte forma:

- o Capítulo 2 define os aspectos teóricos fundamentais dentro do contexto da obra;
- o Capítulo 3 descreve as técnicas utilizadas para representação, recuperação, classificação e combinação de imagens com CNN, *Transformer*, kNN e aprendizado de variedades;
- o Capítulo 4 apresenta os modelos de extração de características de imagens *EfficientNet*, *RestNet* e ViT, e as técnicas de combinação de características através de aprendizado não supervisionado de variedades;
- o Capítulo 5 apresenta a coleção de imagens utilizadas no experimento deste trabalho, assim como o protocolo experimental aplicado e os resultados obtidos nos diversos cenários;
- o Capítulo 6 discute as considerações finais.

# 2 Fundamentação Teórica

Neste capítulo são definidos os principais aspectos teóricos relacionados às tarefas de recuperação e classificação de imagens, que constituem os temas centrais deste trabalho. Enquanto a Seção 2.1 descreve os principais conceitos de um sistema de recuperação de informações, a Seção 2.2 apresenta os principais métodos de aprendizado relevantes à pesquisa proposta, a Seção 2.3 explica as métricas para medir a eficácia e qualidade dos resultados obtidos, e por fim a Seção 2.4 apresenta os métodos de captura de imagens médicas.

## 2.1 Recuperação de Informação

Recuperação de Informação (RI) é uma área da computação que lida com o armazenamento de documentos e a recuperação automática de informação associada a eles. Um problema de recuperação de informações pode ser entendido como uma tarefa de busca de um determinado item a partir de um critério fornecido pelo usuário.

A Recuperação da Informação trata da representação, armazenamento, organização e acesso a itens de informação, como documentos, páginas Web, catálogos online, registros estruturados e semiestruturados, objetos multimídia etc. A representação e organização dos itens de informação devem fornecer aos usuários facilidade de acesso às informações de seu interesse (BAEZA-YATES, 2013).

Nesta seção explicamos sobre os Sistemas de Recuperação de Imagens por Conteúdo (CBIR), descritores de imagens globais e locais e métricas de eficácia.

### 2.1.1 Definição Formal

Seja $\mathcal{C} = \{\text{obj}_1, \text{obj}_2, ..., \text{obj}_n\}$ uma coleção de objetos, onde $N$ é o tamanho da coleção $\mathcal{C}$. Um descritor $\mathcal{D}$ pode ser definido como uma tupla $(\epsilon, \rho)$, onde $\epsilon : \hat{O} \rightarrow \mathbb{R}^m$ é uma função, que extrai o vetor de características $\text{v}_{\hat{O}}$ de um objeto em $\hat{O}$ e $\rho$: $\mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ é uma função de distância que calcula a distância entre dois objetos de acordo com a distância entre seus vetores de características. Sejam dois objetos $obj_i$ e $obj_j$, $\rho\left(\epsilon\left(obj_i\right), \epsilon\left(obj_j\right)\right)$ define a distância entre os mesmos. A notação $\rho\left(i, j\right)$ é utilizada para denotar a distância entre dois objetos $i$ e $j$.

De modo a se obter uma matriz quadrada de distâncias $A$, calcula-se a distância $\rho\left(i, j\right)$ entre todos os objetos $obj_i, obj_j \in \mathcal{C}$, tal que $A_{ij} = \rho\left(i, j\right)$. A matriz de distâncias $A$ é utilizada como entrada para vários métodos de aprendizado não supervisionados.

Uma forma de representar os resultados de recuperação de informações é realizada através das listas ordenadas de resultados ou listas ranqueadas (*ranked lists*). Uma lista ranqueada $\tau_q$ de um determinado objeto de consulta $obj_q$ pode ser calculada baseada na função de distância $\rho$. *Ranked lists* podem incluir informações de uma coleção inteira e em suas primeiras posições é esperado encontrar os objetos mais revelantes associados ao objeto de consulta.

Sendo assim, de modo a acelerar o processo de busca, uma possível estratégia realizada por alguns métodos consiste em considerar um subconjunto dos $L$ objetos mais similares, onde $L \leq N$ é o número de objetos nas primeiras posições das listas ranqueadas. Este tipo de estratégia se mostra extremamente útil para grandes coleções (com $N$ muito grande), em que calcular $\tau_q$ tem um custo considerável.

A *ranked list* $\tau_q = (\text{obj}_1, \text{obj}_2, ..., \text{obj}_L)$ pode ser definida como a permutação de uma coleção de objetos $\mathcal{C}_L \subset \mathcal{C}$, que contêm os objetos mais similares a um determinado objeto de busca $obj_q$, com $|\mathcal{C}_L| = L$. A permutação $\tau_q$ é uma bijeção do conjunto $\mathcal{C}_L$ sobre o conjunto $[L] = \{1, 2, ..., L\}$. Para uma permutação $\tau_q$, $\tau_q(i)$ é a posição (classificação ou *rank*) do objeto $obj_i$ na lista ranqueada $\tau_q$. Pode-se dizer que, se o objeto $obj_i$ é classificado antes do objeto $obj_j$ na *ranked list* do objeto $obj_q$, ou seja, $\tau_q(i) < \tau_q(j)$, então $\rho(q, i) \leq \rho(q, j)$. Se tomarmos cada objeto $obj_i \in \mathcal{C}$ como um objeto de busca $obj_q$, obtemos o conjunto $R = \tau_1, \tau_2, ..., \tau_n$ de listas ranqueadas para cada objeto da coleção $\mathcal{C}$.

### 2.1.2 Sistemas de Recuperação de Imagens pelo Conteúdo

Com o surgimento de novas tecnologias de captura e armazenamento de imagens digitais, observamos um grande aumento de dados disponíveis e a busca por imagens se tornou uma tarefa importante no contexto atual. O modelo tradicional de busca textual por palavras-chave apresenta algumas limitações, e como alternativa, surgiu o CBIR.

Os sistemas de CBIR tem como tarefa principal, encontrar imagens similares a uma dada consulta em uma coleção de imagens. Nesses sistemas, surgiram diversos tipos de desafios. Como as imagens são arquivos grandes, temos a questão de armazenamento do grande volume de dados, assim como a sua recuperação e processamento.

A principal virtude da CBIR está na capacidade de considerar as características visuais das imagens no processo de indexação e recuperação. As características visuais mais comuns de serem analisadas são a cor, textura e forma dos objetos. Por meio dessa abordagem, o processo de busca de imagens não é afetado pela ausência de descrições textuais, ou pela existência de descrições subjetivas e ambíguas. Dessa forma, a busca por conteúdo permite encontrar imagens relevantes, que não seriam encontradas por mecanismos de busca baseados em descrições textuais.

A Figura 1 apresenta a implementação clássica de um sistema CBIR (TORRES;

FALCÃO, 2006). Esta pode ser dividida em três partes: (*i*) a interface, (*ii*) o módulo de processamento da consulta e (*iii*) a base de dados. A interface é responsável pela interação com o usuário, com a inserção e saída dos dados. O módulo de processamento da consulta tem como tarefa extrair as características das imagens e fazer as operações necessárias de modo a realizar as comparações entre elas. Por fim, a base de dados armazena as imagens e seus respectivos vetores de características, de modo a evitar a recalculá-los.

Há diversas formas de se calcular a distância entre dois pontos dos vetores de características de cada imagem. Algumas delas são apresentadas a seguir.

- **Euclidiana**

$$d\left(p,q\right) = \sqrt{\sum_{i=1}^{n}\left(p_i - q_i\right)^2}, \tag{2.1}$$

onde $p$ e $q$ são dois pontos do vetor de características.

- **Chebysev**

$$d\left(p,q\right) = \max_{i=1}^{n}\left|p_i - q_i\right|, \tag{2.2}$$

onde $p$ e $q$ são dois pontos do vetor de características.

- **Manhattan**

$$d\left(p,q\right) = \sum_{i=1}^{n}\left|p_i - q_i\right|, \tag{2.3}$$

onde $p$ e $q$ são dois pontos do vetor de características.

- **Minkowski**

$$d\left(p,q\right) = \left(\sum_{i=1}^{n}\left|x_i - y_i\right|^p\right)^{1/p}, \tag{2.4}$$

onde $p$ e $q$ são dois pontos do vetor de características.

- **Canberra**

$$d\left(p,q\right) = \sum_{i=1}^{n}\frac{\left|p_i - q_i\right|}{\left|p_i\right| + \left|q_i\right|}, \tag{2.5}$$

onde $p$ e $q$ são dois pontos do vetor de características.

Em geral, a maioria dos sistemas CBIR utilizam da distância Euclidiana (PATIL; TALBAR, 2012) para o cálculo das distâncias. O cálculo das distâncias é essencial para uma comparação eficaz entre os elementos da base de dados, porém, os descritores e as extrações de características são os elementos que mais impactam a eficácia de um sistema CBIR como um todo.

### 2.1.3 Descritores de Imagens

Um dos principais componentes de um CBIR é o descritor de imagens. O descritor de imagens é responsável por quantificar o quão semelhante são duas imagens (TORRES; FALCÃO, 2006). Ele pode ser caracterizado por dois componentes:

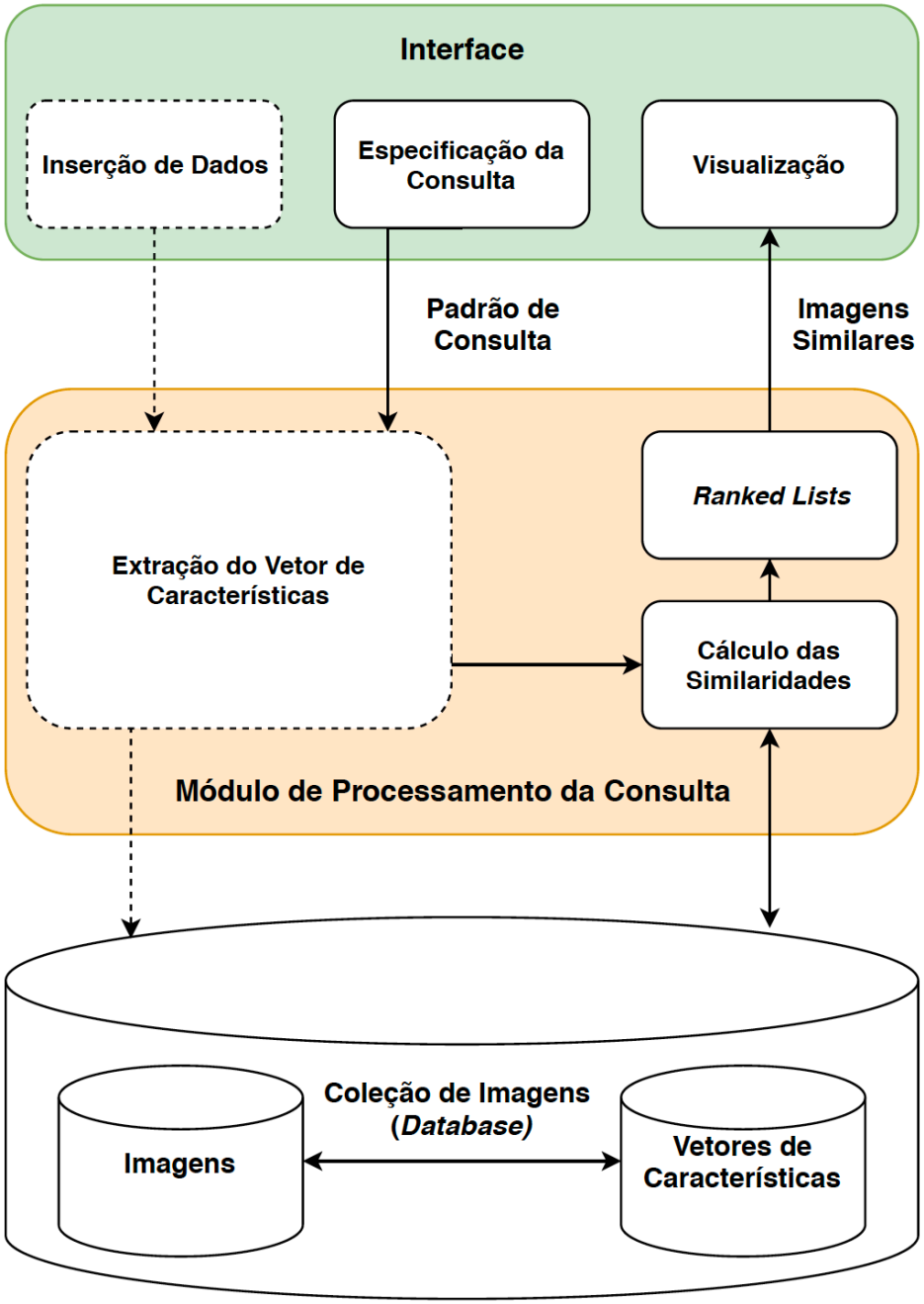



**Figura 1** – Arquitetura típica de um sistema CBIR (TORRES; FALCÃO, 2006)

i. Algoritmo responsável por realizar a extração das propriedades visuais de uma dada imagem de entrada em um vetor de características;

ii. Função para comparação dos vetores de características através de cálculo da distância

entre os vetores. Em geral, funções de distância tradicionais como Euclidiana e Manhattan são comumente utilizadas para comparação dos vetores de características.

As aplicações de processamento de imagens necessitam de um robusto detector de características. As informações das imagens são utilizadas para caracterizar a aparência e formatos das imagens são classificadas como descritores globais ou locais, e podem representar contornos, bordas, pontos, arestas, texturas, cores entre outros (AWAD, 2016).

Uma visão geral das principais categorias são apresentadas a seguir.

- **Globais**

Os sistemas de reconhecimento de imagens utilizam detectores de características globais que descrevem uma imagem inteira. A maioria dos descritores de formato e textura são dessa categoria. Essas caraterísticas produzem representações simples e compactas das imagens, onde cada imagem corresponde a um ponto no espaço de características de alta dimensionalidade. Apesar de ser uma abordagem de extração de caraterísticas simples e rápida, descritores globais são pouco adequados para localizar detalhes nas imagens, além de serem pouco robustos a deformações nas imagens.

Alguns exemplos de descritores globais são: *Global Color Histogram* (GCH) (MEHTRE et al., 1995), *Border/Interior Classification* (BIC) (STEHLING; NASCIMENTO; FALCãO, 2002), *Color Autocorrelogram* (ACC) (MIKOLAJCZYK; SCHMID, 2005), *Joint Autocorrelogram* (JAC) (WILLIAMS; YOON, 2007), *Local Activity Spectrum* (LAS) (TAO; DICKINSON, 2000) e *Quantized Compound Change Histogram* (QCCH) (HUANG; LIU, 2007).

- **Locais**

Um paradigma diferente é utilizar detectores de características locais que são descritores vizinhos de uma imagem, computada em múltiplos pontos de interesse (LISIN et al., 2005).

Os descritores locais são computados sobre as características locais das imagens como regiões, fronteiras ou pontos de interesse. A principal ideia dessa abordagem está em analisar diferentes imagens procurando pontos-chave que sejam de alta relevância para descrever as imagens, e que sejam também invariantes, esses pontos são partes das imagens que podem possuir tamanhos distintos. Essas partes são utilizadas para construir um dicionários de palavras visuais (*bag of visual words* BOVW) (YANG et al., 2007). Um vetor de características pode ser construído, considerando a ocorrência de determinados elementos do dicionário.

Os principais exemplos de descritores locais são: *Scale-Invariant Feature Tranform* (SIFT) (LOWE, 2001), *Speeded Up Robust Features* (SURF) (BAY; TUYTELAARS; GOOL, 2006), *Binary Robust Independent Elementary Features* (BRIEF) (CALONDER

et al., 2010) e *Oriented Fast and Rotated BRIEF* (ORB) (RUBLEE et al., 2011).

- **Baseados em Aprendizado Profundo**

Métodos de recuperação de imagens baseados em aprendizado profundo tem sido propostos nos últimos anos, e tem substituído gradualmente as técnicas manuais de descritores locais e globais. Conforme proposto por (ZHENG; YANG; TIAN, 2017), diferentes descritores locais podem ser combinados para aprimorar a eficácia das redes neurais.

As redes de aprendizado profundo são geralmente treinadas de maneira supervisionada, e o vetor de pesos é utilizado como vetor de características para uma determinada imagem. As redes neurais convolucionais CNN são comumente utilizadas para essa tarefa, conforme será apresentado na Seção 3.2.1.

Os modelos de recuperação baseados em CNN computam representações compactas e empregam a distância Euclidiana ou métodos de busca por aproximação de vizinhos mais próximos (*Approximate Nearest Neighbor* ANN). A literatura atual pode empregar diretamente os modelos de CNN pré-treinados, ou realizar *fine-tunnings* para as tarefas de recuperação, sendo que a maioria desses métodos alimenta a imagem na rede apenas uma vez para obter o descritor.

## 2.2 Aprendizado de Máquina

*Machine learning* (ML) ou aprendizado de máquina é definido como um conjunto de métodos que podem detectar automaticamente padrões nos dados, e depois usar os padrões descobertos para prever dados futuros, ou executar outros tipos de decisão feitas sobre incertezas (MURPHY, 2013).

O aprendizado de máquina é normalmente dividido entre três tipos principais: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. Também existe um quarto tipo conhecido como aprendizado semi-supervisionado, que reside entre aprendizado supervisionado e não supervisionado, sendo esses últimos os mais comuns.

### 2.2.1 Aprendizado Supervisionado

Aprendizado supervisionado é uma abordagem de ML que é definido pelo uso de dados rotulados. Os *datasets* são criados para treinar os algoritmos que são utilizados para classificação de novos dados ou para previsão de resultados com precisão (MURPHY, 2013; GOODFELLOW; BENGIO; COURVILLE, 2016).

Nessa abordagem temos variáveis de entrada ($x$) e variáveis de saída ($y$) e podemos utilizar um algoritmo para aprender a função de mapeamento entre a entrada e a saída. O

objetivo do aprendizado supervisionado é aproximar a função de mapeamento $y = f(x)$ tão bem que quando tivermos um novo dado de entrada $(x)$, podemos prever a variável de saída $(y)$ para aquele dado (MURPHY, 2013).

Os problemas de aprendizado supervisionado podem ser agrupados em problemas de classificação e regressão.

- Classificação tem como objetivo aprender a mapear as entradas $(x)$ em saídas $(y)$, onde $y \in 1 \ldots C$, e $C$ é dado como o número de classes. Se $C = 2$, então chamamos de classificação binária, e se $C > 2$, então chamamos de classificação multi-classe. Caso a classe de saída não for mutuamente exclusiva, então chamamos de classificação multi-rótulo. A classificação pode ser utilizada para reconhecer padrões e determinar os rótulos de imagens em uma rede neural pré-treinada. Alguns algoritmos de classificação são *Support Vector Machines* (SVM), árvores de decisão, *K-nearest neighbor*, *Random Forest* e redes neurais.

- Regressão busca prever o valor de uma ou mais variáveis contínuas $t$ dado o valor de um vetor $D - dimensional$ $x$ de variáveis de entrada (BISHOP, 2006). Nesse problema temos um valor de entrada $x_i \in \mathbb{R}$, e um único valor de saída $y_i \in \mathbb{R}$. A regressão pode ser utilizada, por exemplo, para prever o risco de determinada doença ocorrer, com base em determinadas características do paciente. Os algoritmos de regressão mais utilizados são regressão linear, regressão logística e regressão polinomial.

### 2.2.2 Aprendizado Não Supervisionado

Aprendizado não supervisionado utiliza algoritmos de ML para analisar e agrupar *datasets* não rotulados. Estes algoritmos descobrem padrões escondidos nos dados sem a necessidade de intervenção humana (MURPHY, 2013; GOODFELLOW; BENGIO; COURVILLE, 2016).

Essa abordagem é aplicada quando temos somente os dados de entrada $(x)$ e não temos as variáveis de saída correspondentes. O objetivo do aprendizado não supervisionado é modelar a estrutura ou distribuição para aprender mais sobre os dados avaliados. São chamados de aprendizado não supervisionado porque, ao contrário do aprendizado supervisionado, não há resposta certa, sendo que os algoritmos são deixados para descobrir e apresentar padrões nos dados. Essa técnica é uma abordagem útil para problemas que não possuem dados de saída suficientes para treinamento (MURPHY, 2013).

Os modelos de aprendizado não supervisionado são utilizados em três tipos de tarefas: agrupamento (*clustering*), associação e redução de dimensionalidade.

- Agrupamento é uma técnica para agrupar dados não rotulados baseado em suas similaridades e diferenças. Como exemplo, algoritmos de agrupamento *K-means* atribuem pontos similares em grupos, onde o valor $K$ representa o tamanho do agrupamento ou granularidade. Esta técnica pode ser útil, por exemplo, em tarefas de agrupamento de imagens não rotuladas.

- Associação é outro tipo de método de aprendizado não supervisionado que usa diferentes regras para encontrar relações entre variáveis em um conjunto de dados. Este método é frequentemente utilizado para análise do padrões e motores de recomendação.

- Redução de dimensionalidade é uma técnica de aprendizado utilizada quando o número de características ou dimensões em um dado conjunto de dados é muito grande. Dessa forma, reduz o número de dados para um tamanho gerenciado enquanto preservar a integridade dos dados e de suas características, conforme será apresentado no Capítulo 3.4.

### 2.2.3 Aprendizado Semi-Supervisionado

Aprendizado semi supervisionado é um problema que combina um pequeno número de dados rotulados com um grande número de dados não rotulados. Problemas de aprendizado dessa natureza são desafiadores, e residem entre aprendizado supervisionado e não supervisionado, pois nenhum desses são capazes de fazer uso efetivo dos dados disponíveis para resolver o problema (CHAPELLE BERNHARD SCHöLKOPF, 2006).

Um dos principais objetivos desse tipo de aprendizado é solucionar cenários onde existem poucos dados rotulados, nos quais geralmente os métodos supervisionados não são eficazes (KUNCHEVA, 2004).

Quando estamos trabalhando com problemas de aprendizado semi-supervisionado, o objetivo varia dependendo do tipo de problemas que queremos resolver. Podemos utilizar aprendizado semi-supervisionado para aprendizado indutivo, transdutivo ou os dois juntos.

O objetivo do aprendizado indutivo é generalizar os dados novos. Dessa forma, nesse tipo de técnica, construímos algoritmos que aprendem com o pequeno conjunto de dados rotulados, e generalizam para novos dados. Quando utilizamos o aprendizado transdutivo, temos como objetivo transferir informações de conjuntos de dados de treinamento rotulados para os dados não rotulados.

Podemos considerar um problema de aprendizado de dados rotulados e não rotulados. Dado um conjunto de dados $x = \{x_1, ..., x_l, x_{l+1}, ..., x_n\}$ e um conjunto de rótulos $L = \{1, ..., c\}$, sendo que os primeiros $l$ dados possuem os rótulos $\{y_1, ..., .y_l\} \in L$ e os dados restantes são não rotulados. O objetivo é prever os rótulos dos dados não rotulados.

Este tipo de tarefa é um exemplo de aprendizado semi-supervisionado. Visto que a rotulagem de dados requer um trabalho humano intenso e caro, enquanto os dados não rotulados são mais fáceis de se obter, a aplicação do aprendizado semi-supervisionado é muito útil nessas situações.

## 2.3 Métricas de Eficácia

Métricas de eficácia são fundamentais para avaliar a qualidade dos resultados de busca (BAEZA-YATES, 2013). Nesta subseção são apresentadas as métricas abordadas neste trabalho. Todas as métricas requerem dados rotulados. Os resultados são obtidos no intervalo $[0, 1]$ e quanto maior o valor, melhor o resultado.

- **Acurácia**

Acurácia é a medida mais intuitiva de eficácia, sendo o número de previsões corretas sobre o tamanho total da saída e informa o quanto o seu modelo está acertando, sendo *tp (true positives)*, *tn (true negatives)*, *fp (false positives)* e *fn (false negatives)*.

$$A = \frac{(tp + tn)}{(tp + fp + tn + fn)}, \tag{2.6}$$

- **Precisão**

A precisão pode ser entendida como a fração de *tp (true positives)* pela soma dos *tp (true positives)* e *fp (false positives)*, sendo calculada como:

$$P_n = \frac{tp}{(tp + fp)}, \tag{2.7}$$

onde $n$ é o número de itens recuperados. A precisão é intuitivamente a habilidade do classificador não rotular uma amostra como positiva, sendo que ela é negativa.

- ***Recall***

Diferente da precisão, o *recall* é a fração de *tp (true positives)* pela soma dos *tp (true positives)* e *fn (false negatives)*, sendo definida como:

$$R_n = \frac{tp}{(tp + fn)}, \tag{2.8}$$

onde $n$ é o número de itens recuperados. O *recall* é intuitivamente a habilidade do classificador encontrar todas as amostras positivas.

