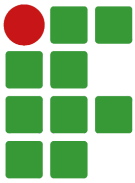


INSTITUTO FEDERAL
São Paulo
Câmpus Campinas

Raphael Tavares Shimamoto

Classificação de Galáxias do tipo Via Láctea (SBb - SBc) com Redes Neurais Convolucionais

Campinas
2024

Raphael Tavares Shimamoto

Classificação de Galáxias do tipo Via Láctea (SBb - SBc) com Redes Neurais Convolucionais

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do diploma do Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia Campus Campinas.

Orientador: Prof. Me. Antonio Queiroz da Silva Neto

Campinas

2024

Ficha Catalográfica
Instituto Federal de São Paulo – Campus Campinas
Biblioteca “Pedro Augusto Pinheiro Fantinatti”
Rosangela Gomes - CRB8/8461

Shimamoto, Raphael Tavares
S556c Classificação de galáxias do tipo via láctea (SBb - SBc) com redes neurais convolucionais / Raphael Tavares Shimamoto. – Campinas, SP: [s.n.], 2024.
41 f. : il.

Orientador: Me. Antônio Queiroz da Silva Neto
Trabalho de Conclusão de Curso (graduação) – Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Campus Campinas. Curso de Tecnologia em Análise e Desenvolvimento de Sistemas, 2024.

1. Astrofísica computacional. 2. Galáxias. 3. Aprendizado de máquina. I. Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Campus Campinas, Curso de Análise e Desenvolvimento de Sistemas. II. Título.

ATA N.º 31/2024 - TADS-CMP/DAE-CMP/DRG/CMP/IFSP

Ata de Defesa de Trabalho de Conclusão de Curso - Graduação

Na presente data, realizou-se a sessão pública de defesa do Trabalho de Conclusão de Curso intitulado "Classificação de Galáxias do tipo Via Láctea (SBb - SBc) com Redes Neurais Convolucionais", apresentado(a) pelo(a) estudante Raphael Tavares Shimamoto do Curso **SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS** (campus Campinas) Os trabalhos foram iniciados às 19:00 h pelo(a) Professor(a) presidente da banca examinadora, constituída pelos seguintes membros:

Membros	Instituição	Presença (Sim/Não)
Antonio Queiroz da Silva Neto (Presidente/Orientador)	IFSP	Sim
Ricardo Barz Sovat	IFSP	Sim
Marcos Brandao Rios	IFSP	Sim

Observações:

A banca examinadora, tendo terminado a apresentação do conteúdo da monografia, passou à arguição do(a) candidato(a). Em seguida, os examinadores reuniram-se para avaliação e deram o parecer final sobre o trabalho apresentado pelo(a) estudante, tendo sido atribuído o seguinte resultado:

[X] Aprovado(a) [ ] Reprovado(a)

Proclamados os resultados pelo presidente da banca examinadora, foram encerrados os trabalhos e, para constar, eu lavrei a presente ata que assino em nome dos demais membros da banca examinadora.

Campus Campinas, 7 de dezembro de 2024

Documento assinado eletronicamente por:

- **Antonio Queiroz da Silva Neto**, **PROF ENS BAS TEC TECNOLOGICO-SUBSTITUTO** , em 07/12/2024 16:07:22.
- **Ricardo Barz Sovat**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 07/12/2024 23:14:48.
- **Marcos Brandao Rios**, **PROF ENS BAS TEC TECNOLOGICO-SUBSTITUTO** , em 09/12/2024 10:19:07.

Este documento foi emitido pelo SUAP em 06/12/2024. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifsp.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 857071
Código de Autenticação: d396483524

RESUMO

As galáxias exibem uma variedade de formas distintas, podendo ser classificadas de acordo com o sistema de categorização proposto por Hubble. Este sistema morfológico reconhece quatro categorias principais: galáxias elípticas (E), galáxias espirais (S), galáxias espirais barradas (SB) e galáxias irregulares (I). Este estudo foca na identificação de galáxias dos tipos SBb e SBc utilizando redes neurais convolucionais (CNNs). Os dados para esta pesquisa são provenientes do Sloan Digital Sky Survey (SDSS) e do catálogo Galaxy Zoo 2, acessados via Astroquery. A classificação eficiente de galáxias é crucial na astronomia moderna para compreender a natureza e a evolução do Universo. A aplicação de CNNs pode aprimorar significativamente o processo de classificação ao automatizar a identificação de tipos de galáxias com base em dados de imagem de alta resolução. O estudo envolve o pré-processamento dos dados para filtrar as galáxias relevantes, aplicando um algoritmo de ranqueamento baseado em redshift para priorizar as galáxias mais próximas, e o treinamento das CNNs em conjuntos de dados balanceados de galáxias SBb e SBc. Esta abordagem visa melhorar a precisão e a confiabilidade da classificação de galáxias, fornecendo insights valiosos sobre as características morfológicas das galáxias espirais barradas.

Palavras-chave: astrofísica computacional; galáxias; aprendizado de máquina.

ABSTRACT

Galaxies exhibit a variety of distinct shapes and can be classified according to the Hubble categorization system. This morphological system recognizes four main categories: elliptical galaxies (E), spiral galaxies (S), barred spiral galaxies (SB), and irregular galaxies (I). This study focuses on the identification of SBb and SBc galaxies using convolutional neural networks (CNNs). The data for this research come from the Sloan Digital Sky Survey (SDSS) and the Galaxy Zoo 2 catalog, accessed via Astroquery. Efficient galaxy classification is crucial in modern astronomy to understand the nature and evolution of the Universe. The application of CNNs can significantly improve the classification process by automating the identification of galaxy types based on high-resolution image data. The study involves preprocessing the data to filter out relevant galaxies, applying a redshift-based ranking algorithm to prioritize the nearest galaxies, and training CNNs on balanced datasets of SBb and SBc galaxies. This approach aims to improve the accuracy and reliability of galaxy classification, providing valuable insights into the morphological characteristics of barred spiral galaxies.

Keywords: computational astrophysics; galaxies; machine learning.