- **MAP**

O MAP (*Mean Average Precision*) é a métrica mais comum para medir a eficácia de listas ranqueadas em tarefas de recuperação. Para cada uma das listas ranqueadas do

conjunto $\mathcal{R}$, pode-se obter a precisão média $AP$ (*Average Precision*). A precisão e o *recall* são calculadas em cada uma das posições da lista ranqueada, dando origem a um gráfico que descreve a função $P(R)$, onde a precisão é fornecida em função do *recall*. A precisão média considera o valor médio de $P(R)$ no intervalo $[0, 1]$ e é definida como:

$$AP = \int_0^1 P(R)dr \tag{2.9}$$

Ao se realizar a média ponderada da precisão média de cada uma das listas ranqueadas do conjunto $\mathcal{R}$, obtêm-se o MAP.

## 2.4 Imagens Médicas

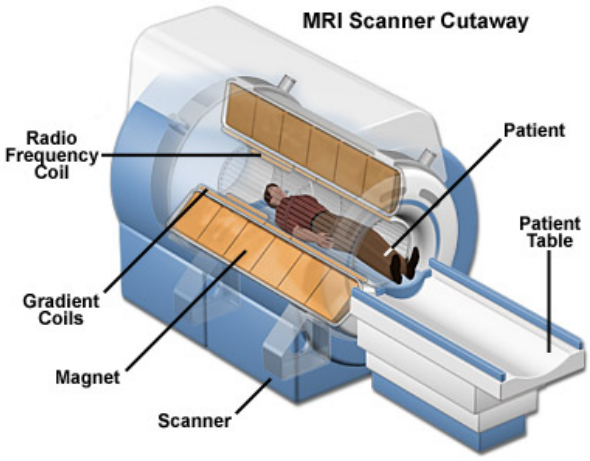
Nessa seção apresentamos as diversas técnicas de captura de imagens e os seus tipos relacionados que são utilizadas para apoiar o diagnóstico médico na prevenção e no tratamento de diversas doenças e anormalidades.

Na avaliação experimental do Capítulo 5 utilizamos a Ressonância Magnética (MRI), por se tratar de uma técnica mais utilizada no espectro de estudos das imagens da doença que foi escolhida para os experimentos. Apesar disso, trazemos uma visão geral das técnicas de captura de imagens que podem ser utilizadas para outras doenças e anormalidades.

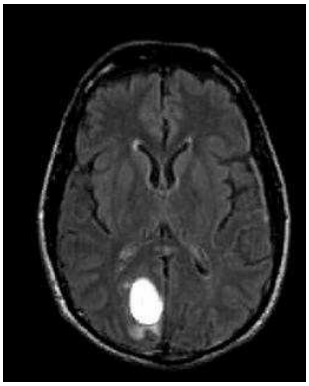
### 2.4.1 Ressonância Magnética (MRI)

A Ressonância Magnética (MRI) é uma tecnologia de diagnóstico de imagem que utiliza campos magnéticos e de radiofrequência para criar imagens dos tecidos do corpo humano, sendo mais comumente utilizada em neurologia e neurocirurgia, fornecendo detalhes extraordinários da anatomia do cérebro. A MRI se baseia na capacidade de detectar mudanças na densidade de prótons e nos tempos de relaxamento do *spin* magnético, que são características presentes nos tecidos doentes.

Um *scanner* de ressonância magnética consiste de um imã principal que gera um campo magnético, um sistema de gradiente de campo magnético que consiste, normalmente, de três bobinas de gradiente ortogonal, utilizados para a localização do sinal, e um sistema de rádio frequência (RF) com uma bobina transmissora capaz de gerar um campo magnético giratório para excitar o sistema de *spin*, e uma segunda bobina receptora que converte a magnetização anterior em um sinal elétrico. A Figura 2a ilustra as partes do *scanner* de ressonância magnética. A Figura 2b apresenta um exemplo de imagem de cérebro obtida por MRI.




**(a)** Equipamento de Ressonância Magnética (COYNE, 2021)



**(b)** MRI de cérebro com tumor maligno (LAKSHMI; ARIVOLI, 2014)

**Figura 2** – Equipamento de Ressonância Magnética e exemplo de imagem de cérebro com tumor maligno.

A MRI é baseada nas propriedades de magnetização dos núcleos atômicos. Um campo magnético externo poderoso e uniforme é empregado para alinhar os prótons que normalmente são orientados aleatoriamente dentro dos núcleos de água do tecido que está sendo examinado. Este alinhamento (ou magnetização) é em seguida perturbado ou interrompido pela introdução de uma energia externa de radiofrequência (RF). Os núcleos retornam ao seu alinhamento de repouso através de vários processos de relaxamento e, ao fazê-lo, emitem energia de RF (VLAARDINGERBROEK, 1999).

Após um certo período da RF inicial, os sinais emitidos são medidos. A transformada de Fourier é usada para converter as informações de frequência contidas no sinal de cada local no plano de imagem em níveis de intensidade correspondentes, que são então exibidos como tons de cinza em um arranjo matricial de *pixels*. Variando a sequência de pulsos de RF aplicados e coletados, diferentes tipos de imagens são criados. Tempo de repetição (*Repetition Time* - TR) é a quantidade de tempo entre sequências de pulso sucessivas aplicadas à mesma fatia. O tempo até o eco (*Time to Echo* - TE) é o tempo entre a entrega do pulso de RF e o recebimento do sinal de eco.

O tecido pode ser caracterizado por dois tempos de relaxamento diferentes – T1 e T2. T1 (tempo de relaxamento longitudinal) é a constante de tempo que determina a taxa na qual os prótons excitados retornam ao equilíbrio. É uma medida do tempo que leva para os prótons em rotação se realinharem com o campo magnético externo. T2 (tempo de relaxamento transversal) é a constante de tempo que determina a taxa na qual os prótons excitados atingem o equilíbrio ou saem de fase entre si. É uma medida de tempo que leva para os prótons em rotação perderem a coerência de fase entre os núcleos girando perpendicularmente ao campo principal.

As sequências de MRI mais comuns são as ponderadas em T1 e T2. Imagens ponderadas em T1 são produzidas usando tempos TE e TR curtos. O contraste e o brilho da imagem são predominantemente determinados pelas propriedades T1 do tecido. Por outro lado, imagens ponderadas em T2 são produzidas usando tempos TE e TR mais longos. Nestas imagens, o contraste e o brilho são predominantemente determinados pelas propriedades T2 do tecido.

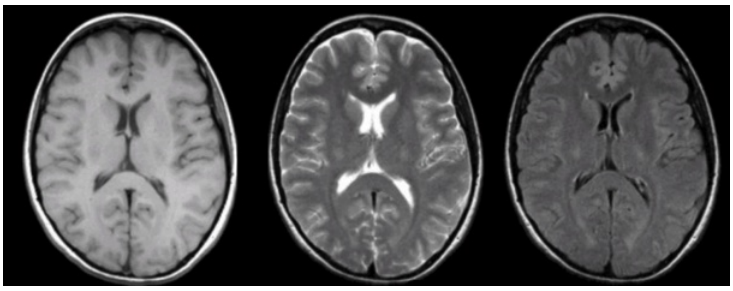


**Figura 3** – Da esquerda para a direita: Imagens de Tumor Cerebral ponderada em T1, T2 e FLAIR.
(ATIA et al., 2022)

Em geral, as imagens ponderadas em T1 e T2 podem ser facilmente diferenciadas pela observação do CSF (*Cerebral Spinal Fluid*). O CSF é escuro nas imagens ponderadas em T1 e claro nas imagens ponderadas em T2.

Imagens ponderadas em T1 também podem ser realizadas durante a infusão de gadolínio (Gad). Gad é um agente de aumento de contraste paramagnético não tóxico. Quando injetado durante a varredura, o Gad altera a intensidade do sinal encurtando T1. Assim, Gad é muito brilhante em imagens ponderadas em T1. Imagens aprimoradas por Gad são especialmente úteis na observação de estruturas vasculares e rupturas na barreira hematoencefálica, como por exemplo, tumores, abscessos, inflamação, etc.

Uma terceira sequência comumente usada é a *Fluid Attenuated Inversion Recovery* (FLAIR). A sequência FLAIR é semelhante a uma imagem ponderada em T2, exceto que os tempos TE e TR são muito longos, conforme apresentado na Figura 3. Ao fazer isso, as anormalidades permanecem brilhantes, mas o líquido normal do CSF é atenuado e escurecido. Esta sequência é muito sensível à patologia e facilita muito a diferenciação entre CSF e uma anormalidade.

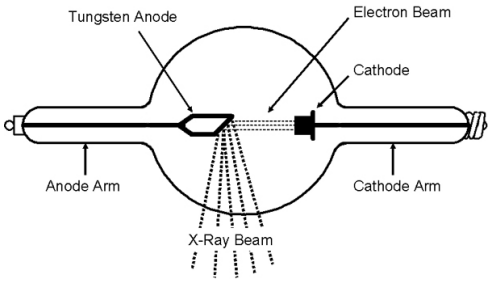
### 2.4.2 Outros Tipos de Imagens

#### 2.4.2.1 Raio-X

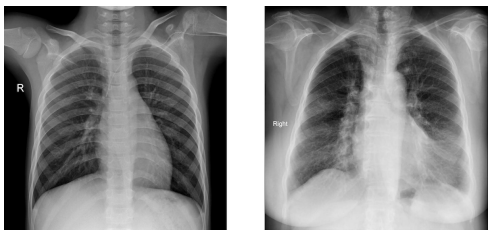
A radiografia é uma técnica utilizada como diagnóstico de imagem que usa radiação eletromagnética ionizante, similar ao raio-X utilizado para visualizar objetos. O raio-X é uma radiação eletromagnética de alta energia, que pode penetrar sólidos e ionizar gases.

Podemos observar na Figura 4a uma ilustração explicativa do funcionamento do tubo de raio-X.

Quando utilizada para imagens médicas, o raio-X passa através do corpo, que é absorvido ou atenuado em diferentes níveis, de acordo com a densidade e número atômico dos diferentes tecidos, criando um perfil. Esse perfil é registrado em um detector, criando a imagem relacionada, conforme apresentado na Figura 4b.




**(a)** Tubo de Raio-X (TEKIN; KARA, 2016)



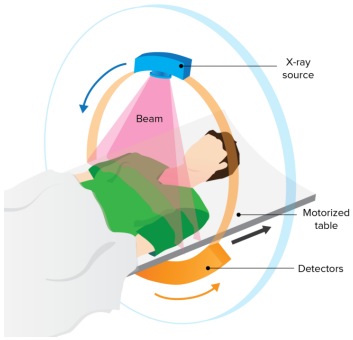
**(b)** Esquerda: pulmão saudável, direita: pulmão com COVID-19 (HAQUE; ABDELGAWAD, 2020)

**Figura 4** – Funcionamento de um tubo de raio-X, e exemplo de imagens de raio-X de pulmão saudável e com COVID-19.

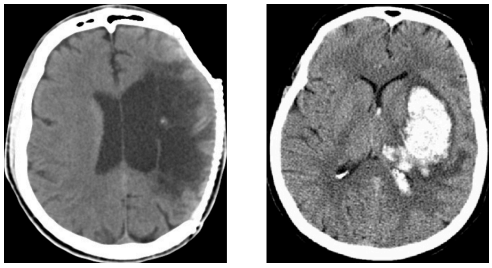
#### 2.4.2.2 Tomografia Computadorizada

Tomografia Computadorizada (*Computed Tomography*) CT é uma tecnologia de imagem que combina um equipamento de raio-X com um computador, e um *display* de tubo de raios catódicos para produzir imagens de seções transversais do corpo humano.

O filme radiográfico é substituído por um detector que mede o perfil de raios-X. Dentro do tomógrafo, há uma estrutura giratória que possui um tubo de raios-X montado em um lado e o detector montado no lado oposto. Um feixe de raios-X é gerado enquanto uma estrutura rotativa gira o tubo de raios-X e o detector ao redor do paciente. Cada vez que o tubo de raio-X e o detector realizar uma rotação completa, uma imagem ou *slice* são capturados. A Figura 5a apresenta o esquema de tomografia computadorizada.




**(a)** Esquema de Tomografia Computadorizada (OISETH LINDSAY JONES, 2021)



**(b)** Esquerda: AVC Isquêmico (área preta escura), direita: AVC Hemorrágico (área branca) (CHAWLA et al., 2009)

**Figura 5** – Esquema de Tomografia Computadorizada, e exemplo de imagens de CT de AVC.

Conforme o tubo de raios-X e o detector fazem essa rotação, o detector obtém vários perfis do feixe de raios-X atenuado. Cada perfil é reconstruído pelo computador em uma imagem 2D do *slice* que foi digitalizado.

A tomografia computadorizada 3D pode ser criada usando um CT espiral, que obtêm um conjunto de dados da anatomia do paciente em uma única posição. Esse conjunto de dados pode ser reconstruído em computador para obter as imagens tridimensionais de estrutura mais complexas. A Figura 5b apresenta um exemplo de imagens de AVC isquêmico e hemorrágico obtido através de uma CT.

#### 2.4.2.3 Ultrassonografia

A ultrassonografia é uma tecnologia de diagnóstico de imagem que utiliza ondas sonoras de banda larga em alta frequência que são refletidas pelo tecido humano em vários graus, para produzir as imagens médicas.

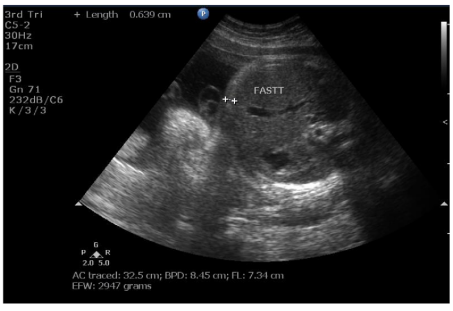



**Figura 6** – Imagem de ultrassonografia fetal (KHALIFA; HASSANEIN; EID, 2019)

O transdutor de ultrassom é colocado contra a pele do paciente próximo à região de interesse. O transdutor produz um fluxo de ondas sonoras de alta frequência que penetram

no corpo e refletem nos órgãos internos. O transdutor detecta as ondas sonoras à medida que ecoam nas estruturas internas dos órgãos.

Diferentes tecidos refletem as ondas sonoras de forma diferente, resultando em uma assinatura que pode ser medida e transformada em uma imagem. Essas ondas são recebidas pela máquina de ultrassom e transformadas em imagens ao vivo. A Figura 6 apresenta uma imagem de ultrassonografia fetal.

#### 2.4.2.4 Tomografia por Emissão de Pósitrons

A imagem por radionuclídeo ou medicina nuclear é uma tecnologia de diagnóstico que usa pequenas quantidades de material radioativo para produzir imagens do corpo interno.

Pequenas quantidades de isótopos radioativos de baixo nível são administradas por injeção ou por via oral. Esses isótopos são atraídos para os órgãos, ossos ou tecidos específicos, que absorvem o material radioativo. Uma vez que um órgão ou tecido tenha absorvido o material radioativo, ele produz emissões, que podem ser detectadas por detectores de radiação especiais. O *scanner* funciona com um computador para converter as emissões em uma imagem.

*Positron Emissions Tomography* (PET) é uma técnica de imagem por radionuclídeo, que fornece informações sobre o metabolismo de uma doença. Os isótopos usados em imagens PET são degradados pela emissão de pósitrons. O pósitron emitido viaja apenas uma distância mínima antes de sofrer uma reação de aniquilação com a produção de dois fótons que viajam em direções opostas um ao outro. A localização do evento de aniquilação é obtida colocando-se dois detectores em lados opostos do paciente. Quando os fótons são detectados ao mesmo tempo, a posição do pósitron emitido pode ser rastreada de volta em uma linha reta. A Figura 7 apresenta uma imagem do cérebro de um paciente com Alzheimer, obtida através de PET.

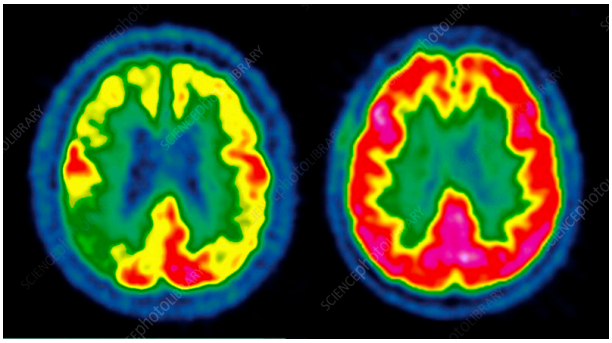


**Figura 7** – Imagem obtida com PET de um paciente com Alzheimer (PERRIN, 2021)

Recentemente, a fusão de imagens foi usada para combinar imagens PET com imagens de CT, ou com MRI, para produzir visualizações especiais capturando informações

de dois exames diferentes para serem correlacionados e interpretados em uma imagem única. Essa técnica fornece informações e diagnósticos mais precisos.

#### 2.4.2.5 Imagem Óptica

A imagem óptica é uma tecnologia não invasiva que utiliza a luz para mostrar a função celular e molecular do corpo vivo. A imagem óptica é considerada uma ferramenta poderosa para sondar tecidos profundos, onde a luz se propaga de maneira difusa.

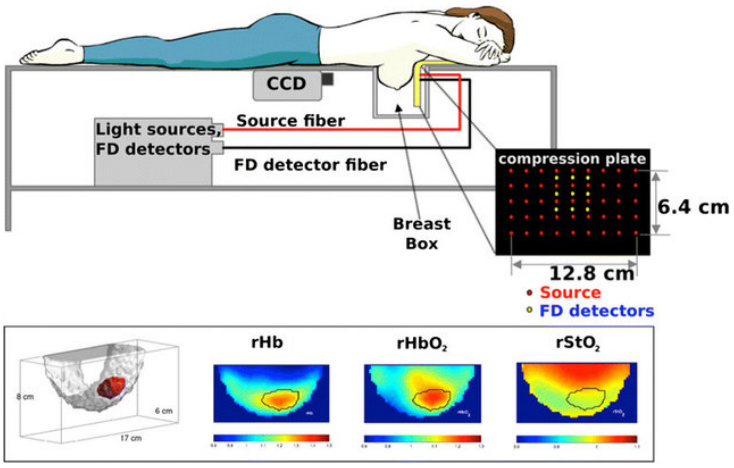



**Figura 8** – Esquema de exame de mamografia (CHUNG et al., 2015)

As informações são derivadas da composição do tecido e dos processos biomoleculares. O contraste é derivado tanto do uso de agentes exógenos que fornecem o sinal ou de moléculas endógenas com assinaturas ópticas. A interação da luz propagada de forma difusa com diferentes componentes do tecido, permite a visualização de anormalidades do tecido ou de processos patológicos.

As imagens ópticas reduzem significativamente a exposição do paciente à radiação prejudicial ao usar radiação não ionizante, que inclui luz visível, ultravioleta e infravermelha. Por ser muito mais seguro do que as técnicas que requerem radiação ionizante, como os raios-X, a imagem óptica pode ser usada para procedimentos repetidos para monitorar a progressão de doenças ou os resultados do tratamento.

Pode ser combinada com outras técnicas de imagem, como MRI ou raios-X, para fornecer informações aprimoradas para médicos que monitoram doenças complexas. Vemos na Figura 8 uma esquema de exame de mamografia utilizando imagem óptica.

# 3 Técnicas de representação, recuperação e classificação de imagens

Apresentamos neste capítulo as técnicas de recuperação e classificação de imagens, principalmente nos métodos de aprendizado supervisionado e não supervisionado, com foco em técnicas de destaque recente na literatura. Na Seção 3.1 apresentamos os métodos de extração de características, na Seção 3.2 apresentamos os métodos aplicados a *Convolutional Neural Networks* (CNN), *Transformer* e kNN, na Seção 3.3 apresentamos as técnicas para combinação de imagens, na Seção 3.4 apresentamos a conceituação teórica sobre aprendizado de variedades, e por fim, na Seção 3.5 apresentamos os trabalhos correlatos.

## 3.1 Métodos de Extração de Características

As características extraídas de imagens são utilizadas em algoritmos de classificação para apoiar os médicos no processo de entendimento e diagnóstico de anormalidades nos órgãos e tecidos humanos, causadas por determinadas doenças. Algumas características das imagens como diferenças de textura e formato, alterações de cores e variações de luminosidade podem fornecer informações relevantes para esses algoritmos.

Cada tipo de imagem e doença que está sendo analisada, requer a seleção de um conjunto específico de extratores de características para resultar em uma classificação com maior acurácia e precisão.

Apesar de aplicarmos neste trabalho a extração de características utilizando aprendizado profundo, procuramos descrever os métodos tradicionais para a extração de características de imagens. Como o *Gray Level Co-occurrence Matrix* (GLCM) é um método bastante utilizado em imagens médicas, ele foi escolhido como *baseline* de comparação com os métodos de aprendizado profundo aplicados nos experimentos.

- **GIST**

O descritor GIST (OLIVA; TORRALBA, 2001) fornece uma representação de baixa dimensão de uma cena que não requer qualquer forma de segmentação. O descritor foca na forma da cena, na relação entre os contornos das superfícies e suas propriedades, e ignora os objetos locais na cena e as relações entre eles. É derivado do redimensionamento de uma imagem para 128x128 *pixels* e iteração em diferentes escalas, onde para cada escala a imagem é dividida em células de 8x8 *pixels*. Para cada célula, são extraídos histogramas de orientação (a cada 45 graus), cor e intensidade. O descritor é uma concatenação de

todos os histogramas, para todas as escalas e células.

- ***Gray Level Co-occurrence Matrix* (GLCM)**

*Gray Level Co-occurrence Matrix* (GLCM) (HARALICK; SHANMUGAM; DINSTEIN, 1973) é um dos métodos mais antigos para extração de características de textura. Tem sido amplamente utilizado em muitas aplicações de análise de textura e permaneceu como um método de extração de característica importante no domínio da análise de textura. O GLCM determina a relação de textura entre os *pixels*, computando para cada posição da janela de filtro, a frequência com que pares específicos de valores de células de imagem ocorrem em posições de células vizinhas. Os resultados são tabulados em uma matriz de coocorrência com o mesmo número de linhas e colunas que os valores de cinza na imagem, e medidas estatísticas específicas são calculadas a partir dessa matriz para produzir o valor filtrado para a célula-alvo. As características extraídas pelo GLCM incluem autocorrelação, contraste, correlação, dissimilaridade, energia, entropia, homogeneidade, probabilidade máxima, soma de quadrados entre outros.

- ***Histogram of Oriented Gradients* (HOG)**

*Histogram of Oriented Gradients* (HOG) são descritores de características usados para detecção de objetos (MCCONNELL, 1986). A técnica funciona contando a ocorrência de orientação de gradiente calculada em uma grade densa de células uniformemente espaçadas em uma imagem. A ideia por trás desse algoritmo é que a aparência local dos objetos em uma imagem pode ser descrita usando a distribuição das direções das bordas.

- ***Pyramid Histogram of Oriented Gradients* (PHOG)**

*Pyramid Histogram of Oriented Gradients* (PHOG) (BOSCH; ZISSERMAN; MUÑOZ, 2007) é uma extensão das características HOG. No PHOG, o *layout* espacial da imagem é preservado dividindo a imagem em sub-regiões em múltiplas resoluções e aplicando o descritor HOG em cada sub-região. Para computar as características do PHOG, o detector de bordas é geralmente aplicado em imagens em tons de cinza, em seguida, uma pirâmide espacial é criada com quatro níveis. O histograma de gradientes de orientação é então calculado para todos os compartimentos em cada nível. Todos os histogramas são então concatenados para criar a representação PHOG da imagem de entrada.

- ***Local Binary Pattern* (LBP)**

*Local Binary Pattern* (LBP) é um algoritmo de extração de características que aplica um operador de textura simples mas eficiente, utilizado para descrever os padrões de texturas locais de uma imagem (OJALA; PIETIKÄINEN; HARWOOD, 1996). É

amplamente utilizado em aplicações de processamento de imagem. O LBP funciona em uma janela de tamanho de 3x3 *pixels* que percorre a imagem, sendo que o seu *pixel* central é utilizado como *threshold* para os *pixels* vizinhos. O *threshold* é aplicado de acordo com os valores dos *pixels* adjacentes ao *pixel* central, então a matriz LBP é calculada de acordo com os valores dos vizinhos locais. Devido ao seu poder discriminativo e simplicidade computacional, o operador de textura LBP se tornou uma abordagem popular em várias aplicações. A propriedade mais importante do operador LBP no uso em aplicações é a sua robustez para mudanças monotônicas de escalas de cinza causadas por variações de iluminação.

- ***Local Binary Gray Level Co-occurrence Matrix* (LBGLCM)**

O método de extração de características *Local Binary Gray Level Co-occurrence Matrix* (LBGLCM) é uma técnica híbrida usada junto com o padrão binário local LBP e GLCM. A técnica LBP é aplicada à imagem em níveis de cinza, em seguida, as características GLCM são extraídas da imagem de textura LBP. O método GLCM leva em consideração os *pixels* vizinhos no estágio de extração de características. Ele não executa nenhuma operação em outros padrões locais na imagem. A informação textural e espacial da imagem é obtida em conjunto com o método LBGLCM.

- ***Gray Level Run Length Matrix* (GLRLM)**

*Gray Level Run Length Matrix* (GLRLM) (TANG, 1998) é uma matriz da qual as características de textura podem ser extraídas para análise de textura, que utiliza métodos estatísticos de ordem superior para extrair as características espaciais de *pixels* de níveis de cinza. Uma execução de nível de cinza (*gray level run*) pode ser descrita como uma linha de *pixels* em uma determinada direção com o mesmo valor de intensidade. O número desses *pixels* define o comprimento de execução do nível de cinza (*gray level run length*) e o número de ocorrências é chamado de valor de comprimento de execução. O comprimento de execução é considerado um número de *pixels* vizinhos que possuem a mesma intensidade de cinza em uma direção específica. Os parâmetros GLRLM mais utilizados são: *Short Run Emphasis* (SRE), *Long Run Emphasis* (LRE), *Gray level non-uniformity* (GLN), *Run length non-uniformity* (RLN), *Run Percentage* (RP), *Low Gray Level Run Emphasis* (LGLRE) e *High Gray Level Run Emphasis* (HGLRE).

- ***Segmentation-based Fractal Texture Analysis* (SFTA)**

O *Segmentation-based Fractal Texture Analysis* (SFTA) é um algoritmo de extração de características que consiste em decompor a imagem de entrada em um conjunto de imagens binárias a partir das quais as dimensões fractais das bordas das regiões são calculadas para descrever os padrões de textura segmentada (COSTA; MAMANI; TRAINA,

2012). Para decompor a imagem de entrada em imagens binárias utiliza-se o algoritmo *Two-Threshold Binary Decomposition* (TTBD). As características são extraídas de cada imagem binária. Medidas fractais são aplicadas no algoritmo SFTA, para aprender a complexidade de contorno de objetos e as estruturas na imagem.

- ***Scale-Invariant Feature Transform* (SIFT)**

*Scale-Invariant Feature Transform* (SIFT) é um algoritmo de visão computacional usado para detectar e descrever as características locais na imagem. Ele procura os pontos extremos na escala espacial e extrai os invariantes de posição, escala e rotação (LOWEDAVID, 2004). A descrição e detecção de características locais da imagem podem ajudar a identificar objetos. As características SIFT são baseadas em alguns pontos de interesse de aparência local no objeto e não têm nenhuma relação com o tamanho e a rotação da imagem. A tolerância de luz, ruído e algumas mudanças no micro ângulo de visão também é bastante alta. A essência do algoritmo SIFT é encontrar pontos-chave em diferentes espaços de escala e calcular a direção dos pontos-chave. Os pontos-chave encontrados pelo SIFT são aqueles que são muito proeminentes e não mudam devido à iluminação, transformações, *data augmentation*, ruído e outros fatores, como pontos de canto, pontos de borda, pontos brilhantes em áreas escuras e pontos escuros em áreas claras.

## 3.2 Métodos de Classificação de Imagens

### 3.2.1 Redes Neurais Convolucionais

*Convolutional Neural Networks* (CNN) ou Redes Neurais Convolucionais são algoritmos de aprendizado supervisionado, e um tipo especializado de redes neurais *feed-forward* para processamento de dados em uma topologia de matriz. São construídas como uma pilha de camadas convolucionais, camadas de *pooling* e camadas totalmente conectadas (*fully-connected*) (LECUN et al., 1998). A Figura 9 apresenta a arquitetura da CNN LeNet-5 (LECUN et al., 1998) contendo a camada de entrada com a imagem 32x32 *pixels*, camadas de convolução (C1, C3, C5), camadas de *subsampling* (S2, S4), camada totalmente conectada (*fully-connected* F6) e camada de saída.

Camadas convolucionais são camadas que podem aprender características das imagens, enquanto a camada de *pooling* é uma camada de redução do tamanho da imagem (*down sampling*) que ajuda a aumentar o nível de abstração entre as camadas que podem ser aprendidas, e reduz a necessidade de poder computacional. As camadas totalmente conectadas (*fully-connected*) são responsáveis pela interpretação das características extraídas.

A aplicação de Redes Neurais Convolucionais foi apresentada em um trabalho que aplicou um algoritmo de *backpropagation* a um problema do mundo real para reconhecimento de dígitos manuscritos de código postal dos correios dos EUA. A rede de aprendizado foi alimentada diretamente com imagens, em vez de vetores de características, demonstrando assim a capacidade das redes de *backpropagation* de lidar com grandes quantidades de informações de baixo nível (LECUN et al., 1989).

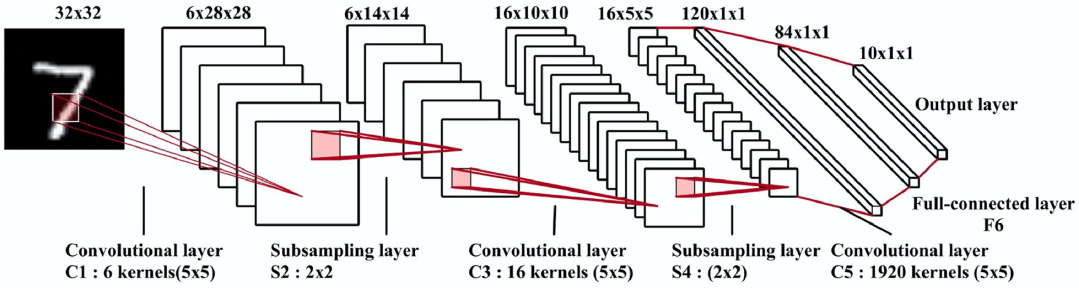



**Figura 9** – Arquitetura da CNN LetNet-5
(LECUN et al., 1998)

A arquitetura de uma CNN é análoga aquela do padrão de conectividade de neurônios no cérebro humano e foi inspirada na organização do córtex visual. Os neurônios individuais respondem a estímulos apenas em uma região restrita do campo visual conhecida como campo receptivo. Uma coleção desses campos se sobrepõe para cobrir toda a área visual (SERRE; WOLF; POGGIO, 2005).

Modelos de classificação baseados em CNN recentes estão cada vez mais mostrando melhorias significativas em uma variedade de tarefas. Exemplos incluem dados de série temporal, que podem ser representados por matriz $1D$, dados de imagens em uma matriz $2D$ de *pixels*. Tem sido utilizada em diversas aplicações práticas como visão computacional, reconhecimento e detecção de objetos e segmentação de imagens (GOODFELLOW; BENGIO; COURVILLE, 2016).

Uma CNN é composta pelas etapas de extração de características (*feature learning*) e pela etapa de classificação (*classification*). Na primeira etapa são aplicados detectores de características (*feature detectors*) nas camadas de convoluções, seguido por camadas de *pooling* (*pooling filters*) e por fim uma camada de *flattening* com vetores resultantes que serão a entrada da segunda etapa, utilizando uma rede neural densa totalmente conectada. A saída dessa rede neural é o resultado da classificação, com a identificação das classes ou rótulos detectados pela CNN. Na Figura 10 podemos observar a saída das camadas convolucionais C1 e C3, com as respectivas representações das características extraídas em cada uma delas.

A entrada da CNN é dada por uma imagem que é representada por uma matriz de *pixels*, sendo que cada posição da matriz é a intensidade de cada *pixel*. Em imagens coloridas, temos o padrão RGB (*Red, Green, Blue*) como três matrizes com a representação

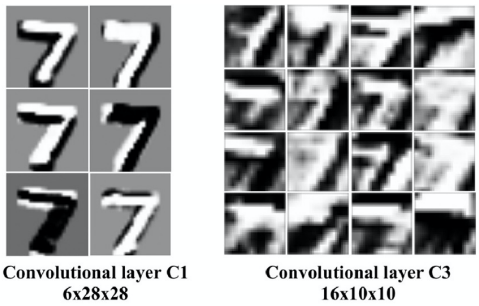



**Figura 10** – Extração de características da CNN LeNet-5 (LECUN et al., 1998)

de intensidade dos *pixels* para formar a imagem colorida. Em uma imagem preto e branco, utiliza-se o padrão *grayscale* (intensidades de cinza), com a representação de uma única matriz de intensidade de *pixels*.

- **Camadas de convolução**

A camada de convolução é um componente fundamental da arquitetura CNN que executa a extração de características, que normalmente consiste em uma combinação de operações lineares e não lineares, ou seja, operação de convolução e função de ativação (ReLU). Ela detecta características da imagem de entrada como bordas, linhas, texturas, alterações de cores, luz entre outros, utilizando filtros também conhecidos como *kernels* ou detectores de características (*feature detectores*).

Um *kernel* é uma matriz de valores com pesos, que possui dimensões $n_k \times n_k$, onde $n_k$ é um inteiro com valores entre 3 e 5, e que são treinados para detectar características específicas. O filtro se move através de cada conjunto de *pixels* da imagem para verificar se a característica a ser detectada está presente. Para fornecer o valor que representa a confiança de presença de uma determinada característica, o filtro realiza a operação de convolução, que é a soma da multiplicação de todos os *pixels* da imagem com o filtro, resultando no mapa de características (*feature map*) (MICHELUCCI, 2019). A Figura 11 apresenta um exemplo de operação de convolução, dado uma matriz de entrada, realiza-se a soma da multiplicação dos pedaços da imagem analisado (*image patch*) com o *kernel*, resultando em uma nova matriz de saída.

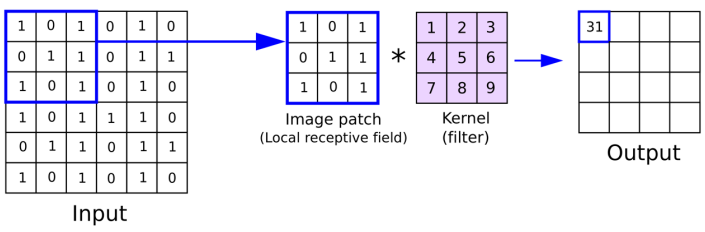



**Figura 11** – Exemplo de operação de convolução

- **Camada de *pooling***

A camada de *pooling* é responsável pela redução do tamanho espacial do mapa de caraterísticas (*down sampling*). Isso é feito para reduzir o poder computacional necessário para processar os dados através da redução de dimensionalidade. Também é muito útil para extrair características dominantes que não mudam com a rotação ou posição da imagem, garantindo a efetividade do processo de treinamento do modelo.

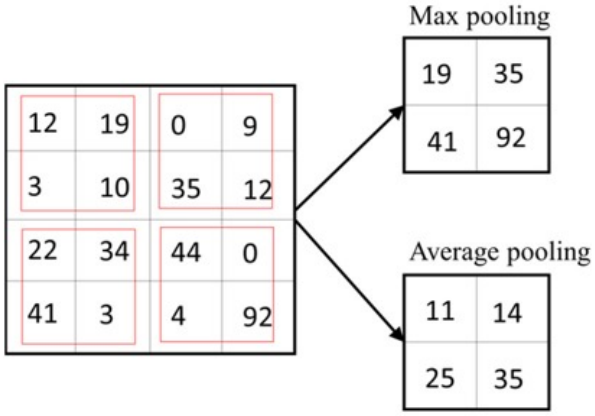



**Figura 12** – Exemplo de operação de *pooling*

Podem ser utilizados dois tipos de operações de *pooling*: *max pooling* e *average pooling*. O *max pooling* retorna o valor máximo da matriz de *pixels* da imagem avaliada pelo filtro em questão, e o *average pooling* retorna a média dos valores da matriz de *pixels* da imagem. O *max pooling* atua como supressor de ruído, pois elimina os ruídos junto com a redução de dimensionalidade (MARCHI, 2019). Podemos ver um exemplo dos dois tipos de operação de *pooling* na Figura 12.

- **Camada totalmente conectada**

A camada totalmente conectada (*fully-connected*) é uma *Multi-Layer Perceptron* (MLP) composta por três tipos de camada: camada de entrada (*input layer*), camadas escondidas (*hidden layers*) e camada de saída (*output layer*) (BISHOP, 2006), conforme ilustrado na Figura 13.

Adicionar uma camada totalmente conectada (*fully-connected*) é uma maneira de aprender combinações não lineares das características de alto nível com custo baixo. A camada totalmente conectada está aprendendo uma função possivelmente não linear nesse espaço. A saída da camada de *pooling* é transformada em um vetor (*flattened*) para ser alimentada em uma rede neural densa aplicada a cada iteração de treinamento.

A entrada da camada totalmente conectada recebe as características aprendidas pela CNN, e ao longo de várias iterações (épocas), o modelo é capaz de distinguir entre as características dominantes e certas características de baixo nível nas imagens e classificá-las usando a técnica de classificação função *softmax*.

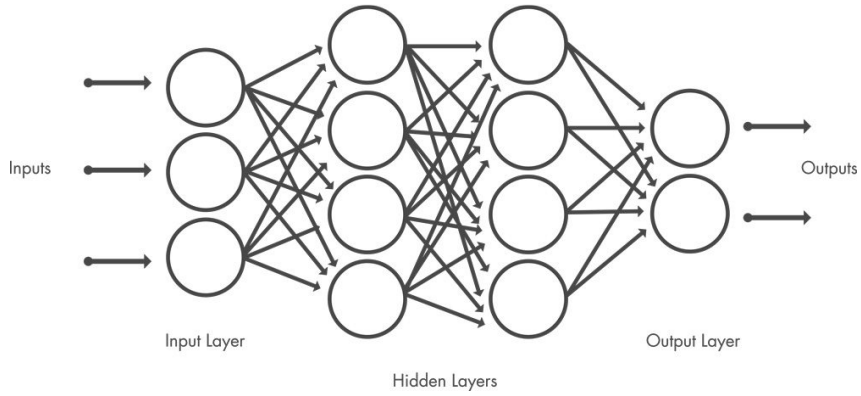



**Figura 13** – Estrutura de uma rede MLP

Na Seção 4.2.1 explicamos em detalhes a arquitetura de CNN *EfficientNet*, que foi utilizada neste trabalho.

### 3.2.2 *Transformer*

O *Transformer* (TR) é um modelo de *Deep Learning* (DL) que foi proposto como uma arquitetura de rede simples, baseada em mecanismos de atenção e pesando diferencialmente a importância de cada parte dos dados de entrada, dispensando inteiramente as recorrências e convoluções (VASWANI et al., 2017). Estudos recentes tem demonstrado excelentes resultados de eficácia, mesmo quando comparados a outros modelos relevantes, como CNNs, enquanto pode ser paralelizado, requerendo significantemente menos tempo para o treinamento. Essa arquitetura pode ser utilizada em diversas tarefas como processamento de linguagem natural (*Natural Language Processing* NLP) e visão computacional (*Computer Vision* CV).

A arquitetura de *transformer* foi projetada para lidar com dados de entrada sequenciais, como linguagem natural, para tarefas como tradução e resumo de texto. Ao contrário de outras redes neurais, os *transformer* não processam necessariamente os dados em ordem. Em vez disso, o mecanismo de atenção fornece o contexto para qualquer posição na sequência de entrada. Por exemplo, se os dados de entrada são uma frase em linguagem natural, o *transformer* não precisa processar o início da frase antes do final. Em vez disso, ele identifica o contexto que confere significado a cada palavra da frase. Esse recurso permite maior paralelização do que outras arquiteturas e, portanto, reduz o tempo de treinamento (ROTHMAN, 2021).

O *transformer* adota uma arquitetura *encoder-decoder*. O *encoder* consiste em camadas de codificação que processam uma sequência de entrada iterativamente e geram representações neurais ou vetores, que são utilizadas como entradas pelo *decoder*, para gerar uma sequência de saída, conforme apresentado na Figura 14. Tanto o *encoder* quanto o *decoder* possuem várias camadas (VASWANI et al., 2017). O número de camadas é um hiper-parâmetro desse modelo e pode ser ajustado para cada propósito.

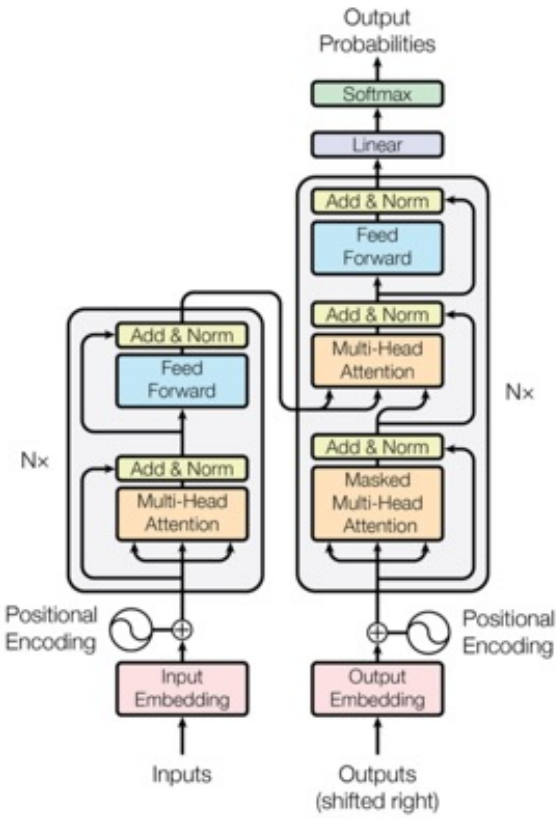



**Figura 14** – Arquitetura do *Transformer*. Camadas do *encoder* (esquerda) e *decoder* (direita) (VASWANI et al., 2017)

A função de cada camada do *encoder* é gerar codificações que contêm informações sobre quais partes das entradas são relevantes entre si. Ele passa suas codificações para a próxima camada do *encoder*. Cada camada *decoder* faz o oposto, pegando todas as codificações e usando suas informações contextuais incorporadas para gerar uma sequência de saída. Para conseguir isso, cada camada de *encoder* e *decoder* faz uso de um mecanismo de atenção.

Para cada entrada, a atenção pesa a relevância de todas as outras entradas e extrai os valores delas para produzir a saída. Cada camada do *decoder* tem um mecanismo de atenção adicional que extrai informações das saídas dos decoders anteriores, antes que a camada do *decoder* extraia informações das codificações.

Ambas as camadas do *encoder* e *decoder* são compostos de módulos que podem ser empilhados um sobre o outro várias vezes, o que é descrito por $Nx$. Possuem uma rede neural *feed-forward* para processamento adicional das saídas e contêm conexões residuais e etapas de normalização de camada.

Os blocos de construção *multi-head attention* do *transformer* são unidades de *scaled dot-product attention*. Quando uma sequência de *tokens* é passada para um modelo de *transformer*, os pesos de atenção são calculados entre cada *token* simultaneamente. A unidade de atenção produz uma saída para cada *token* que contém informações sobre o

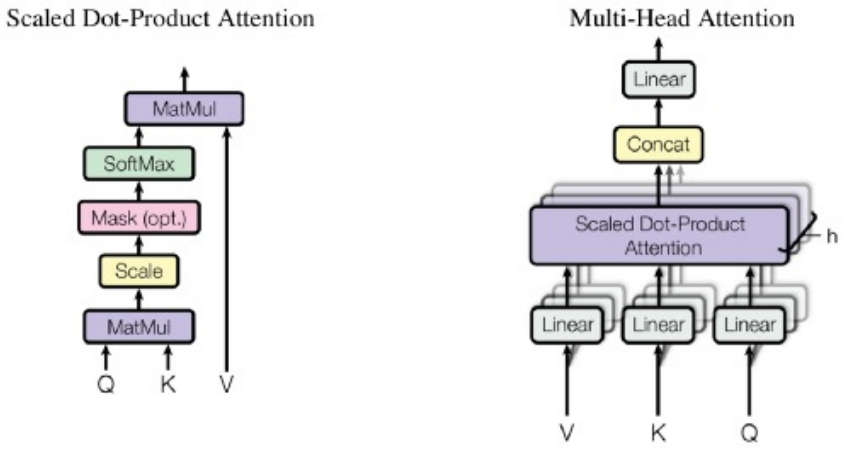



**Figura 15** – *Multi-Head Attention* (direita) consiste em várias camadas de atenção *scaled dot-product attention* (esquerda) executadas em paralelo (VASWANI et al., 2017)

próprio *token* junto com uma combinação ponderada de outros *tokens* relevantes, cada um ponderado por seu peso de atenção. A Figura 15 apresenta os blocos de *multi-head attention* e *scaled dot-product attention*.

Para cada unidade de atenção, o modelo de *transformer* aprende três matrizes de peso: $W_q$, $W_k$ e $W_v$. $W_q$ é uma matriz que contém os pesos da consulta, $W_k$ contém os pesos das chaves e $W_v$ contém os pesos dos valores. Para cada *token* $i$, a palavra de entrada $x_i$ é multiplicada por cada uma das três matrizes de peso para produzir um vetor de consulta $q_i = x_i * W_q$, um vetor-chave $k_i = x_i * W_k$ e um vetor de valor $v_i = x_i * W_v$. As matrizes $Q$, $K$ e $V$ (*query, key e value*) são definidas como matrizes onde as linhas i-ésima são os vetores resultantes $q_i$, $k_i$ e $v_i$ (VASWANI et al., 2017).

O conjunto de matrizes $W_q$, $W_k$ e $W_v$ são chamados de *attention head* e cada camada em um modelo do *transformer* possui múltiplas *attention head (multi-attention head)* executando em paralelo.

O cálculo de atenção para todos os *tokens* pode ser expresso como um grande cálculo de matriz usando a função *softmax*, que é útil para o treinamento devido às otimizações de operação de matriz computacional que calculam rapidamente as operações de matriz.

$$Attention(Q, K, V) = softmax\frac{(Q * K^T)}{\sqrt{d_k}} * V, \tag{3.1}$$

Uma das arquiteturas de *Transformer* utilizadas para imagens é o *Vision Transformer* (ViT). Na Seção 4.2.3 explicamos essa arquitetura em mais detalhes.

### 3.2.3 kNN

*K-Nearest Neighbors* (kNN) é um algoritmo de aprendizado supervisionado usado em aprendizado de máquina e pertence à família de algoritmos *Instance-based Learning* IBL (COVER; HART, 1967).

Os algoritmos desta família armazenam todas as instâncias de treinamento e, quando uma nova instância é apresentada ao algoritmo para ser classificada, um conjunto de instâncias similares (ou próximas) à nova instância é recuperada do conjunto de treinamento e utilizada para classificar a nova instância. Ele é um classificador onde o aprendizado é baseado "no quão similar" é um dado do outro. O treinamento é formado por vetores de n-dimensões.

O kNN tenta classificar cada amostra de um conjunto de dados avaliando sua distância em relação aos vizinhos mais próximos. Se os vizinhos mais próximos forem majoritariamente de uma classe, a amostra em questão será classificada nesta classe.

A Figura 16 apresenta um exemplo para ilustrar o funcionamento do algoritmo kNN. Na figura existe uma instância a ser classificada representada pela interrogação, e instâncias de treinamento já classificadas associadas às classe triângulo e quadrado.

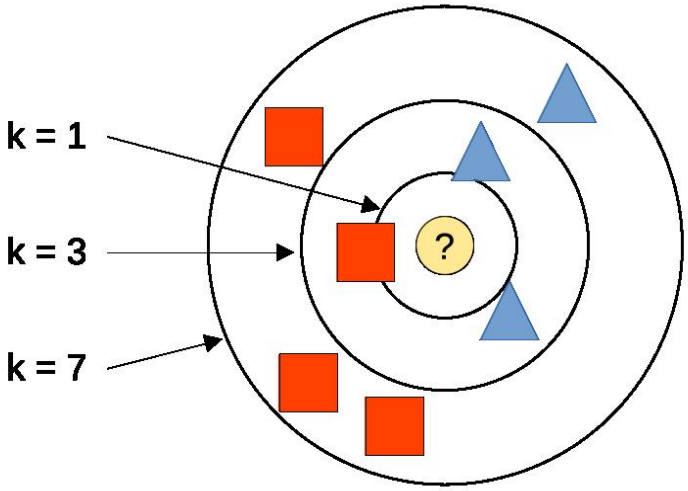



**Figura 16** – Exemplo de algoritmo kNN

Para o valor de $k = 1$, pelo funcionamento do algoritmo kNN, a nova instância será classificada como pertencente à classe quadrado, uma vez que a classe do vizinho mais próximo é a classe quadrado. Caso o valor de $k = 3$, a classe da nova instância será triângulo, dado que duas instâncias dos três vizinhos mais próximos têm classe triângulo e uma tem classe quadrado. Já no caso de $k = 7$, a classe da nova instância será a classe quadrado. Neste algoritmo, o que determina a classificação é a maior frequência das classes dos $k$ vizinhos mais próximos da instância a ser classificada

As etapas do algoritmo kNN são:

1. Recebe um dado não classificado;
2. Mede a distância (Euclidiana, Manhattan, Minkowski ou Ponderada) do novo dado com todos os outros dados que já estão classificados;
3. Obtém as k menores distâncias;
4. Verifica a classe de cada da um dos dados que tiveram a menor distância e conta a quantidade de cada classe que aparece;
5. Toma como resultado a classe que mais apareceu dentre os dados que tiveram as menores distâncias;
6. Classifica o novo dado com a classe tomada como resultado da classificação.