LISTA DE ILUSTRAÇÕES

Figura 1 – Classificação morfológica. 12
Figura 2 – Fluxograma das tarefas de classificação para GZ2 17
Figura 3 – Estrutura de um neurônio artificial 19
Figura 4 – Processo da convolução 21
Figura 5 – Modelo de matriz 28
Figura 6 – Classificação de Arquivos FITS de Acordo com os Valores de Redshift . . . 30
Figura 7 – Exemplo de *Early Stopping* indicando onde o treinamento deve parar para evitar sobreajuste. 33
Figura 8 – Matriz de confusão 37
Figura 9 – Curva ROC 38

LISTA DE TABELAS

Tabela 1 – Tabela reduzida contendo apenas galáxias do tipo SBb e suas variações . . . 29
Tabela 2 – Objetos ranqueados por redshift 31
Tabela 3 – Arquitetura da Rede Neural Convolucional (CNN) 32
Tabela 4 – Distribuição total dos dados entre as classes para a validação cruzada K-Fold. 34
Tabela 5 – Resultados da Validação no Fold 5. 35
Tabela 6 – Resultados das métricas de validação para as *5 folds* do processo de K-Fold Cross Validation. 36

SUMÁRIO

1 INTRODUÇÃO 11

2 JUSTIFICATIVA 14

3 OBJETIVOS 15
3.1 Objetivo Geral 15
3.2 Objetivos Específicos 15

4 FUNDAMENTAÇÃO TEÓRICA 16
4.1 Classificação Morfológica de Galáxias 16
4.1.1 Classificação de Hubble 16
4.1.2 Classificação do Galaxy Zoo 16
4.2 Redshift e sua Relevância na Classificação de Galáxias 18
4.3 Deep Learning 18
4.3.1 Redes Neurais 18
4.3.2 Função de Ativação 20
4.4 Convolução 20
4.4.1 Componentes da Convolução 20
4.5 Redes Neurais Convolucionais 21
4.6 Métricas 22
4.6.1 Matriz de Confusão 22
4.6.2 Acurácia 22
4.6.3 Recall 23
4.6.4 Precisão 23
4.6.5 F1-score 23
4.6.6 Curva ROC 24
4.7 Trabalhos Relacionados 24
4.7.1 Deep Galaxy: Classification of Galaxies based on Deep Convolutional Neural Networks. 24
4.7.2 Machine and Deep Learning Applied to Galaxy Morphology. 25

5 METODOLOGIA 27
5.1 Ferramentas e Etapas 27
5.2 Segmentação do Catálogo Galaxy Zoo 2 por Tipos Morfológicos 28
5.3 Ranking de objetos por Redshift 30
5.4 Arquitetura implementada 31
5.4.1 Treinamento do modelo 32

5.4.2 Divisão dos Dados e Validação Cruzada 34

6 RESULTADOS E DISCUSSÕES 35
6.1 Resultados da Validação 35
6.2 Análise da Matriz de confusão 36
6.2.1 Análise da Curva ROC 38

7 CONCLUSÃO 40
7.1 Conclusão e Trabalhos Futuros 40

REFERÊNCIAS 41

1 INTRODUÇÃO

Desde os primórdios da observação astronômica, a fascinação pelo cosmos tem sido alimentada pela exploração das galáxias, essas vastas coleções de estrelas, poeira cósmica e matéria escura. A jornada para compreender a natureza e a diversidade das galáxias remonta a séculos atrás, quando os primeiros astrônomos começaram a mapear e catalogar esses objetos celestes. O primeiro grande passo na compreensão das galáxias ocorreu no início do século XX, quando o astrônomo Edwin Hubble, utilizando o telescópio Hooker no Observatório Mount Wilson, realizou observações pioneiras que revelaram a verdadeira escala do universo. Ao medir as distâncias para várias galáxias e correlacioná-las com seus espectros, Hubble demonstrou que muitas dessas nebulosas (antes identificadas como tal) eram, na verdade, outras galáxias semelhantes à Via Láctea. Esse marco histórico inaugurou uma nova era na astronomia, marcando o nascimento da cosmologia moderna e estimulando um interesse renovado no estudo das galáxias.

As técnicas comuns que envolvem o estudo dos objetos astronômicos são conhecidas como fotometria e espectroscopia. Os estudos fotométricos e espectroscópicos desempenham papéis fundamentais na astrofísica contemporânea, fornecendo insights cruciais sobre as propriedades físicas e químicas dos objetos celestes. Os estudos fotométricos envolvem a medição da intensidade da luz emitida por fontes astronômicas em diferentes comprimentos de onda, permitindo a determinação de características como temperatura, luminosidade e composição química das estrelas, galáxias e outros corpos celestes. Por outro lado, os estudos espectroscópicos analisam a distribuição de energia luminosa em função do comprimento de onda, permitindo a identificação de elementos químicos presentes em objetos celestes, bem como a medição de velocidades radiais e outras propriedades dinâmicas. Essas abordagens complementares desempenham um papel crucial na compreensão da formação e evolução de estruturas cósmicas, bem como na investigação de processos físicos e fenômenos astrofísicos complexos.

Um dos aspectos chave de qualquer investigação extragaláctica é a definição de uma amostra imparcial que inclua tipos morfológicos confiáveis. As propriedades morfológicas das galáxias resultam não apenas dos processos internos de formação e evolução, mas também da interação com o ambiente. Galáxias em grupos ou aglomerados podem ter caminhos evolutivos diversos em comparação com as isoladas, o que se reflete claramente em sua morfologia. Portanto, a classificação das galáxias em um sistema taxonômico significativo é de importância primordial para estudos de formação e evolução galáctica.

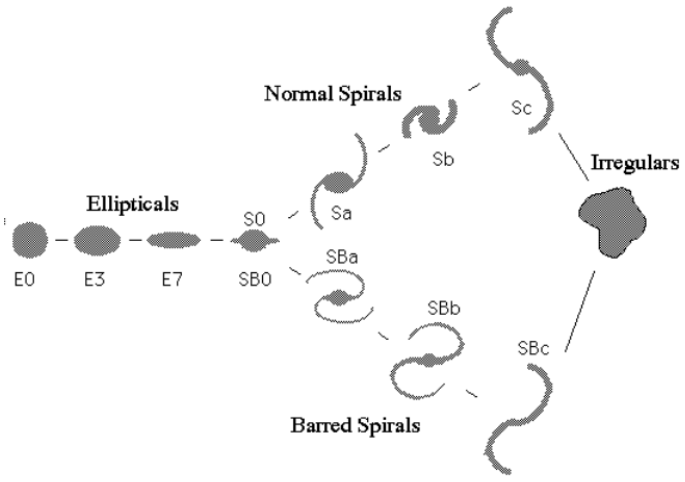
A principal diferenciação das galáxias é entre as Galáxias de Tipo Inicial (ETGs), que são galáxias elípticas com basicamente uma única característica estrutural, e as Galáxias de Tipo Tardio (LTGs), que possuem um disco proeminente (Hubble, 1926). Estudar os tipos e as propriedades das galáxias é importante pois oferece pistas cruciais sobre a origem e o desenvolvimento do universo. A classificação das galáxias desempenha um papel importante no

estudo da formação das galáxias e na avaliação do nosso universo.

Historicamente, a classificação das galáxias envolvia a inspeção visual de imagens bidimensionais das galáxias e sua categorização conforme sua aparência. Embora a classificação humana especializada seja relativamente confiável, é um processo extremamente demorado para grandes quantidades de dados astronômicos coletados recentemente, devido ao aumento do tamanho dos telescópios e das câmeras CCD, que produziram conjuntos de dados extremamente grandes de imagens, como o Sloan Digital Sky Survey (SDSS). Esses dados são muito volumosos para serem analisados manualmente de forma viável.

A classificação das galáxias é baseada em imagens e espectros, sendo essa classificação um objetivo de longo prazo para os astrofísicos. No entanto, a natureza complicada das galáxias e a qualidade das imagens tornaram a classificação das galáxias desafiadora e imprecisa. O sistema de classificação das galáxias ajuda os astrônomos no processo de agrupar as galáxias conforme sua forma visual. O mais famoso é a sequência de Hubble, que é considerada um dos esquemas mais utilizados na classificação morfológica das galáxias. Na Figura 1, podemos observar a sequência de Hubble.

Figura 1 – Classificação morfológica.



Fonte: Disponível em: <https://pages.uoregon.edu/jschombe/ast123/lectures/lec11.html>. Acesso em 16 de março de 2024.

As galáxias possuem características físicas e químicas que podem diferir uma das outras como composições estelares, taxas de formação estelar, composições químicas, presença de gás, poeira cósmica, entre outros aspectos (Kennicutt, 1998).

Trabalhos anteriores consideram a Via Láctea como do tipo SBc como Binney e Tremaine (2011). Investigações mais recentes sugerem que a Via Láctea pode ser uma galáxia do tipo SBb ou SBc (Xu et al., 2023). No entanto, há uma carência de métodos automáticos eficazes para a identificação e diferenciação entre as subcategorias de galáxias espirais barradas, como SBb e SBc, em grandes conjuntos de dados astronômicos.

A classificação precisa das galáxias, particularmente das subcategorias espirais barradas

SBb e SBc, é essencial para desvendar processos fundamentais da formação e evolução galáctica. Contudo, esse é um desafio técnico significativo devido às semelhanças visuais entre essas classes e à falta de métodos automáticos eficazes para analisar os grandes volumes de dados astronômicos disponíveis. Métodos anteriores, como classificadores baseados em Naive Bayes ou Random Forest, não atingiram níveis de precisão suficientemente altos para diferenciar essas subcategorias de maneira confiável. Essa limitação impacta diretamente a capacidade da comunidade científica de testar teorias sobre a evolução do universo.

Para enfrentar esse problema, este trabalho propõe o desenvolvimento de um modelo baseado em Redes Neurais Convolucionais (CNNs) para a classificação automática das galáxias SBb e SBc. As CNNs são particularmente adequadas para essa tarefa, pois conseguem extrair características visuais complexas diretamente das imagens, sem a necessidade de intervenção humana para o design de características específicas. Esse modelo será treinado usando imagens do Sloan Digital Sky Survey Data Release 7 (SDSS-DR7), uma das mais ricas bases de dados astronômicos disponíveis.

A solução proposta utiliza a capacidade das CNNs de lidar com grandes conjuntos de dados multidimensionais para identificar padrões sutis que diferenciam as galáxias SBb e SBc. Com isso, espera-se alcançar uma precisão superior à de métodos tradicionais, melhorando significativamente a eficiência da classificação morfológica e contribuindo para o avanço das investigações sobre a formação e a evolução galáctica.

2 JUSTIFICATIVA

O estudo da classificação morfológica de galáxias é de grande relevância para a astrofísica, uma vez que a morfologia galáctica está diretamente relacionada à formação e evolução das galáxias. No contexto das galáxias espirais barradas, a compreensão detalhada de suas subcategorias (como SBb e SBc) pode fornecer insights valiosos sobre a dinâmica interna dessas estruturas e sua evolução ao longo do tempo.

Além disso, com o crescente volume de dados obtidos por observatórios astronômicos modernos, há uma necessidade urgente de métodos automatizados e precisos para processar e classificar esses dados. As redes neurais convolucionais (CNNs) são particularmente eficazes no reconhecimento de padrões em imagens, tornando-as ferramentas ideais para a análise de dados astronômicos. Sua aplicação neste trabalho visa atender à demanda por métodos mais eficientes e robustos, oferecendo uma abordagem que não apenas aprimora a precisão das classificações, mas também pode ser aplicada em outras áreas da astronomia computacional.

Essa metodologia tem o potencial de melhorar significativamente a qualidade das análises morfológicas e, ao mesmo tempo, fornecer contribuições importantes para a pesquisa astrofísica, especialmente no estudo da estrutura e dinâmica do universo.

3 OBJETIVOS

3.1 Objetivo Geral

O objetivo geral deste trabalho é aplicar redes neurais convolucionais (CNNs) na classificação morfológica de galáxias espirais barradas, aprimorar a precisão e a eficiência das classificações morfológicas e contribuindo para o avanço da pesquisa em galáxias do tipo da nossa galáxia (Via Láctea).

Por meio da utilização de Redes neurais Convolucionais, espera-se aprofundar o conhecimento sobre as subcategorias SBb e SBc, permitindo uma análise mais detalhada dessas estruturas galácticas e promovendo uma maior compreensão da dinâmica e evolução do universo.

3.2 Objetivos Específicos

1. Desenvolver e treinar um modelo de CNNs utilizando dados observacionais de galáxias espirais barradas.
2. Avaliar a precisão e a eficácia do modelo na classificação das subcategorias SBb e SBc.
3. Identificar potenciais limitações e propor melhorias para futuros estudos.
4. Implementar um modelo capaz de classificar as galáxias SBb e SBc, considerando possíveis desafios técnicos ou limitações nos dados.

4 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, são apresentados os conceitos teóricos essenciais que fundamentam o desenvolvimento do trabalho. Serão discutidos os principais métodos de classificação morfológica de galáxias, com destaque para a classificação de *Hubble* e as contribuições do projeto *Galaxy Zoo*. Em seguida, abordaremos a importância do *redshift* na classificação de galáxias, destacando como essa informação auxilia na seleção das imagens de galáxias mais próximas.

Será introduzido o conceito de *Deep Learning*, com ênfase nas redes neurais artificiais e nas funções de ativação. Posteriormente, discutiremos o processo de convolução e sua aplicação em Redes Neurais Convolucionais (*CNNs*), que são amplamente utilizadas para tarefas de classificação de imagens, incluindo a análise morfológica de galáxias.

Além disso, serão apresentadas as métricas utilizadas para avaliar o desempenho dos modelos, incluindo a matriz de confusão, acurácia, *recall*, precisão e F1-score. Por fim, serão revisados trabalhos relacionados, destacando as contribuições existentes na literatura.