## 3.3 Métodos de Combinação de Características de Imagem

Várias técnicas de fusão (combinação) de descritores de imagem têm sido propostas na literatura. A fusão pode ser investigada de formas diferentes de acordo com os descritores envolvidos, a estratégia de combinação e a aplicação alvo. Em geral, as abordagens de fusão existentes podem ser categorizadas como abordagens de fusão precoce (*early fusion*) ou fusão tardia (*late fusion*), como apresentada na Figura 17, que se referem à sua posição relativa a partir da comparação das características ou etapa de aprendizado em toda a cadeia de processamento (BHOWMIK et al., 2014). Alguns autores tem utilizado a fusão híbrida (*hybrid fusion*), que é a união da fusão precoce com a tardia, para resolver vários tipos de problemas de análise multimídia.

### 3.3.1 Fusão Precoce

A fusão precoce normalmente se refere a combinação de características e uma única representação antes do processamento de similaridades entre as imagens. Este tipo de abordagem é muito comum em CBIR, e as soluções mais conhecidas são baseadas em concatenação dos vetores de características em um único vetor, como em características SIFT (LOWEDAVID, 2004), HOG (MCCONNELL, 1986) e LBP (OJALA; PIETIKÄINEN; HARWOOD, 1996). Outras abordagens propostas para CBIR são baseadas em fusão precoce de diferentes espaços de características como cor e formato (KIMURA et al., 2011), ou cor e textura (YUE et al., 2011).

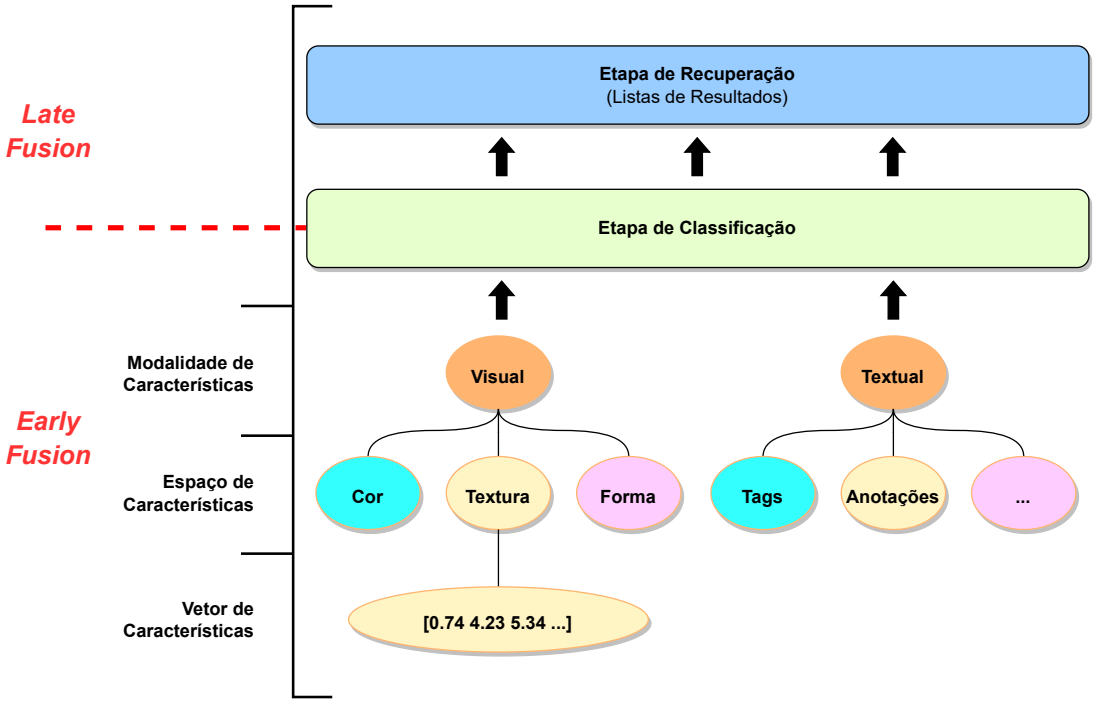



**Figura 17** – CBIR e os Métodos de Fusão
(PIRAS; GIACINTO, 2017)

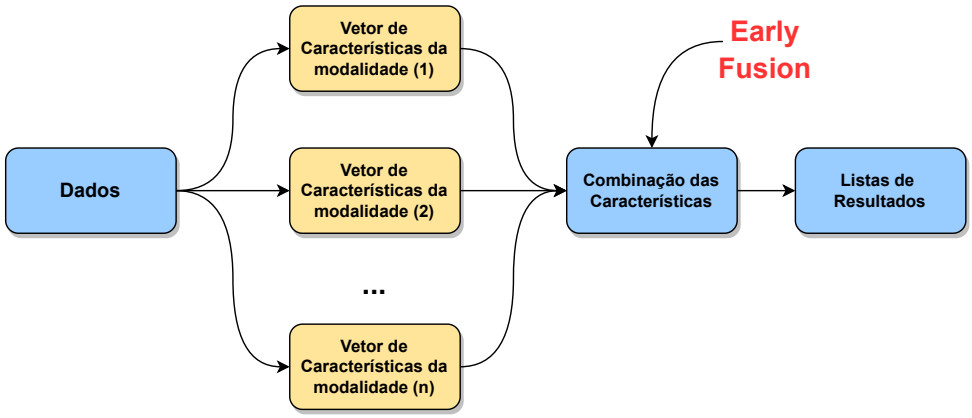



**Figura 18** – Esquema de Fusão Precoce

### 3.3.2 Fusão Tardia

A fusão tardia se refere a combinação de características no último estágio das respostas obtidas após a comparação de características individuais ou aprendizado, como realizado por redes de aprendizado profundo. A combinação de diferentes características ocorre depois do cálculo da distância dos objetos da coleção. Deste modo, a combinação não é realizada sobre vetores de características diretamente, dado que os mesmos já foram processados em alguma representação de similaridade (BHOWMIK et al., 2014), mas sim a partir do vetor resultante do cálculo das distâncias entre eles.

Quando considerados CBIR, múltiplas saídas ranqueadas de múltiplos descritores são agrupados para gerar outra saída ranqueada. Este método de fusão pode ser implemen-

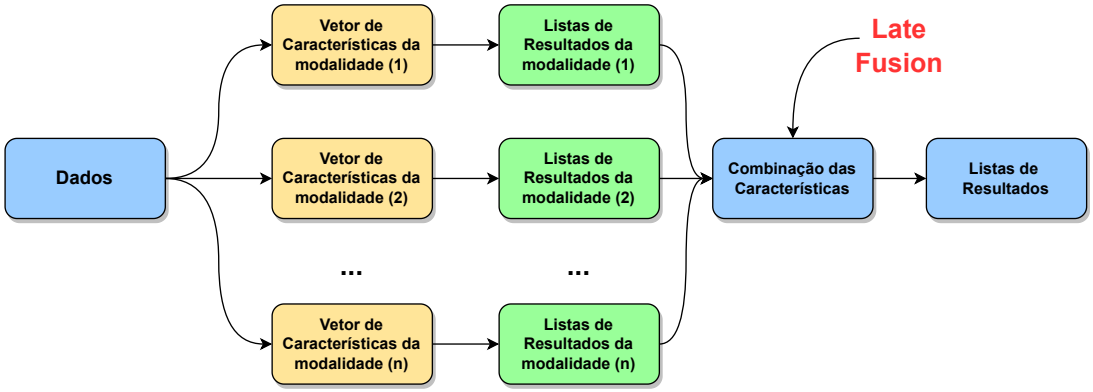



**Figura 19** – Esquema de Fusão Tardia

tado tanto baseado em pontuação (*score-based*) onde combina as similaridades ou vetores de distâncias, ou baseada em ranqueamento (*rank-based*) que considera a combinação dos ranques de resposta. Métodos de re-ranqueamento (VALEM; PEDRONETTE, 2016), como os baseados em vizinhança e aprendizado não supervisionado, comumente realizam formas de fusão tardia. As saídas a serem combinadas podem atribuir pesos distintos para dar mais importância a determinados descritores (ZHANG; QIN; WAN, 2011).

### 3.3.3 Fusão Híbrida

Para explorar as vantagens de ambas as estratégias de fusão, alguns pesquisadores tem optado por usar a fusão híbrida, que é a combinação da fusão precoce com a fusão tardia, onde vetores de características extraídos do conjunto de dados são utilizados como entrada para treinamento de classificadores, como SVM ou CNN. Os classificadores são treinados para escolher a melhor combinação entre os vetores de entrada, criando uma nova representação que combina as múltiplas modalidades.

## 3.4 Métodos de Aprendizado de Variedades

Um dos maiores desafios em algoritmos de aprendizado de máquina são os conjuntos de pontos de alta dimensionalidade. Em conjuntos de pontos com um grande número de dimensões, as combinações de conjuntos distintos de variáveis podem crescer exponencialmente tornando o aprendizado realmente impraticável devido ao custo computacional.

Isso é conhecido na área de aprendizado de máquina como a maldição da dimensionalidade (ZHENG, 2009), e continua sendo um dos maiores desafios que desencadeou o surgimento do aprendizado profundo. *Manifold learning* ou Aprendizado de Variedades se tornou uma abordagem muito popular para lidar com tarefas dessa natureza, ao aplicar técnicas de redução de dimensionalidade.

As abordagens clássicas de redução de dimensionalidade consideram vários métodos para gerar mapas lineares e não lineares. A maioria deles define um problema de otimização onde a solução é obtida por gradiente descendente, que só é garantido por produzir um ótimo local, e no qual o ótimo global é difícil de ser atingido por meios eficientes.

Os métodos de aprendizado de variedades abordam o problema de redução de dimensionalidade, descobrindo a estrutura geométrica intrínseca de baixa dimensão oculta em suas observações de alta dimensão.

Para entender o conceito de *manifold*, podemos imaginar um folha de papel, que é um objeto bidimensional que existe em um mundo tridimensional, e pode ser dobrado ou enrolado nessas duas dimensões. No conceito teórico de aprendizado de variedades podemos considerar essa folha como um *manifold* bidimensional inserido em um espaço tridimensional. Rotacionar, reorientar ou esticar a folha de papel no espaço tridimensional, não muda a geometria plana do papel. Ao dobrar, enrolar ou amassar o papel, ele ainda estará em um *manifold* bidimensional, mas a incorporação no espaço tridimensional não será mais linear. Algoritmos de aprendizado de variedades procuram aprender sobre a natureza bidimensional, mesmo quando ele é contorcido para preencher um espaço tridimensional. A Figura 20 apresenta um exemplo de *manifold* bidimensional inserido em um espaço tridimensional.

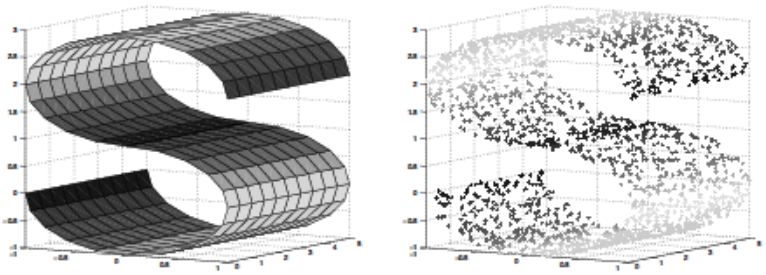


**Figura 20** – A curva S, um *manifold* bidimensional inserido em um espaço tridimensional (esquerda). 2.000 pontos de dados gerados randomicamente para representar a superfície do *manifold* da forma S (MA; FU, 2011)

Alguns métodos lineares de redução de dimensionalidade como *Principal Component Analysis* (PCA) e *Multidimensional Scaling* (MDS) são usados para reduzir a dimensão de grandes conjuntos de dados, transformando um grande conjunto de variáveis em um conjunto menor, mas que ainda contém a maioria das informações do conjunto de dados inicial (MA; FU, 2011).

Quando um *manifold* linear não resulta em uma boa representação de baixa dimensionalidade, então podemos considerar a possibilidade que os dados podem estar perto ou dentro de um *manifold* não linear. Algoritmos de aprendizado de variedades

não lineares incluem *Isometric Mapping* (ISOMAP), *Locally Linear Embedding* (LLE), *Laplacian Eigenmaps* (LE) entre outros.

Enquanto métodos lineares de aprendizado de variedades procuram preservar a estrutura global do *manifold* como é o caso do PCA, a maioria dos métodos não lineares de aprendizado de variedades buscam preservar a estrutura local dos seus vizinhos do *manifold*.

- ISOMAP (*Isometric Mapping* pode ser visto como uma extensão do MDS ou do *Kernel PCA*. Esse algoritmo procura um *embedding* de baixa dimensão que mantém distâncias geodésicas entre todos os pontos (TENENBAUM; SILVA; LANGFORD, 2000). Vemos um exemplo da aplicação do algoritmo Isomap em imagens de faces na Figura 21.

- LLE (*Locally Linear Embedding*) procura por uma projeção de baixa dimensão dos dados que preserva distâncias com a vizinhança local. Pode ser entendido como uma séria de PCA locais que são comparados globalmente para encontrar o melhor *embedding* não linear (ROWEIS; SAUL, 2000).

- LE (*Laplacian Eigenmaps*) encontra uma representação de baixa dimensão dos dados usando uma decomposição espectral do grafo Laplaciano. O grafo gerado pode ser considerado como uma aproximação discreta de um *manifold* de baixa dimensionalidade em um espaço de alta dimensionalidade (BELKIN; NIYOGI, 2001).

- t-SNE (*t-Distribution Stochastic Neighbor Embeddings*) é um dos algoritmos mais populares para visualização de alta dimensão. O algoritmo converte relacionamentos no espaço original em *t-distributions*, ou distribuições normais com tamanhos de amostra pequenos e desvios padrão relativamente desconhecidos. Isso torna o t-SNE muito sensível à estrutura local (MAATEN; HINTON, 2008).

- UMAP (*Uniform Manifold Approximation and Projection*) é um algoritmo de redução de dimensionalidade muito semelhante ao t-SNE, que constrói uma representação gráfica de alta dimensão dos dados e, em seguida, otimiza um grafo de baixa dimensão para ser o mais estruturalmente semelhante possível. Embora a matemática que UMAP usa para construir o grafo de alta dimensão seja avançada, a intuição por trás deles é notavelmente simples (MCINNES; HEALY, 2018).

*Manifolds* generalizam a noção de curvas e superfícies em duas ou três dimensões para alta dimensões. Pode ser considerado em termos similares como um espaço topológico que localmente parece um plano e sem características, e se comporta como uma espaço Euclidiano (MA; FU, 2011).

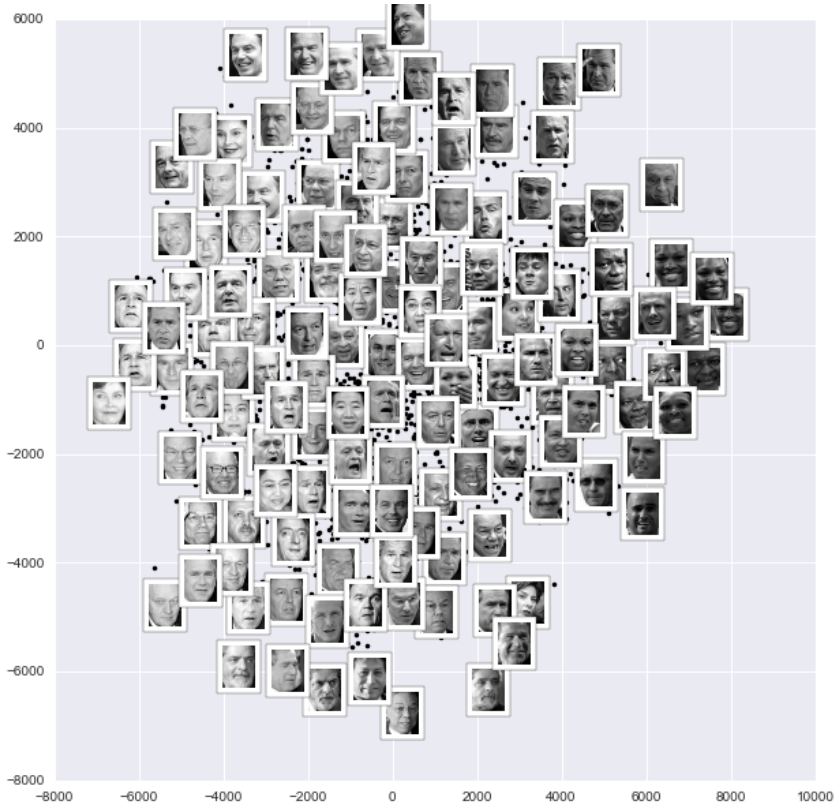



**Figura 21** – Exemplo de aplicação do algoritmo Isomap em imagens de faces (VANDERPLAS, 2016)

Algoritmos de aprendizado de variedades baseados em informações de ranqueamento também tem sido propostos para tarefas de recuperação de imagens pelo conteúdo em cenários de aprendizado não supervisionados. Diferentes modelos tem sido explorados, como grafos (PEDRONETTE; GONÇALVES; GUILHERME, 2018), hiper-grafos (PEDRONETTE et al., 2019) e árvores de busca (PEDRONETTE; VALEM; TORRES, 2021). Métodos similares também têm sido aplicados como etapa de pré-processamento visando aumentar a eficácia de tarefas de classificação supervisionada (AFONSO et al., 2018), semi-supervisionada (VALEM et al., 2018) e agrupamento (ROZIN et al., 2021).

## 3.5 Trabalhos Correlatos

Modelos de aprendizado profundo têm sido amplamente aplicados no domínio de imagens de tumores cerebrais. Existem vários trabalhos com diferentes abordagens que trouxeram excelentes resultados de classificação. Alguns artigos apresentaram aprendizado semi-supervisionado e não supervisionado incorporando recursos extraídos de diferentes tipos de redes de aprendizado profundo.

Alguns trabalhos apresentaram abordagens utilizando aprendizado supervisionado para classificação de tumores cerebrais, incluindo etapas de pré-processamento de imagens, e extração das características das imagens com técnicas como HOG e SURF (AYADI

et al., 2022), enquanto outros exploraram a extração de características utilizando CNN (AYADI et al., 2021) e PCANet (SHAHIN; ALY; ALY, 2023). Para realizar a tarefa de classificação, esses trabalhos utilizaram conjuntos de dados rotulados, atingindo resultados importantes de acurácia.

Outros trabalhos exploraram técnicas de aprendizado semi-supervisionado, com o uso de dados não rotulados, em um cenário mais próximo do encontrado no ambiente real de diagnóstico por imagens médicas. Em (GE et al., 2020), os autores realizaram a extração de características com CNN e incorporação das mesmas em grafo para aprender os rótulos das imagens, e aplicaram um classificador treinado para identificar diferentes tipos de tumores. No trabalho (KANG; ULLAH; GWAK, 2021), os autores exploraram a extração das características com diferentes arquiteturas de CNN pré-treinadas, sendo que as três principais características são concatenadas e utilizadas em um classificador para determinar o rótulo de cada imagem.

Alguns trabalhos aplicaram técnicas de fusão de características, como no trabalho (ÖKSÜZ; URHAN; GÜLLÜ, 2022), onde os autores utilizaram a fusão de características profundas extraídas de modelos de CNN, com as características superficiais extraídas de outra CNN, realizando a classificação com as características fundidas. No trabalho (AMIN et al., 2019), os autores exploraram a extração de características de texturas utilizando os métodos tradicionais como LBP, HOG, SFTA e GWF e fusão dos vetores de características para realizar a classificação utilizando *Random Forest* (RF).

A abordagem proposta neste trabalho apresenta novas contribuições para a exploração de imagens médicas de tumores cerebrais, aplicando a fusão de características extraídas de CNN, amplamente utilizado em outros trabalhos, mas adicionando as características extraídas de *Transformers* como um diferencial para o método. Também empregamos técnicas de aprendizado não supervisionado de variedades e ranqueamento de imagens em um cenário de escassez de dados rotulados, que não foram exploradas nesse contexto por outros autores.

A seguir, analisamos com mais detalhes os trabalhos correlatos.

- **Classificação de tumor cerebral baseada em abordagem híbrida**

Neste trabalho (AYADI et al., 2022), os autores propuseram uma nova técnica que visa classificar três tipos de tumores cerebrais: meningioma, glioma e tumor hipofisário. O esquema proposto faz uso da normalização, características robustas e histograma de abordagens de gradiente para melhorar a qualidade da ressonância magnética e gerar um conjunto de características discriminativas. Eles exploraram SVM na etapa de classificação.

Algumas contribuições foram apresentadas como: (i) novo esquema de classificação com o objetivo de classificar os três tipos de tumores cerebraisl; (ii) o impacto do uso de

imagens completas de MRI, alcançando resultados melhores na classificação multi-classe; (iii) o esquema sugerido classifica os tumores cerebrais em multi-classe, em comparação com os métodos tradicionais, que dependem de classificação binária custosa.

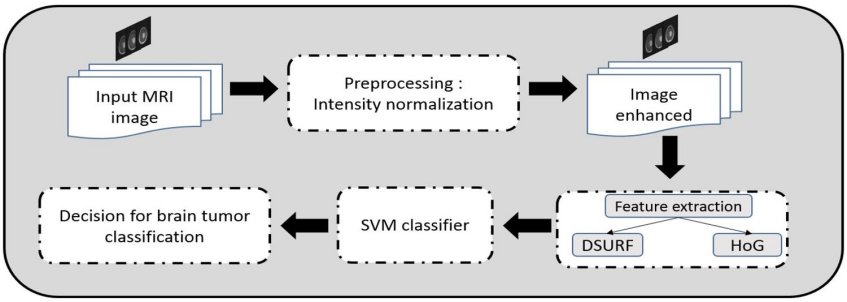



**Figura 22** – Abordagem sugerida no trabalho (AYADI et al., 2022)

O esquema proposto tem o objetivo de gerar um sistema de classificação com maior precisão. A abordagem sugerida é ilustrada na Figura 22. Essa técnica compreende três etapas: (1) normalização para melhorar a qualidade da imagem, (2) aplicação do DSURF e HoG para extrair as características mais importantes da image, e (3) SVM foi explorado como classificador na última etapa.

O estudo experimental da técnica proposta foi realizado com base no conjunto de dados público Cheng (CHENG, 2017). Ele contém 3.064 imagens correspondentes a meningioma, glioma e tumores hipofisários. Os resultados da classificação foram analisados e várias métricas de avaliação calculadas como acurácia, sensibilidade, especificidade, precisão e F1-*score*. A acurácia alcançada com base neste esquema foi de 90,27%, superando o sistema mais recente de acordo com os resultados experimentais.

- **CNN profunda para classificação de tumor cerebral**

Em (AYADI et al., 2021), os autores propuseram um novo modelo para classificação de modelos cerebrais baseado em CNN. O modelo contém várias camadas como convolução, *Rectified Linear Unit* (ReLu) e *pooling*. Esta abordagem não envolve nenhuma segmentação na etapa de pré-processamento, ao contrário de alguns outros métodos, que requerem a segmentação prévia dos tumores.

Um novo modelo para classificação de tumores cerebrais baseado na CNN é discutido neste artigo. Ele contém várias camadas como convolução, unidade linear retificada (ReLu) e *pooling*. Essa nova abordagem não envolve qualquer segmentação na etapa de pré-processamento, ao contrário de alguns métodos anteriores, que requerem segmentação prévia de tumores.

A arquitetura geral do modelo sequencial proposto está descrita na Figura 23. O modelo é composto por várias camadas, cada uma com sua própria funcionalidade. A imagem com tamanho 256 $\times$ 256 representa o modelo de entrada. Dez camadas convolucionais são exploradas para extrair o recurso importante. A camada de *pooling* máximo

após cada duas camadas convolucionais é empregada para reduzir o tamanho dos dados. Cada camada convolucional usa filtros $3 \times 3$ enquanto $2 \times 2$ são aplicados em camadas de *pooling*. Uma camada não linear é adicionada para melhorar a capacidade de ajuste da CNN. Além disso, uma normalização em lote é usada após cada camada de convolução para obter os melhores resultados otimizados e acelerar a convergência da rede. São empregadas camadas totalmente conectadas com 64 neurônios e uma camada de saída baseada em operação de *softmax*.