4.1 Classificação Morfológica de Galáxias

A morfologia das galáxias é uma área central de estudo na astronomia, pois fornece pistas cruciais sobre a formação e evolução das galáxias. A classificação morfológica das galáxias tem sido historicamente baseada em inspeção visual, mas avanços em levantamentos astronômicos e técnicas de aprendizado de máquina têm permitido análises mais detalhadas e em maior escala. Este capítulo discute a classificação morfológica de Hubble, com ênfase nas galáxias estudadas no projeto Galaxy Zoo 2 (GZ2), um esforço de ciência cidadã que utilizou classificações visuais de voluntários para categorizar mais de 300.000 galáxias do Sloan Digital Sky Survey (SDSS).

4.1.1 Classificação de Hubble

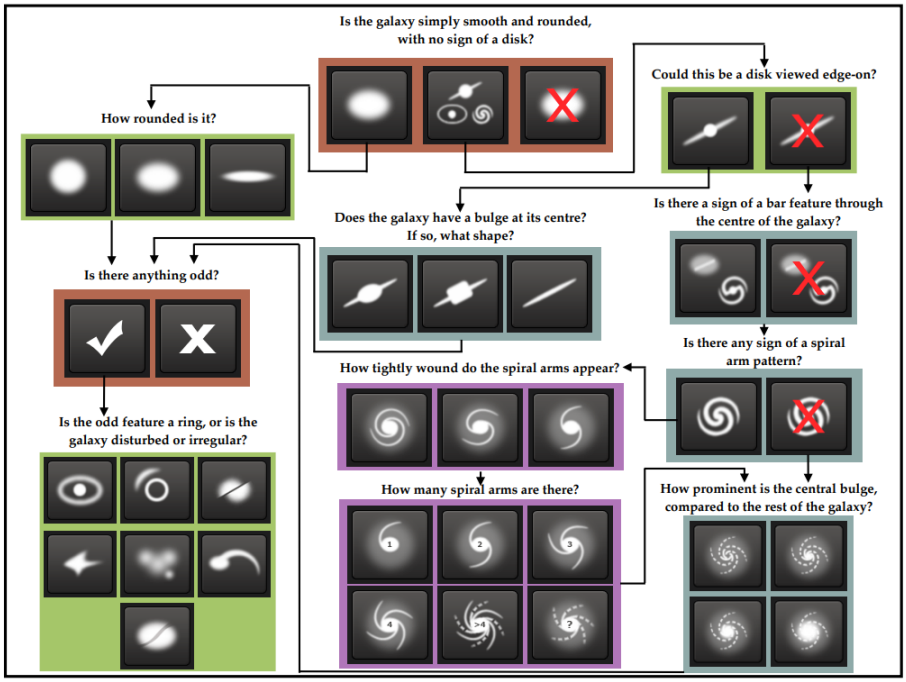
Hubble classificou as galáxias em três classes principais com base em sua morfologia: elípticas (denotadas por "E"), espirais (denotadas por "S"), e irregulares (denotadas por "I"). As galáxias espirais foram ainda subdivididas em duas classes adicionais: "SB"para as espirais barradas e "SA"para as espirais não barradas. As espirais barradas, por sua vez, são divididas em subclassificações SB(a), SB(b) e SB(c), que refletem a abertura e o grau de enrolamento dos braços espirais. As SB(a) possuem braços espirais frouxamente enrolados, enquanto as SB(c) têm braços mais abertos e menos enrolados, sendo as SB(b) intermediárias entre esses dois extremos. Essa classificação é conhecida como Sequência de Hubble e ainda é amplamente utilizada na astronomia.

4.1.2 Classificação do Galaxy Zoo

O Galaxy Zoo, um projeto de ciência cidadã, utilizou classificações visuais de voluntários para categorizar galáxias em várias classes morfológicas. O Galaxy Zoo 2 (GZ2) focou em

características morfológicas detalhadas, como barras, bulbos, e a curvatura dos braços espirais. Isso permitiu uma análise mais rica e detalhada da morfologia galáctica em uma escala sem precedentes. A figura 2 mostra como foram classificados os objetos.

Figura 2 – Fluxograma das tarefas de classificação para GZ2



Fonte: (Willett et al., 2013)

O Galaxy Zoo 2 coletou mais de 16 milhões de classificações morfológicas de 304.122 galáxias do SDSS. A classificação envolveu a identificação de características como a presença de barras, a forma dos discos vistos de lado, e a força relativa dos bulbos e braços espirais. Os dados do GZ2 mostraram uma concordância superior a 90% com as classificações feitas por astrônomos profissionais, especialmente para tipos morfológicos T, barras fortes e curvatura dos braços (**??**).

Para a classificação dos objetos, os voluntários são inicialmente apresentados a uma imagem composta em cores de uma galáxia do SDSS junto com uma pergunta inicial, como "A galáxia é suave, possui características ou disco, ou é uma estrela ou artefato?". Dependendo da resposta dada, o voluntário é guiado por uma série de perguntas subsequentes que coletam informações mais detalhadas sobre a galáxia. Este processo é iterativo, e cada galáxia passa por múltiplas classificações independentes, permitindo uma análise estatística robusta dos dados coletados.

Foi empregado a técnica de "árvores de decisão"para as classificações. As árvores de decisão são fundamentais na metodologia do GZ2 e a resposta (classificação) é o resultado do algoritmo. Cada tarefa dentro da árvore consiste em uma pergunta e um conjunto finito de respostas possíveis. A escolha de uma resposta específica leva o voluntário a uma nova pergunta, configurando uma árvore de decisão que, no caso do GZ2, possui 11 tarefas com um total de 37

respostas possíveis. Este sistema permite uma classificação detalhada e adaptativa de cada galáxia, assegurando que características morfológicas importantes sejam identificadas e registradas.

4.2 Redshift e sua Relevância na Classificação de Galáxias

O redshift, ou desvio para o vermelho, é uma medida da mudança no comprimento de onda da luz emitida por objetos astronômicos, causada pela expansão do universo. Na classificação morfológica de galáxias, o redshift é utilizado para estimar a distância e a velocidade de recessão das galáxias. No contexto do Galaxy Zoo e do SDSS, o redshift fornece uma ferramenta crucial para correlacionar a morfologia galáctica com outros parâmetros físicos, como a massa e a luminosidade das galáxias.

4.3 Deep Learning

Deep Learning (DL), ou aprendizado profundo, é uma subárea da inteligência artificial (IA) e um ramo específico de Machine Learning (ML) que utiliza redes neurais artificiais (ANNs) com múltiplas camadas, conhecidas como deep neural networks. Essas redes são capazes de modelar relações complexas nos dados ao passar informações por diversas camadas hierárquicas, onde cada camada aprende representações progressivamente mais abstratas.

O DL se diferencia de métodos tradicionais de aprendizado de máquina, que geralmente dependem de engenharia manual de características, por automatizar a extração de atributos diretamente dos dados brutos, como imagens, texto e áudio. Isso é particularmente relevante para problemas como classificação de galáxias, onde as características podem ser complexas e de difícil definição manual.