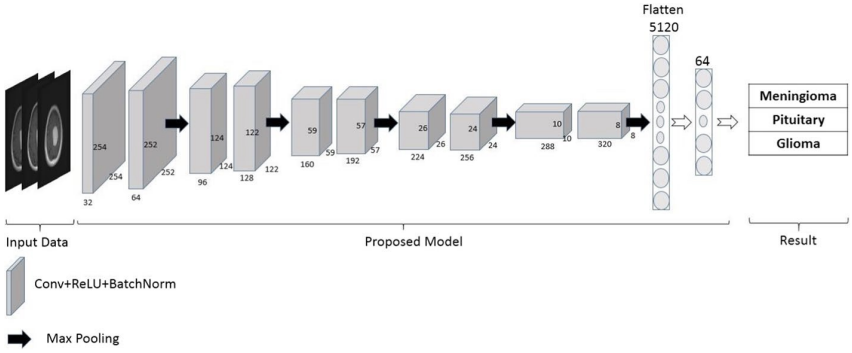



**Figura 23** – Arquitetura CNN sugerida no trabalho (AYADI et al., 2021)

O modelo proposto forneceu uma precisão geral de 94,74%, 93,71% e 95,72%, respectivamente nos conjuntos de dados (CHENG, 2017), Radiopaedia e REMBRANDT.

- **Classificação de tumores cerebrais multi-classe baseado em recursos não supervisionados do PCANet**

Um novo método de classificação híbrido de tumor cerebral multi-classe foi proposto em (SHAHIN; ALY; ALY, 2023). O método híbrido proposto consiste em dois módulos: um módulo de extração de características usando PCANet convolucional simples não supervisionado é seguido por um módulo CNN supervisionado para classificação de características, conforme apresentado na Figura 24. O modelo PCANet foi modificado aplicando *pooling* médio e funções de ativação não lineares após o primeiro e segundo estágios convolucionais, respectivamente. Os filtros convolucionais PCA aprendidos são calculados nos mapas de recursos 3D, ao contrário do modelo PCANet tradicional, que aplica filtros PCA nos mapas 2D. Essas modificações aumentam o poder expressivo das características em comparação com o PCANet tradicional ou as camadas convolucionais que podem ser aprendidas na CNN e requerem menos amostras de treinamento rotuladas. Um módulo simples de classificação CNN é utilizado para aprender recursos de alto nível apropriados para classificação de tumores.

Cada módulo da rede proposta é treinado separadamente utilizando uma estratégia de treinamento diferente. Foram aplicados otimização de pesquisa em grade para obter os valores ideais de hiper-parâmetros para o módulo de classificação que melhoram a precisão da classificação para cada conjunto de dados de tumor cerebral.

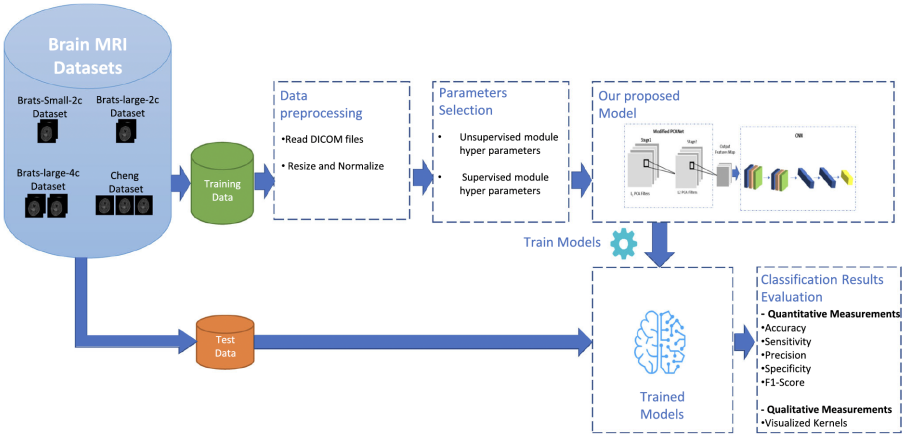



**Figura 24** – Arquitetura híbrida para classificação de tumor cerebral(SHAHIN; ALY; ALY, 2023)

O novo método foi aplicado em vários conjuntos de dados como BRATS e Cheng, e alcançou a maior acurácia de classificação 96,73% para o conjunto de dados multiclasse Cheng (CHENG, 2017).

- **Aprendizado semi-supervisionada para classificação de tumores cerebrais**

Em outro trabalho (GE et al., 2020), os autores propuseram explorar o aprendizado semi-supervisionado profundo para fazer pleno uso dos dados não rotulados. As características profundas da CNN foram incorporadas a uma nova estrutura de aprendizado semi-supervisionado baseada em grafo para aprender os rótulos dos dados não rotulados. Um classificador foi então treinado para classificar diferentes tipos de glioma usando dados rotulados e não rotulados com rótulos estimados.

A ideia principal em que se baseia a estratégia proposta consiste em melhorar o desempenho da classificação do glioma usando os dados não rotulados no conjunto de dados de treinamento, cujos rótulos são estimados por um novo método de aprendizado semi-supervisionado profundo baseado em grafos.

As novidades incluem: (a) O conjunto de dados de treinamento emprega tanto o conjunto de dados rotulado quanto o conjunto de dados não rotulado com rótulos estimados obtidos a partir do método semi-supervisionado proposto. Ao adicionar dados não rotulados e seus rótulos estimados correspondentes ao treinamento CNN, espera-se um melhor desempenho, pois mais dados de treinamento podem mitigar o *overfitting* do aprendizado profundo. Oferece mais robustez e melhor generalização ao classificador CNN. (b) Os rótulos dos dados não rotulados são estimados por um método de aprendizagem semi-supervisionado baseado em grafos. A restrição consistente 3D-2D é introduzida para melhorar a estrutura de propagação de rótulos baseada em grafos convolucionais, com base na intuição de que ressonâncias magnéticas 2D da mesma varredura 3D devem ter o mesmo

rótulo de glioma. Tal restrição é adicionada tanto à forma de construção do grafo quanto à função de custo de propagação de rótulos para aprendizagem semi-supervisionada.

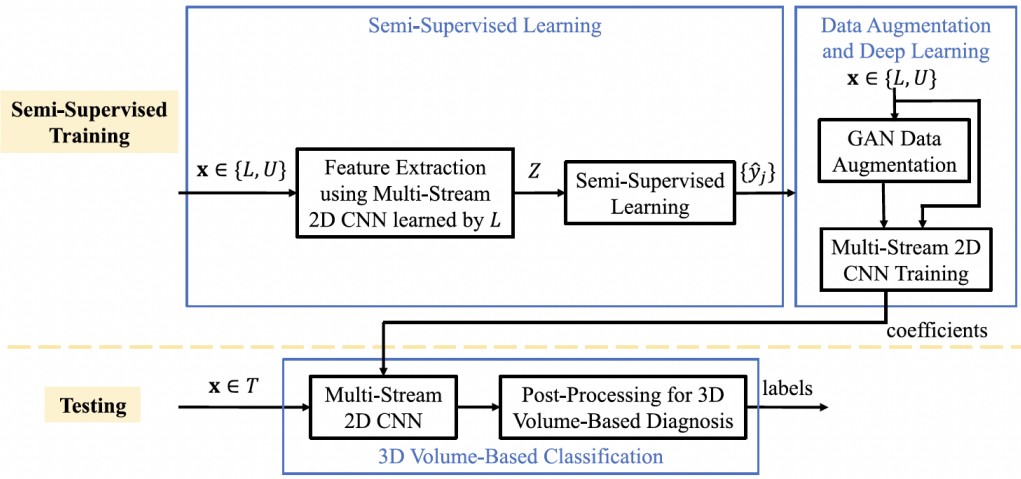



**Figura 25** – Esquema proposto para Aprendizado Semi-Supervisionado para classificação de Glioma (GE et al., 2020)

O pipeline do esquema proposto é mostrado na Figura 25. Ele consiste em três módulos, aprendizado semi-supervisionado, aumento de dados e aprendizado profundo e classificação baseada em volume 3D. CNN 2D *multistream* é primeiro treinado usando apenas os dados rotulados no conjunto de dados de treinamento. Em seguida, é usado para extrair recursos dos dados rotulados e não rotulados no conjunto de dados de treinamento. A aprendizagem semi-supervisionada baseada em grafo é usada para aprender os rótulos estimados dos dados não rotulados. Os dados de treinamento de conjuntos rotulados e não rotulados são alimentados em GANs para gerar ressonâncias magnéticas sintéticas para aumento de dados. O conjunto de dados de treinamento rotulado, o conjunto de dados de treinamento não rotulado com rótulos estimados, bem como os dados aumentados por GAN são usados como entrada para CNN 2D *multi-stream* para aprender as características dos gliomas. Depois disso, na fase de teste, as fatias de ressonância magnética do conjunto de dados de teste são testadas usando a CNN treinada, seguida de pós-processamento para gerar o tipo de glioma para cada varredura cerebral 3D.

Os resultados desse modelo mostraram bom desempenho, com acurácia de 86,53% no conjunto de dados TCGA (TCGA, 2023) e 90,70% no conjunto de dados MICCAI (MICCAI, 2023).

- **Classificação de tumor cerebral baseada em MRI usando conjunto de características profundas e classificadores de aprendizado de máquina**

No trabalho (KANG; ULLAH; GWAK, 2021), os autores propuseram uma solução híbrida que explora (1) várias redes neurais convolucionais profundas (CNNs) pré-treinadas como extratores para extrair características profundas e discriminativas de imagens cere-

brais de ressonância magnética (MRI), e (2) vários classificadores de ML para identificar as imagens de MRI do cérebro normais e anormais. Além disso, para investigar os benefícios da combinação de recursos de diferentes modelos CNN pré-treinados, projetamos o novo método de conjunto de características para a tarefa de classificação de tumores cerebrais baseada em ressonância magnética. Foi proposto o novo mecanismo de avaliação e seleção de características, onde as características profundas de 13 CNNs diferentes pré-treinadas são avaliados usando 9 classificadores de ML diferentes e selecionados com base em nossos critérios de seleção de recursos propostos.

Na nova estrutura proposta, as três principais características profundas selecionados de três CNNs diferentes são concatenados para formar uma característica sintética. O processo de concatenação integra as informações de diferentes CNNs para criar uma representação de características mais discriminativa do que usar a característica extraída de um único modelo de CNN. Um conjunto de característica profundos é então alimentado em vários classificadores de ML para prever o resultado final. O modelo foi avaliado em três conjuntos de dados diferentes: (1) BT-small-2c, o pequeno conjunto de dados com 2 classes (normal/tumor), (2) BT-large-2c, o grande conjunto de dados com 2 classes (normal/tumor) e (3) o grande conjunto de dados com 4 classes (normal, tumor de glioma, tumor de meningioma e tumor hipofisário) para classificação de tumor cerebral.

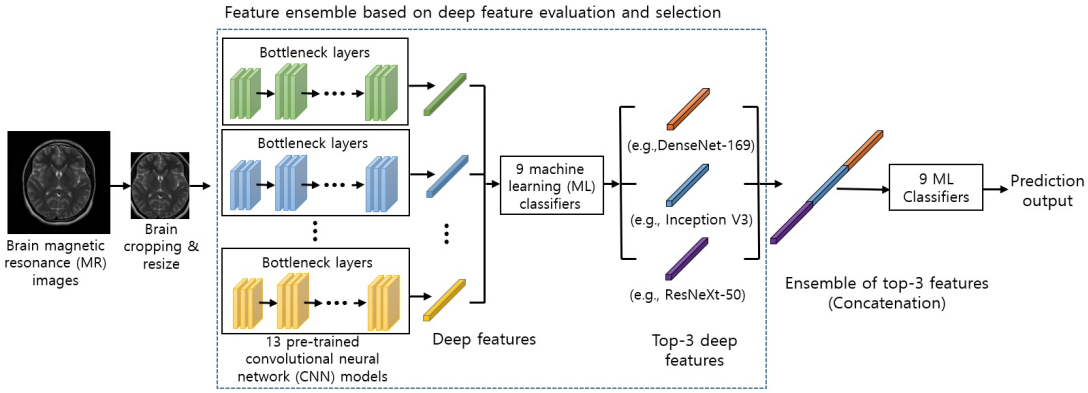



**Figura 26** – Modelo proposto usando conjunto de características baseado em avaliação e seleção profunda de características (KANG; ULLAH; GWAK, 2021)

A arquitetura do método proposto para classificação de tumores cerebrais é ilustrada na Figura 26. Primeiro, as imagens de MRI de entrada são pré-processadas antes de alimentar o modelo. Em segundo lugar, as imagens pré-processadas são usadas como entrada de modelos CNN pré-treinados como extratores de recursos. As características extraídas de modelos CNN pré-treinados são avaliados por vários classificadores ML. As três principais características profundas são selecionados com base nos resultados da avaliação dos classificadores. As três principais características profundos são concatenados no módulo de conjunto, e as características profundos concatenados são posteriormente usadas como entrada para classificadores de ML para prever o resultado final.

O método de conjunto de características proposto ajuda a superar as limitações de um único modelo CNN e produz desempenho superior e robusto, especialmente para grandes conjuntos de dados.

- **Classificação de tumores cerebrais usando características fundidas extraídas da região expandida do tumor**

Em outro estudo (ÖKSÜZ; URHAN; GÜLLÜ, 2022) foi proposto um método de classificação de tumores cerebrais usando a fusão de características profundas e superficiais para distinguir entre meningioma, glioma, tipos de tumores hipofisários. Os tumores cerebrais podem estar localizados em uma região diferente do cérebro e a textura dos tecidos circundantes também pode variar. Portanto, a inclusão de tecidos circundantes na região tumoral (expansão ROI) pode tornar as características mais distintas. Neste trabalho, redes AlexNet, ResNet-18, GoogLeNet e ShuffleNet pré-treinadas foram usadas para extrair características profundas das regiões tumorais, incluindo os tecidos circundantes.

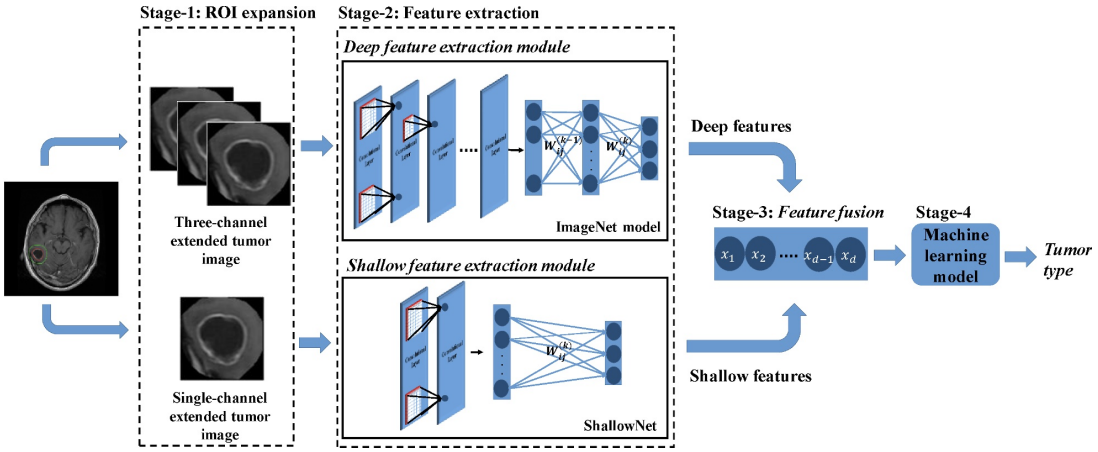



**Figura 27** – Modelo proposto para classificação (ÖKSÜZ; URHAN; GÜLLÜ, 2022)

O método proposto é apresentado na Figura 27 e consiste em uma etapa de expansão da ROI, uma etapa de extração de características com duas sub-redes rodando em paralelo, uma etapa de fusão de características e uma etapa de classificação. Na etapa de expansão da ROI, os tecidos que circundam a região do tumor (ROI expandida) são incorporados a região original do tumor (ROI). Então, o ROI expandido é organizado em três canais para redes profundas e canal único para redes rasas (ShallowNet). Na etapa de extração de características, as características profundas são extraídos da rede profunda e as características superficiais são extraídas da ShallowNet. Finalmente, as características profundas e superficiais são fundidas e a classificação é feita com as características fundidas.

Os resultados experimentais obtidos em dois conjuntos de dados disponíveis publicamente demonstram que usar a fusão de recursos e a expansão do ROI ao mesmo tempo melhora a sensibilidade média em cerca de 11,72% (expansão do ROI: 8,97%, fusão de recursos: 2,75%). Estes resultados confirmam a suposição de que os tecidos que circundam

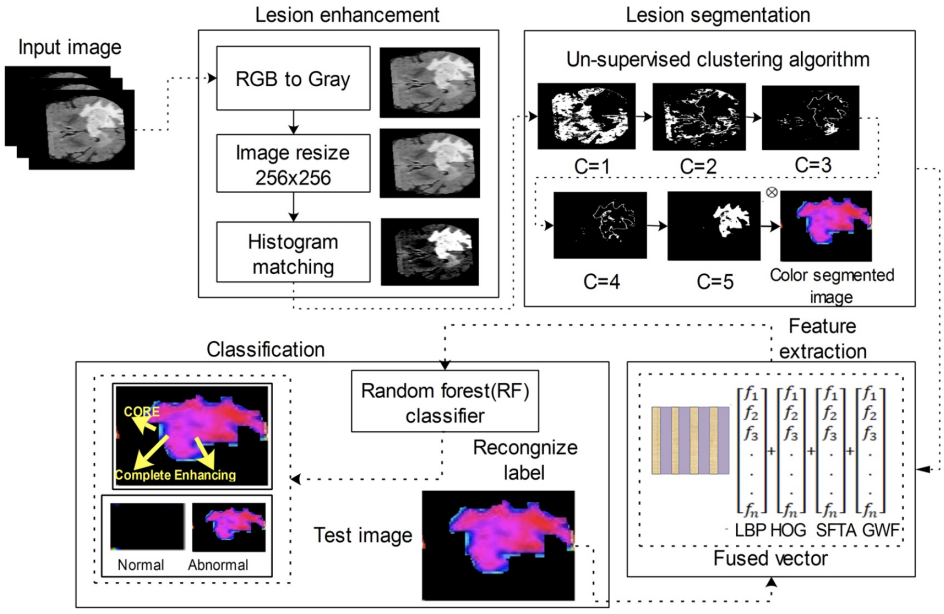



**Figura 28** – Modelo proposto para detecção de tumor cerebral (AMIN et al., 2018)

a região tumoral transportam informações distintas. Além disso, a falta de informações de baixo nível pode ser compensada graças à fusão de recursos. Além disso, resultados competitivos são alcançados em relação a estudos de última geração quando o ResNet-18 é usado como extrator profundo de características de nossa estrutura de classificação.

- **Detecção de Tumor Cerebral baseado em fusão de características e aprendizado de máquina**

Uma abordagem de aprendizado não supervisionado para segmentação de tumores cerebrais foi proposta para identificar a região sub-tumoral, como células centrais (completas e aprimoradas), saudáveis e não saudáveis (AMIN et al., 2018). Conforme apresentado na Figura 28, o método funciona em quatro etapas: (a) realce da lesão; (b) segmentação da lesão; (c) extração de características; e (d) classificação. No bloco de pré-processamento, a imagem de entrada é processada em escala de cinza e redimensionada. Além disso, o algoritmo de correspondência de histograma é usado para normalização de intensidade. No bloco de segmentação da lesão, é utilizado um método não supervisionado que apresenta a vantagem de segmentar com maior precisão a região da lesão. Da mesma forma, na fase de extração de características, quatro características de textura são extraídas em cada lesão candidata utilizando *Local Binary Pattern* (LBP), *Histograms of Oriented Gradient* (HOG), *Segmentation based Fractal Texture Analysis* SFTA e *Gabor wavelet features* (GWF). Na fase de classificação, o classificador *Random Forest* (RF) é usado para discriminação.

O método proposta foi validado em cinco conjuntos de dados públicos de imagens de

MRI, incluindo imagens do BRATS e ISLES. No experimento, a avaliação de desempenho é realizada usando métodos de validação cruzada de *five-fold* para a detecção de tumor em amostras de treinamento selecionadas aleatoriamente. A abordagem de validação cruzada é usada para evitar *overfitting*.

A avaliação foi realizada em dois experimentos, como resultados baseados em características e resultados baseados em *pixels*. Nos resultados baseados em características, a acurácia ficou entre 84,9% com a combinação de características LBP+SFTA, e 92% quando combinadas as características HOG+LBP+SFTA+GWF.

# 4 Recuperação e Classificação de Imagens com Aplicação em Suporte ao Diagnóstico de Tumores Cerebrais

Neste capítulo é apresentada a abordagem proposta para recuperação, classificação e combinação de diferentes características para representação de imagens médicas. Na Seção 4.1 apresentamos uma visão geral da metodologia proposta, na Seção 4.2 são apresentados os modelos de extração de características de imagens *EfficientNet*, *ResNet* e ViT e na Seção 4.3 são discutidas as técnicas de combinação de características através do uso do *framework* UDLF.

## 4.1 Visão Geral

As imagens médicas consolidaram-se como uma relevante ferramenta em amplo espectro de aplicações para visualização dos órgãos humanos com o objetivo de realizar diagnósticos de doenças ou anormalidades em seu funcionamento. Essas imagens podem ser capturadas por vários tipos de dispositivos e serem apresentadas em diferentes formatos, conforme apresentado na Seção 2.4.

Com o resultado desses exames, os médicos buscam entender e interpretar as imagens, com o objetivo de avaliar o funcionamento dos órgãos e detectar possíveis doenças ou anomalias. Essa avaliação é tradicionalmente realizada por médicos treinados com base em sua experiência, casos correlatos e avaliação individual ou de um grupo de médicos. O resultado da avaliação apoia os médicos no diagnóstico e tratamento da doença ou anormalidade, seja em estágios iniciais ou mais avançados. As tarefas realizadas manualmente pelos médicos podem ser sistematicamente auxiliadas por técnicas de aprendizado de máquina, e se apoiar no diagnóstico através de algoritmos de recuperação e classificação das imagens médicas.

Contudo, apesar dos aspectos promissores associados às técnicas de aprendizado de máquina como suporte ao diagnóstico, persistem desafios de grande relevância a serem abordados. A escassez de dados rotulados é um cenário comum, especialmente no domínio médico, e afeta de maneira direta as técnicas baseadas em aprendizado profundo que comumente requerem grande quantidade de dados de treinamento.

As dificuldades associadas à interpretabilidade dos resultados obtidos por métodos de aprendizado profundo é outro desafio fundamental, especialmente em tarefas de classificação em que a saída é apenas uma classe. Além disso, há uma grande diversidade de

modelos disponíveis para representação de imagens e o processo de seleção de um modelo em específico não é uma tarefa simples.

Neste cenário, este trabalho apresenta uma abordagem que pretende atacar os três desafios discutidos. A abordagem proposta utiliza métodos de aprendizado não supervisionado contextual com o objetivo de obter melhores informações de similaridade em cenários de escassez ou ausência de dados rotulados. A abordagem utiliza um arcabouço comum baseado em ranqueamento, seja para tarefas de recuperação ou classificação. Dessa forma, pretende-se melhorar os aspectos de interpretabilidade, uma vez que os resultados de classificação são obtidos a partir dos resultados de recuperação, e as imagens mais similares podem ser analisadas por especialistas. Por fim, o modelo também baseia-se na combinação de diferentes características, combinando representações geradas por diferentes modelos.