Arquiteturas específicas, como redes neurais convolucionais (CNNs), são amplamente utilizadas em tarefas de processamento de imagens devido à sua capacidade de capturar padrões espaciais e hierárquicos. Além disso, a utilização de funções de ativação não lineares, como ReLU, e técnicas de otimização, como backpropagation e gradient descent, contribuem para o treinamento eficiente dessas redes.

Embora extremamente poderoso, o DL enfrenta desafios, como a necessidade de grandes volumes de dados rotulados, alto custo computacional e a dificuldade de compreender como as decisões do modelo são tomadas. Isso ocorre porque os processos internos das redes neurais profundas, distribuídos em diversas camadas e milhões de parâmetros, são complexos e não facilmente interpretáveis por seres humanos. Apesar disso, sua capacidade de aprendizado altamente escalável e eficiente torna-o a escolha ideal para tarefas complexas e de grande dimensão, como as abordadas neste trabalho.

4.3.1 Redes Neurais

As redes neurais são modelos computacionais inspirados no cérebro humano. Elas são compostas por camadas de nós, chamados neurônios, que são interconectados. Cada conexão tem

um peso que é ajustado durante o processo de treinamento. Essas redes são usadas principalmente para reconhecer padrões complexos e fazer previsões baseadas em dados.

Uma rede neural típica é organizada em uma arquitetura de camadas, que inclui uma camada de entrada, várias camadas ocultas e uma camada de saída. Cada neurônio em uma camada está conectado a neurônios na próxima camada, permitindo a passagem de informações e o processamento através da rede (LeCun; Bengio; Hinton, 2015).

- **Camada de entrada**: recebe os dados de entrada, que podem ser qualquer tipo de informação numérica, como valores de pixels em uma imagem ou características numéricas de dados tabulares. Os neurônios de entrada processam os dados, analisam ou categorizam esses dados e os encaminham para a próxima camada.

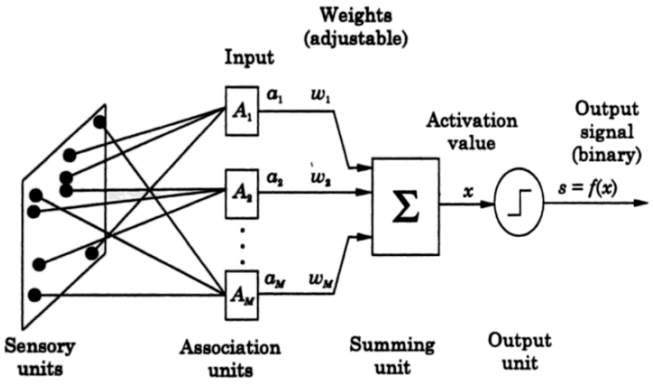
- **Camada oculta**: manipula as entradas recebidas por meio de uma sequência de operações matemáticas, extraindo características progressivamente mais abstratas e complexas. Cada neurônio em uma camada oculta recebe entradas ponderadas de todos os neurônios da camada anterior e as processa através de uma função de ativação.

  As funções de ativação aplicadas às saídas dos neurônios introduzem não-linearidade no modelo, permitindo que ele identifique padrões complexos.

- **Camada de Saída**: Gera o resultado final da rede neural, que pode ser uma única previsão (em casos de regressão) ou uma distribuição de probabilidades entre várias classes (em casos de classificação). Cada neurônio na camada de saída representa uma classe ou valor de saída possível.

A figura 3 ilustra a estrutura de um neurônio artificial.

Figura 3 – Estrutura de um neurônio artificial



Fonte: (Yegnanarayana, 2009)

4.3.2 Função de Ativação

A função de ativação transmite a informação para a próxima camada da rede. Introduz não-linearidade no modelo, permitindo que a rede aprenda e represente relações complexas entre os dados de entrada e saída. Sem a não-linearidade proporcionada pelas funções de ativação, a rede neural se comportaria como um modelo linear, independentemente da quantidade de camadas, limitando sua capacidade de resolver problemas complexos. Algumas das funções de ativação mais comuns incluem:

a) **Sigmóide**

A função sigmóide é uma função que mapeia os valores de entrada em um intervalo entre 0 e 1, tornando-a útil para modelos de classificação binária, onde a saída pode ser interpretada como uma probabilidade. Definida pela seguinte expressão:

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{4.1}$$

b) **ReLU (Rectified Linear Unit)**

ReLU é uma função linear por partes que substitui todos os valores negativos por zero e mantém os valores positivos inalterados. É amplamente utilizada em redes neurais profundas devido à sua capacidade de mitigar o problema do desvanecimento do gradiente e acelerar a convergência. Definida pela seguinte expressão:

$$ReLU(x) = \max(0,x) \tag{4.2}$$

4.4 Convolução

A convolução é uma operação matemática amplamente utilizada em Redes Neurais Convolucionais para processar dados visuais. Ela combina uma matriz de entrada, como uma imagem, com um filtro (ou kernel), gerando um mapa de características (*feature map*) que destaca elementos específicos da entrada original.

Essa operação é realizada aplicando o filtro em diferentes regiões da matriz de entrada, multiplicando seus elementos correspondentes e somando os resultados. O produto final representa as características extraídas, como bordas ou texturas.

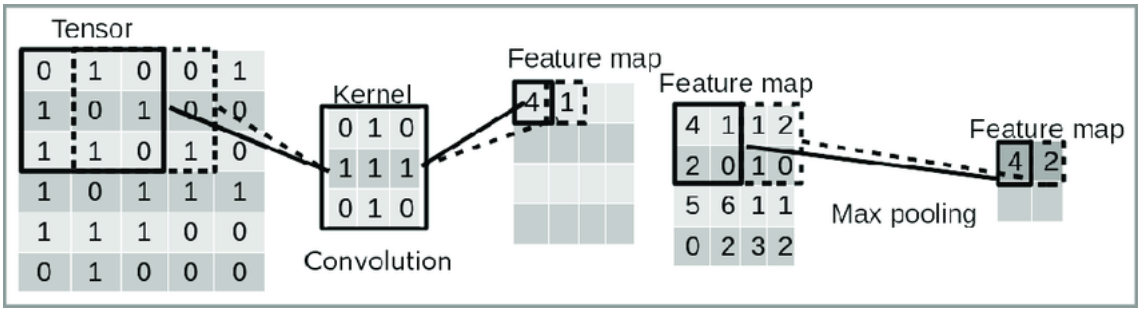
4.4.1 Componentes da Convolução

- **Filtro (Kernel)**: Matriz que define os pesos usados na extração de características.
- **Stride**: Determina o espaçamento entre as aplicações consecutivas do filtro.
- **Padding**: Adiciona elementos ao redor da matriz de entrada para preservar ou alterar suas dimensões após a convolução.