A abordagem proposta é representada na Figura 29 e descrita a seguir. Primeiro, as características da imagem são extraídas do conjunto de dados fornecido, usando diferentes modelos de aprendizado profundo no estado-da-arte treinados por *transfer learning*, considerando os modelos *Transformers* e CNN. Com base nas características extraídas, as distâncias Euclidianas são calculadas e dão origem aos resultados de ranqueamento iniciais. Tais resultados são utilizados como entrada para aplicar vários métodos de aprendizado não supervisionados e fusão de características. A saída define os dados para recuperação e classificação com base na classificação kNN para determinar a classe final da instância de teste. Cada etapa é definida a seguir:

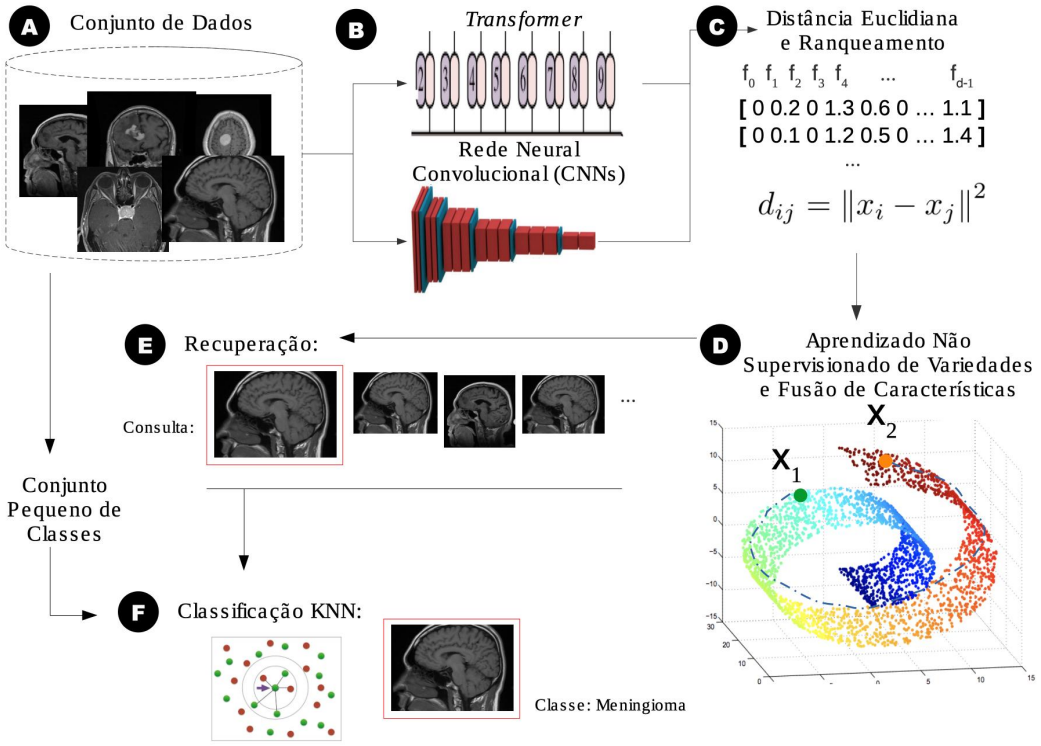



**Figura 29** – Representação das etapas da abordagem proposta

(A) O **Conjunto de Dados** de imagens foi organizado de forma a emular cenários de escassez de dados rotulados. A maioria das imagens é usada para teste e algumas imagens são consideradas para fins de treinamento. Também consideramos este protocolo para treinar os modelos de aprendizado profundo de classificação para serem usados como *baseline* para a avaliação experimental.

(B) Os modelos ***Transformers* e CNNs** treinados por *transfer learning* no ImageNet (DENG et al., 2009) foram usados para extração de características da imagem. Para obter a matriz de características para cada modelo, utilizamos a última camada antes da camada de classificação.

(C) Aplicando a **Distância Euclidiana** na matriz de extração de características da imagem, obtivemos uma nova matriz de distâncias para cada imagem do conjunto de dados. Considerando uma coleção de $n$ pontos em um espaço Euclidiano $d$ dimensional, atribuídos às colunas da matriz $X \in \mathbb{R}^{d \times n}$, $X = [x_1, x_2, \cdots, x_n]$, $x_i \in \mathbb{R}^d$. A distância entre $x_i$ e $x_j$ é dada como:

$$d_{ij} = \|x_i - x_j\|^2 \tag{4.1}$$

onde $\|\|$ denota a norma Euclidiana.

(D) As matrizes de distância extraídas de diferentes modelos de aprendizado profundo são fundidas com **Aprendizado Não Supervisionado de Variedades e Fusão de Características**. Métodos de fusão tardia sensíveis ao contexto são explorados com o objetivo de melhorar a eficácia dos resultados de classificação. Também executamos os experimentos sem fusão das características da imagem.

(E) A **Recuperação** utiliza as listas calculadas pelo aprendizado não supervisionado de variedades. Os resultados também são usados como entrada da etapa de classificação final.

(F) A **Classificação kNN** é realizada considerando as classes mais comuns na lista de imagens do ranqueamento *top-k*, resultando na classe de imagem final para cada imagem de consulta.

As etapas principais da abordagem proposta são discutidas nas seções seguintes.

## 4.2 Extração de Características

Nesta seção apresentamos as arquiteturas de aprendizado profundo, que foram utilizados para a extração de características do experimento. Aplicamos os diferentes modelos de aprendizado profundo EfficientNetB3 (TAN; LE, 2020), ResnetV2 (HE et al., 2015) e

Vision Transformers (ViT) (DOSOVITSKIY et al., 2020) para extrair as características da imagem. Obtemos o aprendizado de características da última camada para esses modelos antes da camada totalmente conectada e, em seguida, calculamos as Distâncias Euclidianas para cada uma, tendo a matriz final a ser usada no modelo de classificação. Extraímos 1.536 características para o modelo EfficientNetB3; 4.096 características para o modelo ResetNetV2; e 768 características para o modelo ViTBasePatch32.

Para a extração de características da imagem, utilizamos sete configurações diferentes, conforme a seguir.

- **EFNB3** - EfficientNetB3
- **RSNV2** - ResnetV2
- **VIT32** - ViTBasePatch32
- **EF+RS** - EfficientNetB3 + ResnetV2
- **EF+VT** - EfficientNetB3 + ViTBasePatch32
- **RS+VT** - ResnetV2 + ViTBasePatch32
- **E+R+V** - EfficientNetB3 + ResnetV2 + ViTBasePatch32

O *Gray Level Co-occurrence Matrix* (GLCM) (HARALICK; SHANMUGAM; DINSTEIN, 1973) foi incluído como *baseline* de extração de características tradicionais para os experimentos.

### 4.2.1 EfficientNet

O modelo *EfficientNet* foi proposto no trabalho (TAN; LE, 2019) e apresentado na *International Conference on Machine Learning* em 2019. Os pesquisadores estudaram o dimensionamento do modelo e identificaram que equilibrar cuidadosamente a profundidade, largura e resolução da rede pode levar a um melhor desempenho.

Com base nessa observação, eles propuseram um novo método de escalonamento composto simples, mas eficaz. Ao contrário da prática convencional que dimensiona arbitrariamente esses fatores, nosso método dimensiona uniformemente a largura, a profundidade e a resolução da rede com um conjunto de coeficientes de dimensionamento fixos. Eles usaram a pesquisa de arquitetura neural para projetar uma nova *baseline* para a rede e a ampliaram para obter uma família de modelos de aprendizado profundo, chamados *EfficientNets*, que alcançam precisão e eficiência muito melhores em comparação com as redes neurais convolucionais anteriores. A Figura 30 ilustra a diferença entre o método de escalonamento da *EfficientNet* e os métodos convencionais.

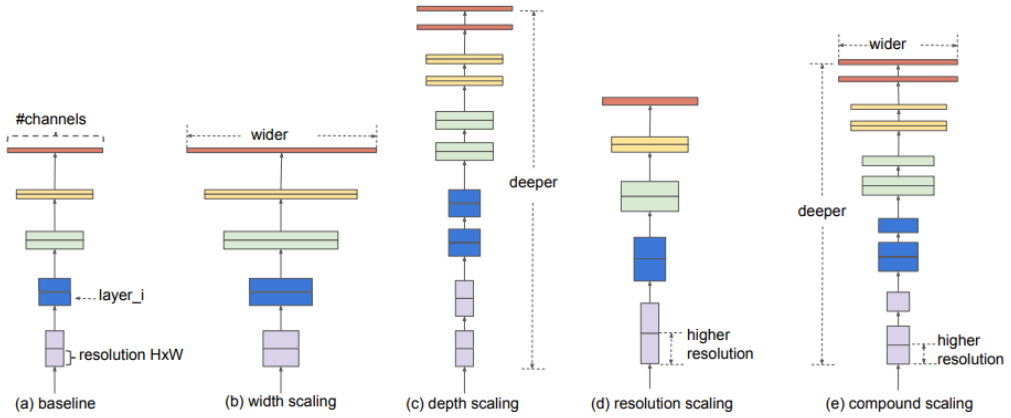



**Figura 30** – Modelos de dimensionamento. (a) é um exemplo de *baseline* rede; (b)-(d) são dimensionamentos convencionais que aumentam apenas uma dimensão da largura, profundidade ou resolução da rede. (e) método de dimensionamento composto da *EfficientNet* que dimensiona uniformemente todas as três dimensões com uma razão fixa. (TAN; LE, 2019)

### 4.2.2 ResNet

*Residual Network* (ResNet) é um modelo de aprendizado profundo usado para aplicações de visão computacional que foi proposto no trabalho (HE et al., 2015). É uma arquitetura de CNN projetada para suportar centenas ou milhares de camadas convolucionais. As arquiteturas CNN anteriores não eram capazes de escalar para um grande número de camadas, o que resultava em desempenho limitado. No entanto, ao adicionar mais camadas, os pesquisadores enfrentaram o problema do "gradiente de fuga".

As redes neurais são treinadas por meio de um processo de retro-propagação que se baseia na descida do gradiente, diminuindo a função de perda e encontrando os pesos que a minimizam. Se houver muitas camadas, as multiplicações repetidas acabarão reduzindo o gradiente até que ele "desapareça" e o desempenho sature ou piore a cada camada adicionada.

A ResNet oferece uma solução inovadora para o problema do gradiente de fuga, conhecido como "pular conexões", e empilha vários mapeamentos de identidade (camadas convolucionais que não fazem nada no início), ignora essas camadas e reutiliza as ativações da camada anterior. Ao ignorar essas camadas, acelera o treinamento inicial comprimindo a rede em menos camadas.

Então, quando a rede é retreinada, todas as camadas são expandidas e as partes restantes da rede, conhecidas como partes residuais, podem explorar mais o espaço de características da imagem de entrada.

A maioria dos modelos ResNet pula duas ou três camadas por vez com não linearidade e normalização em lote entre elas. As arquiteturas ResNet mais avançadas, conhecidas como *HighwayNets*, podem aprender a pular pesos, que determinam dinamicamente o

número de camadas a serem ignoradas.

Os blocos residuais são uma parte importante da arquitetura ResNet. Em arquiteturas mais antigas, as camadas convolucionais são empilhadas com normalização em lote e camadas de ativação não linear, como ReLu, entre elas.

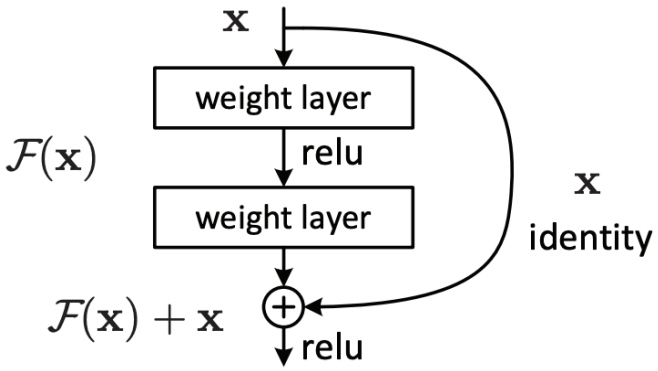



**Figura 31** – Bloco Residual

A arquitetura ResNet apresenta o conceito simples de adicionar uma entrada intermediária à saída de uma série de blocos de convolução. A Figura 31 mostra um bloco residual típico, que pode ser expresso usando a expressão $F(x) + x$, onde $x$ é uma entrada para o bloco residual e saída da camada anterior, e $F(x)$ é parte de uma CNN que consiste em vários blocos convolucionais.

Essa técnica suaviza o fluxo de gradiente durante a retro-propagação, permitindo que a rede seja dimensionada para 50, 100 ou até 150 camadas. Ignorar uma conexão não adiciona carga computacional adicional à rede.

### 4.2.3 ViT

*Vision Transformer* (ViT) é um modelo de visão baseado na arquitetura de *Transformer*, que foi originalmente proposta para tarefas baseadas em texto. ViT representa uma imagem de entrada como uma sequência de pedaços de imagens de tamanho fixo (*tokens*) e realiza operações para prever os rótulos de classe dessa imagem (DOSOVITSKIY et al., 2020), processo muito parecido com a sequência de *tokens* de palavras utilizadas quando aplicando *Transformer* em um texto.

ViT divide a imagem em pequenos pedaços (*tokens*) quadrados da imagem de entrada. Cada pedaço da imagem é transformado em um vetor, concatenando os canais de todos os *pixels* em um pedaço (*patch*) e depois projetando linearmente para a dimensão de entrada desejada (DOSOVITSKIY et al., 2020).

Como os *Transformer* são agnósticos à estrutura dos elementos de entrada, são adicionados *tokens* de posição em cada pedaço da imagem, o que permite que o modelo

aprenda sobre a estrutura da imagem. Inicialmente, o ViT não conhece sobre a localização relativa dos *patches* da imagem. Ele aprende as informações relevantes dos dados de treinamento e codifica as informações estruturais nos *tokens* de posição (WU et al., 2020). A Figura 32 ilustra a visão geral da arquitetura de *Vision Transformer*.

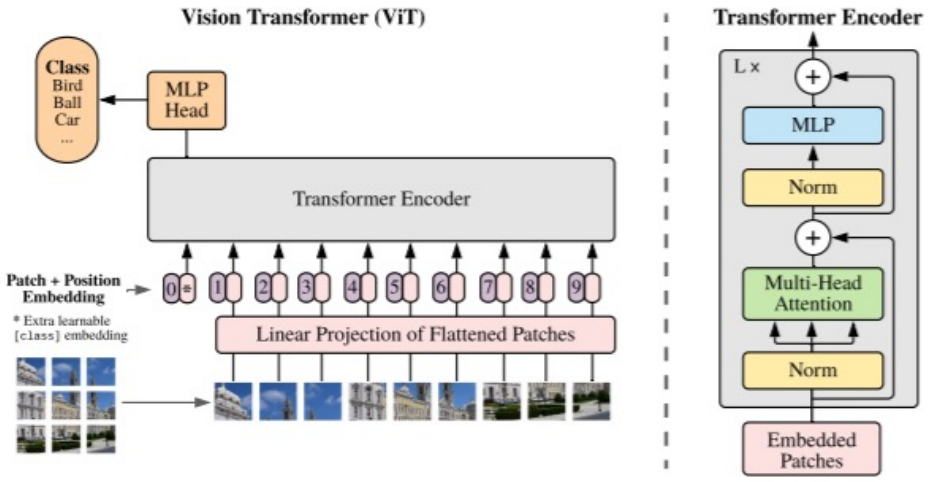



**Figura 32** – Visão geral da arquitetura de *Vision Transformer* (DOSOVITSKIY et al., 2020)

## 4.3 Aprendizado não supervisionado de Variedades e Combinação de Características

Três diferentes abordagens baseadas em ranqueamento foram consideradas para a etapa de aprendizado não supervisionado de variedades: *Cartesian Product of Ranking References* (CPRR) (VALEM; PEDRONETTE, 2016), *Log-based Hypergraph of Ranking References* (LHRR) (PEDRONETTE et al., 2019) e *Rank-Based Diffusion Process with Assured Convergence* (RDPAC) (PEDRONETTE; VALEM; LATECKI, 2021).

Também combinamos as características extraídas usando abordagens de fusão tardia disponíveis nos métodos. Para a classificação, aplicamos uma formulação kNN tradicional para selecionar a classe mais comum para cada lista de *top-k* classificada de instâncias de imagem de teste.

Para todos os experimentos, as mesmas configurações foram usadas. A maioria dos valores dos parâmetros usados pelos métodos seguiram os valores padrão disponíveis no *framework Unsupervisioned Distance Learning Framework* (UDLF) (VALEM; PEDRONETTE, 2017).

### 4.3.1 CPRR

O objetivo principal do *Cartesian Product of Ranking References* (CPRR) é maximizar a informação de similaridade codificada no ranqueamento por meio de operações de produtos cartesianos (VALEM; PEDRONETTE, 2017).

Enquanto o algoritmo CPRR considera apenas um subconjunto de listas ranqueadas para redução de custos computacionais, o produto cartesiano é utilizado para expandir as relações de similaridade. A ideia central consiste no uso de consultas kNN e kNN reversa para computar conjuntos de imagens, que são utilizadas para operações de produtos cartesianos. O CPRR pode ser usado em tarefas de agregação de *ranks* e pode ser calculado de forma eficiente por meio de computação paralela.

A abordagem proposta pode ser dividida em duas etapas principais:

- Normalização de *Rank*: as referências de *rank* recíprocas são analisadas visando melhorar a simetria da areia de vizinhança e, consequentemente, a eficácia das listas ranqueadas;
- *Cartesian Product of Ranking References*: o produto cartesiano é calculado considerando os conjuntos de vizinhança superior e vizinhança reversa. Os resultados obtidos são usados para definir uma medida de similaridade iterativa.

### 4.3.2 LHRR

O *Log-based Hypergraph of Ranking References* (LHRR) (PEDRONETTE et al., 2019) busca encontrar uma relação global de similaridade entre os elementos do conjunto explorando informações estruturais do conjunto de dados. Para isso, o método constrói um hiper-grafo baseado nas relações presentes no ranqueamento dos elementos, recebido como entrada do algoritmo.

O fluxo de processamento executado pelo LHRR, que resulta na obtenção de uma nova medida de similaridade para o conjunto de dados, pode ser dividido em cinco etapas, descritas a seguir:

1. Normalização dos Ranqueamentos: um procedimento de normalização é realizado para melhorar a simetria das referências de ranqueamento;
2. Construção do Hipergrafo: o hipergrafo modela a estrutura de similaridade global do conjunto de dados usando as informações de classificação como entrada;
3. Similaridade das Hiperarestas: as relações codificadas nas hiperarestas são usadas para computar uma nova similaridade entre objetos multimídia;

4. Produto Cartesiano das Hiperarestas: uma operação de produto cartesiano é computada para maximizar a informação de similaridade de hiperaresta;

5. Similaridade baseado em Hipergrafo: as semelhanças entre hiperarestas e as operações do produto cartesiano são combinadas para calcular uma similaridade baseada em hipergrafo, o que leva a novas classificações.

### 4.3.3 RDPAC

O *Rank Diffusion Process with Assured Convergence* (RDPAC) é um método baseado em uma formulação eficiente capaz de evitar o cálculo de valores de similaridade pequenos e irrelevantes. A ideia principal consiste em explorar as informações de ranqueamento para identificar e indexar os altos valores de similaridade nas matrizes de transição e afinidade. Desta forma, o método admite uma solução algorítmica eficiente capaz de computar uma aproximação efetiva dos processos de difusão (PEDRONETTE; VALEM; LATECKI, 2021).

A apresentação do método está organizada em quatro etapas principais: (i) é definida uma medida de similaridade com base nas informações de classificação; (ii) é realizada uma normalização para melhorar a simetria das referências de classificação; (iii) o processo de difusão de postos é realizado, exigindo um pequeno número de iterações; (iv) uma etapa de pós-difusão é realizada para explorar as informações de classificação recíproca.

### 4.3.4 UDLF

O *Unsupervised Distance Learning Framework* (UDLF) é um software que permite uma fácil utilização e avaliação de métodos de aprendizagem não supervisionada. A estrutura define um modelo amplo, permitindo a implementação de diferentes métodos não supervisionados e suportando diversos formatos de arquivo para entrada e saída (VALEM; PEDRONETTE, 2017).

O UDLF fornece um ambiente de software para implementar, usar e avaliar facilmente métodos de pós-processamento não supervisionados. O framework define um modelo geral, permitindo a implementação de diferentes métodos, baseados em medidas de distância ou informações de classificação. O usuário pode selecionar facilmente o método a ser executado e definir os respectivos parâmetros. Diferentes formatos de arquivo (para entrada e saída) são suportados, e avaliação de eficácia e eficiência também estão disponíveis. A estrutura inclui a implementação de sete métodos não supervisionados diferentes. O projeto está disponível publicamente sob os termos da licença GPLv2 de tal forma que a comunidade científica pode acessar, utilizar e compartilhar mudanças.

O uso do *framework* é baseado no arquivo de configuração, conforme ilustrado na Figura 33, que especifica todas as informações sobre a execução: a tarefa desejada, método

```
0  #The comments follow the structure:
1  #PARAMETER = VALUE #(regular expression): Explanation about the parameter
2  #If a regular expression is not specified, any input string can be used
3  #To simplify the expressions, we adopt:
4  #TBool   = (TRUE|FALSE)
5  #TUInt   = (0-9)*
6  #TFloat = ["+"|"-"] [0-9]* ["."] [0-9]+
7
8  #CATEGORY 1. GENERAL CONFIGURATION
9  UDL_TASK = UDL #(UDL|FUSION): Selection of task to be executed
10 UDL_METHOD = CPRR #(NONE|CPRR|RLRECOM|RLSIM|CONTEXTRR|RECKNNGRAPH|RKGRAPH|
   CORGRAPH): Selection of method to be executed
11 #CATEGORY 2. INPUT FILE SETTINGS
12 SIZE_DATASET = 1400 #(TUint): Number of images in the dataset
13 INPUT_FILE_FORMAT = MATRIX #(MATRIX|RK): Format of input file
14 INPUT_MATRIX_TYPE = DIST #(DIST|SIM): Type of matrix file
15 INPUT_RK_FORMAT = NUM #(NUM|STR): Format of ranked list file
16 MATRIX_TO_RK_SORTING = HEAP #(HEAP|INSERTION): Convert matrix to rks
17 NUM_INPUT_FUSION_FILES = 2 #(TUint): Number of files for FUSION tasks
18 INPUT_FILES_FUSION_1 = input1.txt #Path of the first input file
19 INPUT_FILES_FUSION_2 = input2.txt #Path of the second input file
20 #INPUT_FILES_FUSION_* = input*.txt #Path of the *th input file
21 INPUT_FILE = input.txt #Path of the main input file (matrix/rks)
22 INPUT_FILE_LIST = list.txt #Path of the list file
23 INPUT_FILE_CLASSES = classes.txt #Path of the classes file
24 INPUT_IMAGES_PATH = images/ #Dataset images path
25 #CATEGORY 3. OUTPUT FILE SETTINGS
26 OUTPUT_FILE = TRUE #(TBool): Generate output file(s)
27 OUTPUT_FILE_FORMAT = MATRIX #(RK|MATRIX): Format of output file
28 OUTPUT_MATRIX_TYPE = DIST #(DIST|SIM): Type of matrix file to output
29 OUTPUT_RK_FORMAT = ALL #(NUM|STR|HTML|ALL): Output format for rks
30 OUTPUT_FILE_PATH = output #Path of the output file(s)
31 OUTPUT_HTML_RK_PER_FILE = 1 #(TUint): Number of rks for each html file
32 OUTPUT_HTML_RK_SIZE = 20 #(TUint): Number of images per ranked list
33 OUTPUT_HTML_RK_COLORS = TRUE #(TBool): Color borders around images
34 OUTPUT_HTML_RK_BEFORE_AFTER = TRUE #(TBool): Comparison of rks
35 #CATEGORY 4. EVALUATION SETTINGS
36 EFFICIENCY_EVAL = TRUE #(TBool): Enable efficiency evaluation
37 EFFECTIVENESS_EVAL = TRUE #(TBool): Enable effectiveness evaluation
38 EFFECTIVENESS_COMPUTE_PRECISIONS = TRUE #(TBool): Compute precisions
39 EFFECTIVENESS_COMPUTE_MAP = TRUE #(TBool): Compute MAP
40 EFFECTIVENESS_COMPUTE_RECALL = TRUE #(TBool): Compute recall
41 EFFECTIVENESS_RECALL_AT = 40 #(TUint): Position to compute recall
42 EFFECTIVENESS_PRECISIONS_TO_COMPUTE = 5, 20 #(TUint ["," TUint]*):
   Precisions to be computed (unsigned integers separated by commas)
43 #CATEGORY 5. METHOD PARAMETERS
44 PARAM_CPRR_L = 400 #(TUint): Size of ranked lists to consider
45 PARAM_CPRR_K = 20  #(TUint): Number of nearest neighbors
46 PARAM_CPRR_T = 2   #(TUint): Number of iterations
```

**Figura 33** – Arquivo de configuração do UDLF

utilizado, informações do conjunto de dados, arquivos de entrada e saída, configurações de avaliação e outros detalhes. O software considera apenas um único arquivo de configuração por execução, permitindo que o usuário tenha arquivos de configuração distintos para diferentes execuções.

# 5 Avaliação Experimental

Neste capítulo discutimos a avaliação experimental realizada neste trabalho. Na Seção 5.1 apresentamos os conjuntos de dados utilizados na avaliação experimental, na Seção 5.2 detalhamos o protocolo experimental, e por fim na Seção 5.3 avaliamos os resultados obtidos no experimento.