- **Feature Map**: Resultado da operação, contendo as características detectadas na entrada.

A figura 4 ilustra a aplicação de um filtro a uma matriz de entrada, resultando no mapa de características.

Figura 4 – Processo da convolução



Fonte: (Schäfer et al., 2019)

Essa abordagem permite identificar padrões locais na entrada e criar representações que são processadas em etapas posteriores da rede neural.

4.5 Redes Neurais Convolucionais

As Redes Neurais Convolucionais (CNNs) são uma classe de redes neurais artificiais projetadas para reconhecer padrões em dados visuais. Inspiradas na organização do córtex visual dos animais, as CNNs são amplamente utilizadas em tarefas como classificação de imagens, detecção de objetos e segmentação semântica.

Sua estrutura consiste em camadas convolucionais para extração de características locais, seguidas por camadas de pooling que reduzem a dimensionalidade e por camadas totalmente conectadas para a classificação final.

- **Camadas Convolucionais**: Extraem padrões como bordas e texturas.
- **Pooling**: Max pooling e average pooling são usados para reduzir a dimensionalidade, selecionando valores representativos ou médias.
- **Camadas Totalmente Conectadas**: Integram as características extraídas para realizar classificações finais.

Com vantagens como a invariância a translações e a capacidade de construir hierarquias de características, as CNNs permitem reconhecer padrões complexos em imagens. Por exemplo, essas propriedades são usadas em diagnósticos médicos, como a detecção de tumores em imagens de ressonância magnética.

4.6 Métricas

Em problemas de classificação, a avaliação do desempenho do modelo é essencial para entender sua eficácia em identificar corretamente as classes. Para isso, utilizamos várias métricas que permitem verificar não apenas a taxa de acertos gerais, mas também a qualidade das previsões feitas para cada classe em particular. Entre as principais métricas para avaliar modelos de classificação estão a matriz de confusão, acurácia, recall, precisão e F1-score, cada uma com um papel específico.

4.6.1 Matriz de Confusão

Uma matriz de confusão (ou matriz de erro) é uma técnica de visualização utilizada para avaliar o desempenho de um modelo de classificação. Trata-se de uma tabela que relaciona o número de previsões corretas e incorretas feitas pelo modelo para cada classe, dividida em categorias específicas de verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos.

Em uma matriz de confusão, cada coluna representa as previsões feitas pelo modelo, enquanto cada linha corresponde aos valores reais (a "verdade fundamental") para uma classe específica. Observe que o inverso também aparece na pesquisa. Essa estrutura de grade é uma ferramenta conveniente para visualizar a precisão da classificação do modelo, exibindo o número de previsões corretas e incorretas para todas as classes, uma ao lado da outra (Stehman, 1997).

Por exemplo, se uma galáxia do tipo SBb é corretamente identificada como SBb, ela é classificada como um verdadeiro positivo (TP). Por outro lado, uma galáxia do tipo SBc prevista incorretamente como SBb é um falso positivo (FP).

4.6.2 Acurácia

A acurácia mede a proporção de previsões corretas em relação ao total de previsões feitas. Ela fornece uma visão geral do quanto o modelo está correto em suas classificações. É dada pela fórmula:

$$Acurcia = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.3}$$

A acurácia é uma métrica confiável quando as classes estão balanceadas, oferecendo uma visão clara do desempenho do modelo, pois ele não favorece uma classe em detrimento de outra. Em um conjunto balanceado, como na classificação de galáxias SBb e SBc em proporções semelhantes, a acurácia é um bom indicador de quão bem o modelo separa as classes de maneira global.

Entretanto, em cenários desbalanceados, a acurácia pode ser enganosa. Por exemplo, se 90% das galáxias forem do tipo SBb e 10% do tipo SBc, um modelo que sempre prevê SBb teria 90% de acurácia, mas sem realmente diferenciar as classes. Nesses casos, é essencial complementar a avaliação com outras métricas.

4.6.3 Recall

O recall é uma métrica que mede a capacidade do modelo em identificar corretamente as instâncias positivas, ou seja, quantas das instâncias que realmente pertencem a uma determinada classe foram corretamente classificadas como tal. É especialmente útil em problemas onde minimizar os falsos negativos (FN) é crucial (Powers, 2020). O recall é definido como:

$$Recall = \frac{VP}{VP + FN} \tag{4.4}$$

No caso das galáxias SBb e SBc, é crucial que o modelo minimize a quantidade de galáxias que são erroneamente classificadas como de outra classe, ou seja, evitar que galáxias relevantes sejam negligenciadas no processo de classificação.

4.6.4 Precisão

Precisão (precision) é a proporção de previsões positivas que realmente pertencem à classe em questão. Em outras palavras, mede a probabilidade de uma instância escolhida aleatoriamente, que foi classificada como positiva pelo modelo, realmente pertencer à classe-alvo. A fórmula da precisão é:

$$Precision = \frac{TP}{TP + FP} \tag{4.5}$$

No contexto da classificação de galáxias, a precisão indica a capacidade do modelo de identificar corretamente galáxias, as galáxias sem confundi-las com outros tipos. Isso é importante, pois na astronomia é desejável minimizar falsos positivos para garantir que apenas galáxias com características morfológicas específicas, que correspondem às classes-alvo, sejam classificadas como SBb ou SBc.

4.6.5 F1-score

A métrica *F1-Score* é uma medida de avaliação de modelos de classificação, especialmente útil quando as classes estão desbalanceadas. Ela combina a precisão e a revocação (ou recall) em um único valor, equilibrando a importância de ambas (Taha; Hanbury, 2015). A fórmula do *F1-Score* é dada por:

$$F1 = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{4.6}$$

onde:

- Precisão (ou precisão) é a proporção de previsões corretas entre todas as previsões positivas feitas pelo modelo:

$$\text{Precisão} = \frac{TP}{TP + FP} \tag{4.7}$$

- Revocação (ou *recall*) é a proporção de casos positivos que foram corretamente identificados pelo modelo:

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{4.8}$$

Com *TP* sendo o número de verdadeiros positivos, *FP* os falsos positivos e *FN* os falsos negativos.

4.6.6 Curva ROC

A Curva ROC (*Receiver Operating Characteristic*) é uma ferramenta amplamente utilizada para avaliar o desempenho de modelos de classificação binária. Ela ilustra a relação entre a taxa de verdadeiros positivos (*True Positive Rate* - TPR) e a taxa de falsos positivos (*False Positive Rate* - FPR) em diferentes limiares de decisão (Junge; Dettori, 2018).

O eixo *x* da curva representa a taxa de falsos positivos, calculada como:

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{4.9}$$

Enquanto o eixo *y* representa a taxa de verdadeiros positivos, definida como:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.10}$$

A principal métrica associada à Curva ROC é a *Área sob a Curva* (AUC - *Area Under the Curve*), que fornece um valor escalar para quantificar o desempenho do modelo. Um AUC próximo de 1 indica um modelo com excelente capacidade discriminativa, enquanto um valor de 0,5 sugere desempenho semelhante ao acaso.

A análise da Curva ROC é particularmente útil em cenários com classes desbalanceadas ou quando os custos associados a falsos positivos e falsos negativos são diferentes. Dessa forma, a curva auxilia na escolha de limiares de decisão mais adequados às necessidades do problema em questão.

4.7 Trabalhos Relacionados

Diversos trabalhos exploram diferentes metodologias, desde a extração manual de características morfológicas até o uso de redes neurais convolucionais para identificação direta de padrões nas imagens. A seguir, são apresentados dois estudos relevantes que aplicam essas técnicas na classificação morfológica de galáxias.

4.7.1 Deep Galaxy: Classification of Galaxies based on Deep Convolutional Neural Networks.

O artigo *Deep Galaxy: Classification of Galaxies based on Deep Convolutional Neural Networks*, de Khalifa et al. (2017), aborda o desafio da classificação morfológica de galáxias, que é essencial para estudos sobre a formação e evolução do universo. A classificação visual tradicional de galáxias, embora confiável, tornou-se inviável com o crescente volume de dados astronômicos, impulsionado pela evolução dos telescópios e câmeras CCD. Assim, a pesquisa

de Khalifa et al. explora o uso de redes neurais convolucionais profundas (CNNs) como uma alternativa automatizada e eficiente.

A arquitetura implementada é composta por 8 camadas, incluindo uma camada convolucional principal com 96 filtros para extração de características, seguida por camadas totalmente conectadas para a classificação. O modelo foi treinado com 1.356 imagens do catálogo EFIGI, que reúne amostras de galáxias dos tipos Hubble, como Elípticas, Espirais e Irregulares. Os resultados foram promissores, com o modelo alcançando uma precisão de 97,27% na classificação, superando técnicas anteriores, como redes neurais feedforward e Random Forest, que atingiram no máximo 93% de acurácia.

A pesquisa de Khalifa et al. (2017) destaca a importância das CNNs para a análise de grandes volumes de dados astronômicos, demonstrando que esses modelos não só automatizam o processo, mas também aprimoram a precisão da classificação. A metodologia apresentada permite uma análise mais rápida e confiável de catálogos de galáxias, contribuindo para o avanço de estudos sobre a morfologia galáctica e para a exploração de novas hipóteses sobre a evolução do universo.

4.7.2 Machine and Deep Learning Applied to Galaxy Morphology.

O trabalho de Barchi (2020), intitulado *Machine and Deep Learning Applied to Galaxy Morphology*, explora abordagens de aprendizado de máquina e aprendizado profundo aplicadas à classificação morfológica de galáxias. Este estudo ressalta a importância da classificação precisa para a compreensão da estrutura do universo, uma vez que as características morfológicas das galáxias refletem seus processos de formação internos e interações com o ambiente.

Barchi desenvolve duas metodologias principais para a classificação morfológica. A primeira abordagem utiliza características morfológicas extraídas por meio do sistema CyMorph, que calcula métricas não-paramétricas como Concentração, Assimetria e Suavidade, além de padrões de gradiente e entropia. Essas características são então aplicadas a modelos de aprendizado de máquina supervisionado, como árvores de decisão e máquinas de vetores de suporte. A segunda abordagem é baseada em redes neurais convolucionais profundas (CNNs), que realizam a extração automática de características a partir das imagens das galáxias, dispensando a necessidade de extração manual de atributos.

As redes neurais convolucionais (CNNs) empregadas por Barchi conseguem extrair automaticamente características relevantes diretamente das imagens, identificando padrões estruturais que ajudam a distinguir entre galáxias elípticas, espirais e suas subcategorias, incluindo espirais barradas. Estas redes foram treinadas com dados rotulados do projeto Galaxy Zoo, que fornece rótulos visuais detalhados e confiáveis para cada galáxia. Utilizando milhares de imagens, a arquitetura convolucional foi ajustada para maximizar a precisão na classificação dos subtipos morfológicos, alcançando acurácia superior a 94,5% na classificação entre galáxias elípticas e espirais e até 99% ao considerar apenas duas classes (elípticas e espirais).

Um dos aspectos mais significativos do trabalho foi a implementação da arquitetura

GoogleNet Inception, uma rede convolucional profunda de 22 camadas, que permite a detecção de padrões complexos em múltiplas escalas. Através do módulo Inception, que realiza convoluções em paralelo com diferentes tamanhos de filtro (1x1 a 5x5), a rede consegue capturar tanto detalhes estruturais finos quanto padrões mais amplos, essenciais para a classificação de galáxias do tipo SBb e SBc, nas quais variam a presença e a intensidade da barra e dos braços espirais.

Além disso, Barchi valida o desempenho dos modelos utilizando dados espectroscópicos, o que permite garantir a precisão das classificações morfológicas. Esse rigor metodológico reforça a confiabilidade dos resultados e demonstra a robustez da abordagem aplicada, especialmente em grandes conjuntos de dados astronômicos. O trabalho também aborda o problema do desbalanceamento de classes, discutindo como diferentes métricas e ajustes de amostras ajudam a manter a qualidade das classificações.

Com essa combinação de abordagens tradicionais e avançadas, Barchi destaca que, embora as CNNs obtenham uma alta precisão e automatizem a classificação, métodos tradicionais ainda oferecem insights valiosos para a compreensão de características específicas de subtipos de galáxias. Essa abordagem híbrida reforça o potencial das CNNs para classificar com eficácia galáxias do tipo Via Láctea, como SBb e SBc, ao mesmo tempo em que aponta para a importância de métodos que preservam as características fundamentais das galáxias, permitindo análises morfológicas detalhadas.

5 METODOLOGIA

No presente capítulo, abordaremos os processos e métodos para a redução dos dados e criação do modelo. Para a classificação de galáxias dos tipos SBb e SBc utilizando redes neurais convolucionais, adotamos uma abordagem sistemática baseada em dados observacionais provenientes do catálogo Galaxy Zoo 2 (Willett et al., 2013). Inicialmente, filtramos as galáxias mais próximas por meio da extração dos valores de redshift. Portanto, priorizamos galáxias com valores mais baixos de redshift. Em seguida, as galáxias selecionadas foram rankeadas de acordo com o valor de redshift z. Utilizando a biblioteca Python Astroquery, realizamos a busca dos objetos e realizamos o download dos dados correspondentes, acessando as coordenadas RA e DEC e extraindo as imagens em formato FITS.