## 5.1 Coleção de Imagens

Esta seção apresenta os conjuntos de dados utilizadas durante a avaliação experimental. Os conjuntos de dados selecionados são distintos entre si, de modo a avaliar os métodos nos mais variados cenários. A abordagem proposta é flexível para uso em outros domínios médicos, mas foram selecionados conjuntos de dados de diagnóstico de tumor cerebral para validação da proposta, que apresenta um aspecto de multi-classe com vários tipos de tumores cerebrais.

### 5.1.1 Tumor Cerebral

O tumor cerebral é considerado uma das doenças mais agressivas entre crianças e adultos. Os tumores cerebrais representam 85% a 90% de todos os tumores primários do Sistema Nervoso Central (SNC). Todos os anos, cerca de 11.700 pessoas são diagnosticadas com um tumor cerebral (GCO, 2020). A taxa de sobrevivência de 5 anos para pessoas com câncer cerebral ou tumor do SNC é de aproximadamente 34% para homens e 36% para mulheres (CANCER.NET, 2023).

Os tumores cerebrais são classificados em: tumor benigno, tumor maligno, tumor hipofisário, etc (AANS, 2023). Tratamento adequado, planejamento e diagnósticos precisos devem ser implementados para melhorar a expectativa de vida dos pacientes. A melhor técnica para detectar tumores cerebrais é a Ressonância Magnética (MRI). Uma enorme quantidade de dados de imagem é gerada através das varreduras. Essas imagens são examinadas pelo médico. Um exame manual pode ser propenso a erros devido ao nível de complexidade envolvido nos tumores cerebrais e suas propriedades.

Dois conjuntos de dados diferentes com imagens de tumores cerebrais de MRI foram utilizados nos experimentos.

- **Bhuvaji Dataset**: este conjunto de dados consiste em imagens de MRI de glioma, meningioma, tumor hipofisário e imagens não tumorais, com um total de 3.338 imagens. O conjunto de dados foi construído pelos pesquisadores Navoneel Chakrabarty

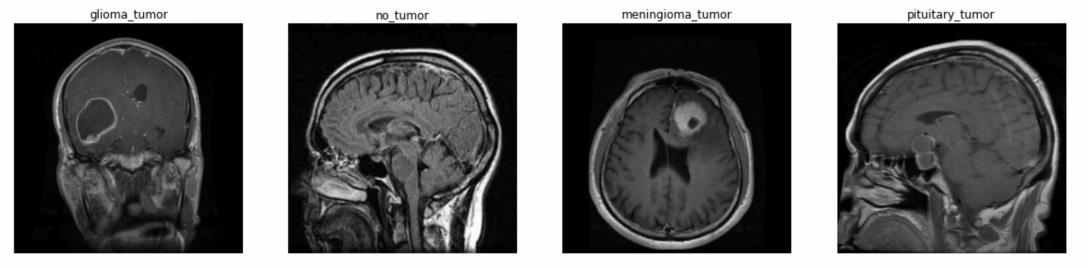



**Figura 34** – Amostras da coleção de dados. Da esquerda para a direita. Glioma Cerebral; Sem Tumor Cerebral; Meningioma Cerebral; Pituitário Cerebral.

e Swati Kanchan e organizado por Sartaj Bhuvaji. As imagens foram rotuladas e validadas pelo Dr. Shashikant Ubhe, e estão acessíveis online (BHUVAJI, 2020).

- **Cheng Dataset**: este conjunto de dados contém 3.064 imagens com contraste ponderado em T1 de 233 pacientes com três tipos de tumor cerebral: meningioma, glioma e tumor hipofisário. Este conjunto de dados foi adquirido do Hospital Nanfang, Guangzhou, China, e Hospital Geral, Tianjing Medical University, China, de 2005 a 2010 pelo Dr. Jun Cheng e está acessível online (CHENG, 2017).

A seguir explicamos em detalhes os tipos de tumor cerebral, contidos nas imagens apresentadas na coleção de dados utilizadas nesse experimento.

- **Glioma Cerebral**

O glioma é um tipo comum de tumor que se origina no cérebro. Cerca de 33% de todos os tumores cerebrais são gliomas, que originam-se nas células gliais que residem no cérebro. Os gliomas são chamados de tumores cerebrais intra-axiais, porque eles crescem dentro da substância do cérebro e frequentemente misturam-se com o tecido normal do cérebro (HOPKINS, 2020a).

Alguns gliomas não apresentam sintomas, mas, quando eles apresentam, é porque já estão pressionando o cérebro ou a medula espinhal. Os sintomas mais comuns do glioma são dores de cabeça; convulsões; mudanças de personalidade; fraqueza nos braços, na face ou pernas; dormência; tontura; problemas na fala; náusea; vômito.

Os gliomas são descobertos através de Ressonância Magnética ou outro exame com imagem do cérebro, e diagnosticado por meios cirúrgicos para determinar o tipo e o grau. O tratamento pode incluir observação, cirurgia, terapia de radiação, quimioterapia ou uma combinação dessas modalidades. Não existem causas óbvias para o glioma. Eles podem ocorrer em pessoas de todas as idades, mas, são mais comuns em adultos.

- **Meningioma Cerebral**

O meningioma é o tipo mais comum de tumor cerebral primário e é responsável por aproximadamente 30% dos tumores cerebrais. Um tumor cerebral primário é um tumor que se origina no cérebro (HOPKINS, 2020b).

Esses tumores originam-se na meninge, que é a terceira camada mais externa de tecido entre o crânio e o cérebro, bem debaixo do crânio, que cobre e protege o cérebro. Eles podem crescer lentamente ou não crescer, e existir por anos antes de serem detectados.

Às vezes os médicos podem descobrir um meningioma acidentalmente em uma Ressonância Magnética ou Tomografia Computadorizada da cabeça ou da medula espinhal. A maioria deles podem ser removidos com cirurgia. Se os sintomas do meningioma ocorrerem, eles podem ser súbitos e começar lentamente, conforme o tumor cresce e pressiona o cérebro ou a medula espinhal.

Dependendo do tamanho e da localização do meningioma, os sintomas comuns podem incluir: dores de cabeça; convulsões; turvação visual; fraqueza nos braços ou pernas; dormência; perda de equilíbrio; perda de audição; e perda de memória.

- **Pituitário Cerebral**

O tumor na hipófise é um tumor benigno resultante do crescimento de uma massa anormal na hipófise, glândula localizada na base do cérebro responsável por controlar outras glândulas e produzir hormônios. O tumor na hipófise também é conhecido como tumor pituitário ou adenoma hipofisário (D'OR, 2022).

O tumor na hipófise não tem uma causa definida, no entanto acredita-se que certas alterações no DNA das células da medula óssea podem ser responsáveis pelo aparecimento do tumor.

Tumor na hipófise tem cura, pois é um tumor benigno, ou seja, não cancerígeno. O tratamento é feito com a remoção cirúrgica do tumor, pelo nariz ou por uma incisão no crânio. Caso o tumor seja muito grande e afete outras áreas do cérebro, a cirurgia é mais arriscada. O tratamento com radioterapia ou medicamentos para reduzir o tumor podem ser indicados antes da cirurgia, mas tudo depende de seu tamanho e da decisão do médico.

## 5.2 Protocolo Experimental

Conforme apresentado no Capítulo 4, utilizamos os conjuntos de dados de tumor cerebral para realizar a extração de características utilizando a arquitetura de CNN *EfficientNet*, *ResNet* e *transformer* ViT com *transfer learning* do Imagenet (DENG et al., 2009). Utilizamos a última camada totalmente conectada *fully connected layer* da *EfficientNet* com 1.536 características extraídas para cada imagem e da *ResNet* com 4.096 características, e a camada de extração do *token* da ViT com 768 características extraídas.

As seguintes definições foram usadas para os experimentos.

**Formulação de experimentos e *baseline***: os conjuntos de dados foram divididos para emular cenários com escassez de rótulos. Portanto, 80% do total de imagens foram usadas para teste e apenas 20% para treinamento. Comparamos os métodos propostos com modelos de classificação supervisionados com conjuntos de dados de tumores cerebrais de ressonância magnética nos mesmos cenários. Os mesmos modelos usados para extração de características (EfficientNet, ResNet e ViT) foram considerados como *baselines*.

**Configurações dos parâmetros de Aprendizado de Variedades**: foram utilizados os métodos *Unsupervisioned Distance Learning* (UDL) com as abordagens CPRR, LHRR e RDPAC com as seguintes configurações: tamanho das listas ranqueadas $L = 400$; tamanho da vizinhança $k = 80$; o número de iterações $T = 2$.

**Classificação kNN**: utilizamos o $k = 20$ para a etapa de classificação das listas ranqueadas.

**Métricas**: utilizamos as métricas de precisão e MAP para as tarefas de recuperação, e a métrica de acurácia para as tarefas de classificação.

## 5.3 Resultados Experimentais

Nesta seção, apresentamos os resultados de recuperação e classificação para os dois conjuntos de dados considerados.

- **Conjunto de dados Bhuvaji**

A Tabela 1 apresenta os resultados de recuperação para diferentes métodos de aprendizado não supervisionado de variedades usando as características extraídas por GLCM, EfficientNetB3, ResnetV3 e ViTBasePatch32 do conjunto de dados Bhuvaji. Os melhores resultados foram encontrados na configuração EFNB3 com 0,2031 de MAP no método LHRR, configuração RSNV2 com 0,2636 de MAP no método RDPAC e configuração VITB32 com 0,2615 de MAP no método RDPAC.

A Tabela 2 mostra os resultados da classificação kNN re-ranqueados. Podemos observar que os melhores resultados foram obtidos por RSNV2 em RDPAC. Todos os resultados superaram os resultados dos modelos de aprendizado profundo da *baseline* ao usar o método de classificação conforme mostrado na Tabela 5.

A Tabela 3 apresenta os resultados com tarefas de fusão usando as características extraídas de EfficientNetB3, ResnetV2 e ViTBasePatch32. Podemos ver que a maioria dos resultados superam os resultados se comparados com os experimentos que não utilizaram fusão da Tabela 1. Podemos destacar a configuração RS+VT no método LHRR com $0,2790$ de MAP e configuração E+R+V no método LHRR com $0,2846$ de MAP.

**Tabela 1** – Resultados para EfficientNetB3, ResnetV2 e ViTBasePatch32 com Aprendizado não supervisionado de variedades para o conjunto de dados Bhuvaji.

| Configuração | Método | P@5 | P@10 | P@20 | MAP |
|---|---|---|---|---|---|
| GLCM | - | 0.5170 | 0.4673 | 0.4331 | 0.1628 |
| GLCM | CPRR | 0.5389 | 0.4732 | 0.4326 | 0.1642 |
| GLCM | LHRR | 0.5301 | 0.4646 | 0.4253 | **0.1645** |
| GLCM | RDPAC | **0.5468** | **0.4804** | **0.4368** | 0.1643 |
| EFNB3 | - | 0.6606 | 0.5735 | 0.5140 | 0.1659 |
| EFNB3 | CPRR | 0.6387 | 0.5832 | 0.5428 | 0.1817 |
| EFNB3 | LHRR | 0.6303 | 0.5714 | 0.5373 | **0.2031** |
| EFNB3 | RDPAC | **0.6771** | **0.5976** | **0.5481** | 0.1829 |
| RSNV2 | - | 0.7398 | 0.6694 | 0.6160 | 0.2471 |
| RSNV2 | CPRR | 0.7161 | 0.6677 | 0.6319 | 0.2584 |
| RSNV2 | LHRR | 0.6942 | 0.6547 | 0.6247 | 0.2609 |
| RSNV2 | RDPAC | **0.7488** | **0.6890** | **0.6471** | **0.2636** |
| VIT32 | - | 0.7387 | 0.6611 | 0.6081 | 0.2477 |
| VIT32 | CPRR | 0.7059 | 0.6541 | 0.6147 | 0.2544 |
| VIT32 | LHRR | 0.6925 | 0.6488 | 0.6143 | 0.2525 |
| VIT32 | RDPAC | **0.7430** | **0.6782** | **0.6302** | **0.2615** |

**Tabela 2** – Resultados para EfficientNetB3, ResnetV2 e ViTBasePatch32 na classificação kNN para o conjunto de dados Bhuvaji.

| Configuração | Método | Acurácia |
|---|---|---|
| EFNB3 | - | 0.2905 |
| EFNB3 | CPRR | 0.6189 |
| EFNB3 | LHRR | 0.6045 |
| EFNB3 | RDPAC | **0.6628** |
| RSNV2 | - | 0.2891 |
| RSNV2 | CPRR | 0.7134 |
| RSNV2 | LHRR | 0.7125 |
| RSNV2 | RDPAC | **0.7243** |
| VIT32 | - | 0.2783 |
| VIT32 | CPRR | 0.6798 |
| VIT32 | LHRR | 0.6742 |
| VIT32 | RDPAC | **0.7212** |

A Tabela 4 mostra os resultados da classificação kNN usando os métodos de fusão no aprendizado não supervisionado com as características extraídas. Os melhores resultados foram encontrados na configuração EF+RS com acurácia de 0,7116 no método CPRR, configuração EF+VT com acurácia de 0,6898 no método CPRR, configuração RS+VT com acurácia de 0,7164 no método CPRR e finalmente configuração E+R+V com acurácia de 0,7155 no método CPRR. Também podemos confirmar que todos os cenários superaram os modelos de aprendizado profundo da Tabela 5.

A Tabela 5 mostra os resultados da classificação em EFNB3, RSNV2 e VIT32, e foi usada como *baseline* para as demais configurações.

- **Cheng resultados do conjunto de dados**

**Tabela 3** – Resultados para os cenários de fusão com EfficientNetB3, ResnetV2 e ViTBasePatch32 no Aprendizado não supervisionado de variedades e extração de características do conjunto de dados Bhuvaji.

| Configuração | Método | P@5 | P@10 | P@20 | MAP |
|---|---|---|---|---|---|
| EF+RS | CPRR | 0.7418 | 0.6949 | **0.6566** | 0.2614 |
| EF+RS | LHRR | 0.7368 | 0.6882 | 0.6556 | **0.2740** |
| EF+RS | RDPAC | **0.7644** | **0.6960** | 0.6409 | 0.1981 |
| EF+VT | CPRR | 0.7272 | 0.6781 | **0.6396** | 0.2554 |
| EF+VT | LHRR | 0.7203 | 0.6779 | 0.6421 | **0.2650** |
| EF+VT | RDPAC | **0.7627** | **0.6886** | 0.6353 | 0.1977 |
| RS+VT | CPRR | 0.7572 | 0.7114 | **0.6809** | 0.2735 |
| RS+VT | LHRR | 0.7398 | 0.7017 | 0.6698 | **0.2790** |
| RS+VT | RDPAC | **0.7886** | **0.7225** | 0.6798 | 0.2710 |
| E+R+V | CPRR | 0.7599 | 0.7145 | **0.6860** | 0.2768 |
| E+R+V | LHRR | 0.7466 | 0.7069 | 0.6805 | **0.2846** |
| E+R+V | RDPAC | **0.7791** | **0.7031** | 0.6469 | 0.1985 |

**Tabela 4** – Resultados para os cenários de fusão com EfficientNetB3, ResnetV2 e ViTBasePatch32 na classificação kNN para o conjunto de dados Bhuvaji.

| Configuração | Método | Acurácia |
|---|---|---|
| EF+RS | CPRR | **0.7116** |
| EF+RS | LHRR | 0.7012 |
| EF+RS | RDPAC | 0.6881 |
| EF+VT | CPRR | **0.6898** |
| EF+VT | LHRR | 0.6702 |
| EF+VT | RDPAC | 0.6742 |
| RS+VT | CPRR | **0.7164** |
| RS+VT | LHRR | 0.7060 |
| RS+VT | RDPAC | 0.7012 |
| E+R+V | CPRR | **0.7155** |
| E+R+V | LHRR | 0.6977 |
| E+R+V | RDPAC | 0.6777 |

**Tabela 5** – Resultados de classificação como *baseline* para os experimentos do conjunto de dados Bhuvaji.

| Configuração | Método | Acurácia |
|---|---|---|
| EFNB3 | Classification | 0.6558 |
| RSNV2 | Classification | 0.4418 |
| VIT32 | Classification | 0.4162 |

A Tabela 6 apresenta os resultados para o conjunto de dados Cheng. Os melhores resultados foram encontrados na configuração EFNB3 com 0,3361 de MAP no método LHRR, configuração RSNV2 com 0,3759 de MAP no método LHRR e configuração VIT32 com 0,3573 de MAP no método RDPAC.

Como podemos observar na Tabela 7, os melhores resultados para a etapa final da Classificação kNN foram encontrados na configuração EFNB3 no método CPRR com acurácia de 0,7029, configuração RSNV2 com acurácia de 0,7360 no método RDPAC e

**Tabela 6** – Resultados para EfficientNetB3, ResnetV2 e ViTBasePatch32 com Aprendizado não supervisionado de variedades no conjunto de dados Cheng.

| Configuração | Método | P@5 | P@10 | P@20 | MAP |
|---|---|---|---|---|---|
| GLCM | - | 0.6312 | 0.5780 | **0.5445** | 0.3003 |
| GLCM | CPRR | 0.6185 | 0.5682 | 0.5420 | **0.3010** |
| GLCM | LHRR | 0.6223 | 0.5720 | 0.5407 | 0.3003 |
| GLCM | RDPAC | **0.6389** | **0.5781** | 0.5438 | 0.3001 |
| EFNB3 | - | 0.7257 | 0.6580 | 0.6125 | 0.2720 |
| EFNB3 | CPRR | 0.7228 | 0.6785 | 0.6490 | 0.2965 |
| EFNB3 | LHRR | 0.7092 | 0.6691 | 0.6392 | **0.3361** |
| EFNB3 | RDPAC | **0.7445** | **0.6900** | **0.6544** | 0.2964 |
| RSNV2 | - | **0.7836** | 0.7272 | 0.6880 | 0.3434 |
| RSNV2 | CPRR | 0.7525 | 0.7142 | 0.6882 | 0.3577 |
| RSNV2 | LHRR | 0.7540 | 0.7191 | 0.6875 | **0.3759** |
| RSNV2 | RDPAC | 0.7820 | **0.7351** | **0.7008** | 0.3612 |
| VIT32 | - | 0.7723 | 0.7168 | 0.6721 | 0.3444 |
| VIT32 | CPRR | 0.7450 | 0.7007 | 0.6707 | 0.3514 |
| VIT32 | LHRR | 0.7354 | 0.6964 | 0.6666 | 0.3565 |
| VIT32 | RDPAC | **0.7747** | **0.7248** | **0.6896** | **0.3573** |

configuração VIT32 com acurácia de 0,7209 no método RDPAC. Todos as configurações, exceto o EFNB3, superaram os modelos de aprendizagem profunda do estado-da-arte, conforme mostrado na Tabela 10.

**Tabela 7** – Resultados para EfficientNetB3, ResnetV2 e ViTBasePatch32 na etapa de classificação kNN para o conjunto de dados Cheng.

| Configuração | Método | Acurácia |
|---|---|---|
| EFNB3 | - | 0.3806 |
| EFNB3 | CPRR | **0.7029** |
| EFNB3 | LHRR | 0.6580 |
| EFNB3 | RDPAC | 0.6944 |
| RSNV2 | - | 0.3643 |
| RSNV2 | CPRR | 0.7156 |
| RSNV2 | LHRR | 0.7082 |
| RSNV2 | RDPAC | **0.7360** |
| VIT32 | - | 0.3549 |
| VIT32 | CPRR | 0.6931 |
| VIT32 | LHRR | 0.6854 |
| VIT32 | RDPAC | **0.7209** |

Para as tarefas de fusão, podemos ver na Tabela 8 que a configuração EF+RS resultou em 0,3983 de MAP usando o método LHRR, configuração EF+VT com 0,3869 de MAP no método LHRR, configuração RS+VT com 0,3832 de MAP no método LHRR e configuração E+R+V com 0,3915 de MAP no método CPRR. Todos os cenários superaram os resultados de MAP em comparação com as características individuais da Tabela 6.

Podemos observar que os resultados finais para a classificação kNN com a combinação de características extraídas na Tabela 9 superaram os demais cenários com

**Tabela 8** – Resultados para os cenários de fusão com EfficientNetB3, ResnetV2 e ViT-BasePatch32 na etapa de Aprendizado não supervisionado de variedades e fusão de características para o conjunto de dados Cheng.

| Configuração | Método | P@5 | P@10 | P@20 | MAP |
|---|---|---|---|---|---|
| EF+RS | CPRR | 0.7752 | 0.7416 | 0.7159 | 0.3826 |
| EF+RS | LHRR | 0.7604 | 0.7398 | 0.7177 | **0.3983** |
| EF+RS | RDPAC | **0.8054** | **0.7545** | **0.7216** | 0.3104 |
| EF+VT | CPRR | 0.7683 | 0.7304 | 0.7041 | 0.3707 |
| EF+VT | LHRR | 0.7601 | 0.7314 | 0.7089 | **0.3869** |
| EF+VT | RDPAC | **0.8112** | **0.7601** | **0.7220** | 0.3086 |
| RS+VT | CPRR | 0.7811 | 0.7517 | 0.7302 | 0.3767 |
| RS+VT | LHRR | 0.7806 | 0.7507 | 0.7284 | **0.3832** |
| RS+VT | RDPAC | **0.8228** | **0.7778** | **0.7425** | 0.3731 |
| E+R+V | CPRR | 0.7892 | 0.7559 | **0.7378** | **0.3915** |
| E+R+V | LHRR | 0.7827 | 0.7515 | 0.7305 | 0.3870 |
| E+R+V | RDPAC | **0.8199** | **0.7693** | 0.7355 | 0.3099 |

características individuais. Obtivemos na configuração EF+RS no método CPRR a acurácia de 0,7368, configuração EF+VT no método RDPAC a acurácia de 0,7286, configuração RS+VT no método CPRR a acurácia de 0,7274 e configuração E+R+V no método LHRR a acurácia de 0,7319. Apenas a configuração EFNB3 no método de classificação apresentou os melhores resultados de acurácia.

**Tabela 9** – Resultados para os cenários de fusão com EfficientNetB3, ResnetV2 e ViTBasePatch32 na classificação kNN para o conjunto de dados Cheng.

| Configuração | Método | Acurácia |
|---|---|---|
| EF+RS | CPRR | 0.7319 |
| EF+RS | LHRR | **0.7368** |
| EF+RS | RDPAC | 0.7205 |
| EF+VT | CPRR | 0.6882 |
| EF+VT | LHRR | 0.6988 |
| EF+VT | RDPAC | **0.7286** |
| RS+VT | CPRR | **0.7274** |
| RS+VT | LHRR | 0.7237 |
| RS+VT | RDPAC | 0.6976 |
| E+R+V | CPRR | 0.7290 |
| E+R+V | LHRR | **0.7319** |
| E+R+V | RDPAC | 0.7168 |

**Tabela 10** – Resultados de classificação como *baseline* para os experimentos usando o conjunto de dados Cheng.

| Configuração | Método | Acurácia |
|---|---|---|
| EFNB3 | Classification | 0.7324 |
| RSNV2 | Classification | 0.6975 |
| VIT32 | Classification | 0.5508 |

A Tabela 10 mostra os resultados da classificação nas configurações EFNB3, RSNV2

e VIT32, e foi utilizada como *baseline* para as demais configurações. Podemos observar que a maioria dos resultados superaram os modelos de *baseline* de aprendizado profundo.

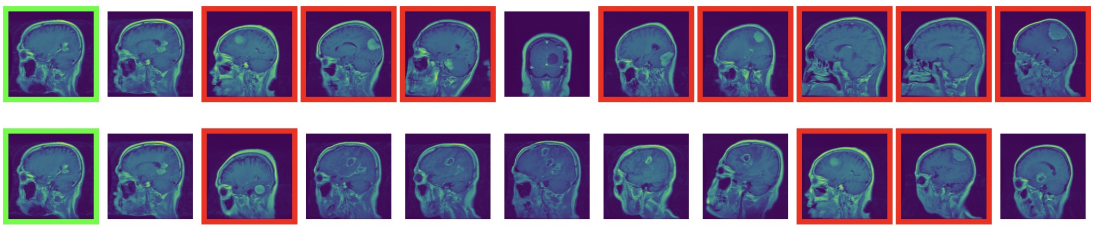


**Figura 35** – Análise visual: resultados de recuperação antes e depois do uso do Aprendizado Não Supervisionado de Variedades (resultados errados em bordas vermelhas).

Na Figura 35 podemos ver uma amostra de RSNV2 com o método RDPAC das listas classificadas antes e depois da avaliação das imagens da etapa Aprendizado Não Supervisionado de Variedades. As imagens da consulta são apresentadas em bordas verdes, enquanto as erradas resultam em bordas vermelhas. Podemos confirmar que após esta etapa temos resultados mais precisos, com menos resultados errados.

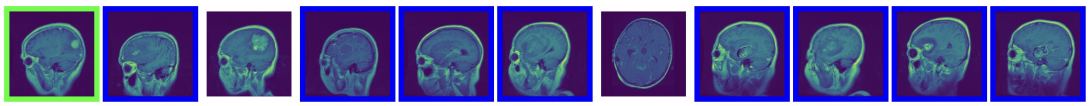


**Figura 36** – Resultados de recuperação usados para classificação kNN: a classe mais comum em bordas azuis.

Na Figura 36 há um exemplo da etapa de Classificação kNN para a configuração RSNV2 com método RDPAC. A imagem de consulta é apresentada em bordas verdes, enquanto as imagens com rótulo mais comum são apresentadas em bordas azuis. Podemos ver neste exemplo que a lista classificada é muito precisa e poderia ser usada pelo kNN para definir o rótulo da imagem de consulta.

# 6 Conclusão

Neste trabalho apresentamos uma nova abordagem para recuperação e classificação de imagens médicas, considerando um cenário não supervisionado ou fracamente supervisionado, avaliado para imagens de ressonância magnética no suporte ao diagnóstico de tumores cerebrais. O cenário em que os médicos têm poucas ou nenhuma imagem para comparar é frequente e representa um desafio fundamental no meio médico.

A avaliação experimental conduzida considerou diferentes características de imagens, desde características globais tradicionais até modelos CNN e *Transformer* do estado-da-arte treinados por meio de *transfer learning*. Os mesmos modelos também foram usados como classificadores de *baseline* considerando o mesmo protocolo experimental com um pequeno conjunto de rótulos para treinamento (apenas 20%).

Considerando os resultados obtidos, algumas conclusões pertinentes podem ser tiradas: (*i*) o uso de múltiplos métodos de aprendizado não supervisionados produz ganhos substanciais tanto em tarefas de recuperação quanto de classificação; (*ii*) na maioria dos cenários, a fusão de características supera os melhores resultados das características individuais, especialmente para tarefas de recuperação; (*iii*) no cenário fracamente supervisionado, a metodologia proposta baseada em características de *transfer learning* e aprendizado de variedades sensível ao contexto obteve melhores resultados de eficácia do que os resultados de *baseline* (dadas pelos mesmos modelos de aprendizado profundo usados por características como classificadores). As contribuições obtidas deram origem a uma publicação em conferência internacional (ANTONIO; PEDRONETTE, 2023).

Apesar dos avanços e contribuições relevantes, alguns trabalhos futuros importantes podem ser destacados. A área médica geralmente exige alta confiança e, portanto, resultados de alta eficácia. Nesse sentido, ainda há espaço para melhorias no cenário desafiador definido onde apenas algumas imagens rotuladas estão disponíveis. Ainda que avaliamos o método com imagens de ressonância magnética de tumor cerebral, novos trabalhos podem aplicar o mesmo método para outros domínios de imagens médicas.

# Referências

AANS. *American Association of Neurological Surgeons*. 2023. Disponível em: <https://https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Brain-Tumors>. Citado na página 72.

AFONSO, L. C. S. et al. Improving optimum- path forest classification using unsupervised manifold learning. In: *24th International Conference on Pattern Recognition, ICPR 2018, Beijing, China, August 20-24, 2018*. IEEE Computer Society, 2018. p. 560–565. Disponível em: <https://doi.org/10.1109/ICPR.2018.8546061>. Citado na página 52.

AGGARWAL, R. et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digital Medicine*, v. 4, n. 1, p. 65, 2021. Disponível em: <https://doi.org/10.1038/s41746-021-00438-z>. Citado na página 17.

AMIN, J. et al. Detection of brain tumor based on features fusion and machine learning. *Journal of Ambient Intelligence and Humanized Computing*, 11 2018. Citado 2 vezes nas páginas 9 e 60.

AMIN, J. et al. Brain tumor classification: Feature fusion. In: *2019 International Conference on Computer and Information Sciences (ICCIS)*. [S.l.: s.n.], 2019. p. 1–6. Citado na página 53.

ANTONIO, A. L. T. D.; PEDRONETTE, D. C. G. Manifold learning for brain tumor mri image retrieval and classification. In: *IEEE International Conference on Bioinformatics and Bioengineering (BIBE)*. [S.l.: s.n.], 2023. Citado na página 81.

ATIA, N. et al. Particle swarm optimization and two-way fixed-effects analysis of variance for efficient brain tumor segmentation. *Cancers*, v. 14, p. 4399, 09 2022. Citado na página 31.

AWAD, M. H. e. A. I. *Image Feature Detectors and Descriptors : Foundations and Applications*. 1. ed. [S.l.]: Springer International Publishing, 2016. (Studies in Computational Intelligence 630). ISBN 978-3-319-28852-9,978-3-319-28854-3. Citado na página 24.

AYADI, W. et al. Brain tumor classification based on hybrid approach. *The Visual Computer*, v. 38, n. 1, p. 107–117, 2022. Citado 3 vezes nas páginas 8, 53 e 54.

AYADI, W. et al. Deep cnn for brain tumor classification. *Neural Processing Letters*, v. 53, n. 1, p. 671–700, 2021. Citado 4 vezes nas páginas 8, 53, 54 e 55.

BAEZA-YATES, B. R.-N. R. *Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca*. [S.l.]: Editora Bookman, 2013. Citado 2 vezes nas páginas 20 e 28.

BAY, H.; TUYTELAARS, T.; GOOL, L. V. Surf: Speeded up robust features. In: . [S.l.: s.n.], 2006. v. 3951, p. 404–417. ISBN 978-3-540-33832-1. Citado na página 24.

BELKIN, M.; NIYOGI, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. In: *NIPS*. [S.l.: s.n.], 2001. Citado na página 51.

BHOWMIK, N. et al. Efficient fusion of multidimensional descriptors for image retrieval. *2014 IEEE International Conference on Image Processing (ICIP)*, p. 5766–5770, 2014. Citado 2 vezes nas páginas 47 e 48.

BHUVAJI, S. *Brain tumor dataset*. 2020. Disponível em: <https://github.com/SartajBhuvaji/Brain-Tumor-Classification-DataSet>. Citado na página 73.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. Citado 2 vezes nas páginas 26 e 42.

BOSCH, A.; ZISSERMAN, A.; MUÑOZ, X. Representing shape with a spatial pyramid kernel. In: *CIVR '07*. [S.l.: s.n.], 2007. Citado na página 37.

CALONDER, M. et al. Brief: Binary robust independent elementary features. In: . [S.l.: s.n.], 2010. v. 6314, p. 778–792. ISBN 978-3-642-15560-4. Citado na página 25.

CANCER.NET. *American Society of Clinical Oncology - Brain Tumor: Statistics*. 2023. Disponível em: <https://www.cancer.net/cancer-types/brain-tumor/statistics>. Citado na página 72.

CHAPELLE BERNHARD SCHöLKOPF, A. Z. O. *Semi-Supervised Learning*. [S.l.]: The MIT Press, 2006. (Adaptive Computation and Machine Learning series). ISBN 0262033585; 9780262033589. Citado 2 vezes nas páginas 16 e 27.

CHAWLA, M. et al. A method for automatic detection and classification of stroke from brain ct images. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, v. 2009, p. 3581–4, 09 2009. Citado na página 33.

CHENG, J. *Brain tumor dataset*. 2017. Disponível em: <https://doi.org/10.6084/m9.figshare.1512427.v5>. Citado 2 vezes nas páginas 56 e 73.

CHENG, J. *Sartajbhuvaji/brain-tumor-classification-dataset*. 2017. (Acessado em 07/31/2022). Disponível em: <https://doi.org/10.6084/m9.figshare.1512427.v5>. Citado 2 vezes nas páginas 54 e 55.

CHUNG, S. H. et al. Macroscopic optical physiological parameters correlate with microscopic proliferation and vessel area breast cancer signatures. *Breast cancer research : BCR*, v. 17, p. 72, 05 2015. Citado na página 35.

COSTA, A. F.; MAMANI, G. H.; TRAINA, A. An efficient algorithm for fractal analysis of textures. *Brazilian Symposium of Computer Graphic and Image Processing*, 08 2012. Citado na página 39.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, v. 13, n. 1, p. 21–27, 1967. Citado na página 46.

COYNE, K. *MRI: A Guided Tour - MagLab*. 2021. (Acessado em 10/31/2021). Disponível em: <https://nationalmaglab.org/magnet-academy/read-science-stories/science-simplified/mri-a-guided-tour/>. Citado na página 30.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2009. p. 248–255. Citado 2 vezes nas páginas 64 e 74.

DONOSER, M.; BISCHOF, H. Diffusion processes for retrieval revisited. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2013. p. 1320–1327. Citado na página 17.

D'OR, R. *Tumor na hipófise: O que É, sintomas, tratamentos e causas*. 2022. (Acessado em 07/25/2022). Disponível em: <https://www.rededorsaoluiz.com.br/doencas/tumor-na-hipofise>. Citado na página 74.

DOSOVITSKIY, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. Disponível em: <http://dblp.uni-trier.de/db/journals/corr/corr2010.html#abs-2010-11929>. Citado 4 vezes nas páginas 9, 65, 67 e 68.

GCO. *Global Cancer Observatory*. 2020. Disponível em: <https://gco.iarc.fr/today/online-analysis-pie?v=2020&mode=population&cancer=31>. Citado na página 72.

GE, C. et al. Deep semi-supervised learning for brain tumor classification. *BMC Medical Imaging*, v. 20, n. 1, p. 87, 2020. Citado 4 vezes nas páginas 9, 53, 56 e 57.

GOODFELLOW, I. J.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. <http://www.deeplearningbook.org>. Citado 4 vezes nas páginas 16, 25, 26 e 40.

HAQUE, K. F.; ABDELGAWAD, A. A deep learning approach to detect covid-19 patients from chest x-ray images. *AI*, v. 1, n. 3, p. 418–435, 2020. ISSN 2673-2688. Disponível em: <https://www.mdpi.com/2673-2688/1/3/27>. Citado na página 32.

HARALICK, R.; SHANMUGAM, K.; DINSTEIN, I. Textural features for image classification. *IEEE Trans Syst Man Cybern*, SMC-3, p. 610–621, 01 1973. Citado 2 vezes nas páginas 37 e 65.

HE, K. et al. *Deep Residual Learning for Image Recognition*. 2015. Citado 2 vezes nas páginas 64 e 66.

HOPKINS, J. *Johns Hopkins Medicine*. 2020. (Acessado em 07/20/2022). Disponível em: <https://www.hopkinsmedicine.org/international/portugues/conditions-treatments/neurosurgery/gliomas.html>. Citado na página 73.

HOPKINS, J. *Johns Hopkins Medicine*. 2020. (Acessado em 07/20/2022). Disponível em: <https://www.hopkinsmedicine.org/international/portugues/conditions-treatments/neurosurgery/meningioma.html>. Citado na página 74.

HUANG, C.-B.; LIU, Q. An orientation independent texture descriptor for image retrieval. In: *2007 International Conference on Communications, Circuits and Systems*. [S.l.: s.n.], 2007. p. 772–776. Citado na página 24.

KANG, J.; ULLAH, Z.; GWAK, J. Mri-based brain tumor classification using ensemble of deep features and machine learning classifiers. *Sensors*, v. 21, n. 6, 2021. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/21/6/2222>. Citado 4 vezes nas páginas 9, 53, 57 e 58.

KHALIFA, E. A.; HASSANEIN, S. A. hamid; EID, H. Ultrasound measurement of fetal abdominal subcutaneous tissue thickness as a predictor of large versus small fetuses for gestational age. *Egyptian Journal of Radiology and Nuclear Medicine*, v. 50, p. 1–8, 2019. Citado na página 33.

KIMURA, P. A. S. et al. Evaluating retrieval effectiveness of descriptors for searching in large image databases. *Journal of Information and Data Management*, v. 2, n. 3, p. 305, 2022/07/25 2011. Disponível em: <https://sol.sbc.org.br/journals/index.php/jidm/article/view/1411>. Citado na página 47.

KUNCHEVA, L. I. *Combining Pattern Classifiers: Methods and Algorithms*. [S.l.]: Wiley, 2004. Citado na página 27.

LAKSHMI, A.; ARIVOLI, T. Computer aided diagnosis system for brain tumor detection and segmentation. *Journal of Theoretical and Applied Information Technology*, v. 64, p. 561–567, 06 2014. Citado na página 30.

LECUN, Y. et al. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, v. 1, p. 541–551, 1989. Citado 2 vezes nas páginas 16 e 40.

LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998. ISSN 0018-9219. Citado 3 vezes nas páginas 39, 40 e 41.

LISIN, D. A. et al. Combining local and global image features for object class recognition. In: *CVPR Workshops*. IEEE Computer Society, 2005. p. 47. ISBN 0-7695-2372-2. Disponível em: <http://dblp.uni-trier.de/db/conf/cvpr/cvprw2005.html#LisinMBLB05>. Citado na página 24.

LOWE, D. Object recognition from local scale-invariant features. *Proceedings of the IEEE International Conference on Computer Vision*, v. 2, 01 2001. Citado na página 24.

LOWEDAVID, G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004. Citado 2 vezes nas páginas 39 e 47.

MA, Y.; FU, Y. *Manifold Learning Theory and Applications*. Taylor & Francis, 2011. ISBN 9781439871096. Disponível em: <https://books.google.de/books?id=LjeGZwEACAAJ>. Citado 3 vezes nas páginas 8, 50 e 51.

MAATEN, L. van der; HINTON, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, v. 9, p. 2579–2605, 11 2008. Citado na página 51.

MARCHI, L. M. L. D. *Hands-On Neural Networks: Learn how to build and train your first neural network model using Python*. [S.l.]: Packt Publishing, 2019. ISBN 1788992598, 978-01788992596. Citado na página 42.

MCCONNELL, R. K. Method of and apparatus for pattern recognition. 1 1986. Disponível em: <https://www.osti.gov/biblio/6007283>. Citado 2 vezes nas páginas 37 e 47.

MCINNES, L.; HEALY, J. Umap: Uniform manifold approximation and projection for dimension reduction. 02 2018. Citado na página 51.

MEHTRE, B. M. et al. Color matching for image retrieval. *Pattern Recognit. Lett.*, v. 16, p. 325–331, 1995. Citado na página 24.

MICCAI. *The Medical Image Computing and Computer Assisted Intervention Society*. 2023. (Acessado em 06/06/2023). Disponível em: <http://www.miccai.org/special-interest-groups/challenges/miccai-registered-challenges/>. Citado na página 57.

MICHELUCCI, U. *Advanced Applied Deep Learning: Convolutional Neural Networks and Object Detection*. [S.l.]: Apress, 2019. ISBN 1484249755,9781484249758. Citado na página 41.

MIKOLAJCZYK, K.; SCHMID, C. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, v. 27, p. 1615–30, 11 2005. Citado na página 24.

MURPHY, K. P. *Machine learning: a probabilistic perspective*. Cambridge, MA: [s.n.], 2013. Citado 4 vezes nas páginas 16, 17, 25 e 26.

OISETH LINDSAY JONES, E. M. S. *Computed Tomography (CT) | Concise Medical Knowledge*. 2021. (Acessado em 10/31/2021). Disponível em: <https://www.lecturio.com/concepts/computed-tomography-ct/>. Citado na página 33.

OJALA, T.; PIETIKÄINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognit.*, v. 29, p. 51–59, 1996. Citado 2 vezes nas páginas 37 e 47.

ÖKSÜZ, C.; URHAN, O.; GÜLLÜ, M. K. Brain tumor classification using the fused features extracted from expanded tumor region. *Biomedical Signal Processing and Control*, v. 72, p. 103356, 2022. ISSN 1746-8094. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1746809421009538>. Citado 3 vezes nas páginas 9, 53 e 59.

OLIVA, A.; TORRALBA, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, v. 42, p. 145–175, 05 2001. Citado na página 36.

PATIL, S.; TALBAR, S. Content based image retrieval using various distance metrics. In: KANNAN, R.; ANDRES, F. (Ed.). *Data Engineering and Management*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. p. 154–161. ISBN 978-3-642-27872-3. Citado na página 22.

PEDRONETTE, D. et al. Multimedia retrieval through unsupervised hypergraph-based manifold ranking. *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, PP, 06 2019. Citado 3 vezes nas páginas 17, 68 e 69.

PEDRONETTE, D.; VALEM, L.; LATECKI, L. J. Efficient rank-based diffusion process with assured convergence. *Journal of Imaging*, v. 7, p. 49, 03 2021. Citado 2 vezes nas páginas 68 e 70.

PEDRONETTE, D. C. G.; GONÇALVES, F. M. F.; GUILHERME, I. R. Unsupervised manifold learning through reciprocal knn graph and connected components for image retrieval tasks. *Pattern Recognit.*, v. 75, p. 161–174, 2018. Disponível em: <https://doi.org/10.1016/j.patcog.2017.05.009>. Citado na página 52.

PEDRONETTE, D. C. G. et al. Multimedia retrieval through unsupervised hypergraph-based manifold ranking. *IEEE Trans. Image Process.*, v. 28, n. 12, p. 5824–5838, 2019. Disponível em: <https://doi.org/10.1109/TIP.2019.2920526>. Citado na página 52.

PEDRONETTE, D. C. G.; VALEM, L. P.; TORRES, R. da S. A bfs-tree of ranking references for unsupervised manifold learning. *Pattern Recognit.*, v. 111, p. 107666, 2021. Disponível em: <https://doi.org/10.1016/j.patcog.2020.107666>. Citado na página 52.

PERRIN, I. C. J. *Alzheimer's disease, PET scan - Stock Image - C038/4517 - Science Photo Library*. 2021. (Acessado em 10/31/2021). Disponível em: <https://www.sciencephoto.com/media/910382/view/alzheimer-s-disease-pet-scan>. Citado na página 34.

PIRAS, L.; GIACINTO, G. Information fusion in content based image retrieval: A comprehensive overview. *Inf. Fusion*, v. 37, p. 50–60, 2017. Citado na página 48.

ROTHMAN, D. *Transformers for Natural Language Processing Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more*. [S.l.]: Packt Publishing Ltd, 2021. (Expert Inside). Citado na página 43.

ROWEIS, S. T.; SAUL, L. K. Nonlinear dimensionality reduction by locally linear embedding. *Science*, v. 290 5500, p. 2323–6, 2000. Citado na página 51.

ROZIN, B. et al. A rank-based framework through manifold learning for improved clustering tasks. *Inf. Sci.*, v. 580, p. 202–220, 2021. Disponível em: <https://doi.org/10.1016/j.ins.2021.08.080>. Citado na página 52.

RUBLEE, E. et al. Orb: an efficient alternative to sift or surf. In: . [S.l.: s.n.], 2011. p. 2564–2571. Citado na página 25.

SERRE, T.; WOLF, L.; POGGIO, T. Object recognition with features inspired by visual cortex. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, v. 2, p. 994 – 1000 vol. 2, 07 2005. Citado na página 40.

SHAHIN, A. I.; ALY, S.; ALY, W. A novel multi-class brain tumor classification method based on unsupervised pcanet features. *Neural Computing and Applications*, v. 35, n. 15, p. 11043–11059, 2023. Citado 4 vezes nas páginas 8, 53, 55 e 56.

STEHLING, R.; NASCIMENTO, M.; FALCãO, A. A compact and efficient image retrieval approach based on border/interior pixel classification. In: . [S.l.: s.n.], 2002. p. 102. Citado na página 24.

TAN, M.; LE, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. arXiv, 2019. Disponível em: <https://arxiv.org/abs/1905.11946>. Citado 3 vezes nas páginas 9, 65 e 66.

TAN, M.; LE, Q. V. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*. 2020. Citado na página 64.

TANG, X. Texture information in run-length matrices. *Image Processing, IEEE Transactions on*, v. 7, p. 1602 – 1609, 12 1998. Citado na página 38.

TAO, B.; DICKINSON, B. Texture recognition and image retrieval using gradient indexing. *Journal of Visual Communication and Image Representation*, v. 11, p. 327–342, 09 2000. Citado na página 24.

TCGA. *The Cancer Genoma Atlas*. 2023. (Acessado em 06/06/2023). Disponível em: <https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas>. Citado na página 57.

TEKIN, H. O.; KARA, Analysis of filtering material and its effect on x-ray features by using monte carlo method for medical imaging applications. *RAD Conference*, 01 2016. Citado na página 32.

TENENBAUM, J. B.; SILVA, V. D.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science*, v. 290 5500, p. 2319–23, 2000. Citado na página 51.

TORRES, R. da S.; FALCÃO, A. X. Content-based image retrieval: Theory and applications. *RITA*, v. 13, p. 161–185, 2006. Citado 2 vezes nas páginas 22 e 23.

VALEM, L.; PEDRONETTE, D. Unsupervised similarity learning through cartesian product of ranking references for image retrieval tasks. In: . [S.l.: s.n.], 2016. p. 249–256. Citado na página 49.

VALEM, L.; PEDRONETTE, D. An unsupervised distance learning framework for multimedia retrieval. In: . [S.l.: s.n.], 2017. p. 107–111. Citado 2 vezes nas páginas 69 e 70.

VALEM, L. P.; PEDRONETTE, D. C. G. Unsupervised similarity learning through cartesian product of ranking references for image retrieval tasks. In: *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. [S.l.: s.n.], 2016. p. 249–256. Citado na página 68.

VALEM, L. P. et al. Manifold correlation graph for semi-supervised learning. In: *2018 International Joi Rio de Janeiro, Brazil, July 8-13, 2018*. IEEE, 2018. p. 1–7. Disponível em: <https://doi.org/10.1109/IJCNN.2018.8489487>. Citado na página 52.

VALEM, L. P.; PEDRONETTE, D. C. G. a. An unsupervised distance learning framework for multimedia retrieval. In: *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. New York, NY, USA: ACM, 2017. (ICMR '17), p. 107–111. ISBN 978-1-4503-4701-3. Citado na página 68.

VANDERPLAS, J. *Python Data Science Handbook: Essential Tools for Working with Data*. 1. ed. [S.l.]: O'Reilly Media, 2016. ISBN 1491912057,978-1-491-91205-8,137-140-141-1. Citado 2 vezes nas páginas 8 e 52.

VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). *NIPS*. [s.n.], 2017. p. 5998–6008. Disponível em: <http://dblp.uni-trier.de/db/conf/nips/nips2017.html#VaswaniSPUJGKP17>. Citado 4 vezes nas páginas 8, 43, 44 e 45.

VLAARDINGERBROEK, D. I. J. A. d. B. a. D. I. M. T. *Magnetic Resonance Imaging: Theory and Practice*. [S.l.]: Springer Berlin Heidelberg, 1999. ISBN 978-3-662-03802-4,978-3-662-03800-0. Citado na página 30.

WILLIAMS, A.; YOON, P. Content-based image retrieval using joint correlograms. *Multimedia Tools Appl.*, Kluwer Academic Publishers, USA, v. 34, n. 2, p. 239–248, ago. 2007. ISSN 1380-7501. Disponível em: <https://doi.org/10.1007/s11042-006-0087-2>. Citado na página 24.

WU, B. et al. Visual transformers: Token-based image representation and processing for computer vision. *CoRR*, abs/2006.03677, 2020. Disponível em: <http://dblp.uni-trier.de/db/journals/corr/corr2006.html#abs-2006-03677>. Citado na página 68.

XU, R.; WUNSCH, D. Survey of clustering algorithms. *IEEE Transactions on neural networks*, Ieee, v. 16, n. 3, p. 645–678, 2005. Citado na página 16.

YANG, G.; YE, Q.; XIA, J. Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion*, v. 77, p. 29–52, 2022. ISSN 1566-2535. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1566253521001597>. Citado na página 17.

YANG, J. et al. Evaluating bag-of-visual-words representations in scene classification. In: . [S.l.: s.n.], 2007. p. 197–206. Citado na página 24.

YUE, J. et al. Content-based image retrieval using color and texture fused features. *Mathematical and Computer Modelling*, v. 54, n. 3, p. 1121–1127, 2011. ISSN 0895-7177. Mathematical and Computer Modeling in agriculture (CCTA 2010). Disponível em: <https://www.sciencedirect.com/science/article/pii/S0895717710005352>. Citado na página 47.

ZHANG, W.; QIN, Z.; WAN, T. Image scene categorization using multi-bag-of-features. In: . [S.l.: s.n.], 2011. p. 1804–1808. Citado na página 49.

ZHENG, A. J. X. a. P. N. *Statistical learning and pattern analysis for image and video processing*. 1. ed. [S.l.]: Springer-Verlag London, 2009. (Advances in pattern recognition). ISBN 1848823118,9781848823112,9781848823129,1848823126. Citado na página 49.

ZHENG, L.; YANG, Y.; TIAN, Q. *SIFT Meets CNN: A Decade Survey of Instance Retrieval*. 2017. Citado na página 25.