5.1 Ferramentas e Etapas

Para o desenvolvimento do projeto, utilizamos as seguintes ferramentas:

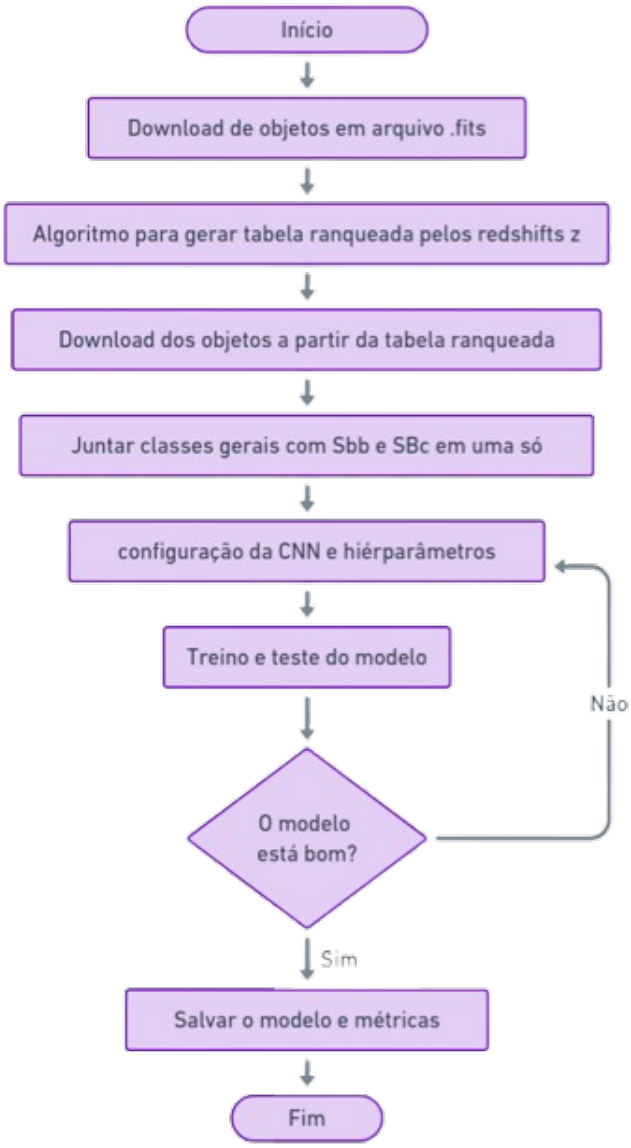
- **Python**: Linguagem de programação utilizada para todo o desenvolvimento.
- **Astroquery**: Biblioteca Python usada para acessar bancos de dados astronômicos.
- **Astropy**: Biblioteca Python para manipulação de dados astronômicos, como arquivos FITS.
- **TensorFlow**: Biblioteca usada para o desenvolvimento das Redes Neurais Convolucionais.

O desenvolvimento do modelo proposto foi conduzido de forma estruturada e sistemática, seguindo um conjunto de etapas bem definidas. Inicialmente, procedeu-se à aquisição dos dados por meio do download dos arquivos no formato *FITS* (Flexible Image Transport System), utilizando as coordenadas celestes fornecidas no catálogo em termos de Ascensão Reta (RA) e Declinação (DEC), garantindo a obtenção precisa das imagens de interesse no catálogo. Após a coleta dos arquivos, foi realizada uma etapa de pré-processamento que envolveu o ranqueamento e a validação dos arquivos, assegurando os objetos mais próximos para melhor qualidade.

Na sequência, efetuou-se o *download* das imagens correspondentes, seguido por um refinamento na estrutura das classes. Nesse processo, as imagens das galáxias pertencentes às classes elípticas (E), espirais do tipo Sa, Sb, Sc e espirais barradas do tipo SBa foram unificadas em uma única classe. Adicionalmente, as galáxias espirais barradas dos tipos SBa e SBb foram agrupadas em outra classe.

Com os dados devidamente organizados e as classes redefinidas, foi elaborada a implementação do algoritmo responsável pelo desenvolvimento do modelo de aprendizado profundo. Nessa etapa foi implementado o algoritmo de redes neurais convolucionais (*CNNs*). Esse procedimento permitiu a construção de um modelo para a classificação morfológica de galáxias. Na figura 5, apresenta-se um fluxograma que descreve, de forma breve, o passo a passo do trabalho realizado:

Figura 5 – Modelo de matriz



Fonte: Elaborado pelo autor (2024)

5.2 Segmentação do Catálogo Galaxy Zoo 2 por Tipos Morfológicos

Na segmentação das imagens, foi desenvolvido um algoritmo em Python com o objetivo de segmentar o catálogo Galaxy Zoo 2 em diferentes tipos morfológicos. Utilizando classificações morfológicas detalhadas, o algoritmo categorizou as galáxias nos seguintes tipos: Elípticas (E), Espirais do tipo Sa, Sb e Sc, e Espirais Barradas do tipo SBa, SBb e SBc. Este processo de categorização resultou na criação de sete novos catálogos, cada um representando um tipo morfológico específico. Esses catálogos foram fundamentais para a organização e análise subsequente dos dados, permitindo um tratamento diferenciado e adequado conforme as características morfológicas de cada grupo de galáxias. Este refinamento inicial dos dados garante a precisão

e a relevância dos conjuntos de dados utilizados nos modelos de classificação, otimizando a eficiência e a eficácia das análises realizadas. Foram extraídas as seguintes colunas do catálogo Galaxy Zoo 2:

- **specobjid**: No catálogo Galaxy Zoo 2, "specobjid"é um identificador único que refere-se a uma entrada específica no banco de dados de espectroscopia do Sloan Digital Sky Survey (SDSS).

- **RA**: Ascensão Reta (Right Ascension) é a coordenada equivalente à longitude na superfície da Terra, mas aplicada à esfera celeste.

- **DEC**: Declinação (Declination) é a coordenada equivalente à latitude na superfície da Terra, aplicada à esfera celeste.

- **gz2class**: o campo "gz2class"refere-se à classificação morfológica das galáxias realizada pelos voluntários participantes do projeto. Esse campo contém as informações sobre as características visuais das galáxias

A tabela 1 contém uma amostra do catálogo para as galáxias do tipo SBb:

Tabela 1 – Tabela reduzida contendo apenas galáxias do tipo SBb e suas variações

specobjid	ra	dec	gz2class
1.8026749296451523e+18	160.9904	11.70379	SBb?t
3.7950636743566336e+17	192.05688	-3.3328447	SBb2l(o)
2.5130803994128814e+18	184.40326	29.60802	SBb(r)
6.891150992977695e+17	227.44485	57.0002	SBb?t
1.9906883998240215e+18	186.76057	15.46148	SBb
2.813697604540459e+18	171.62413	16.863422	SBb+t

Fonte: Elaborado pelo autor (2024)

Após a execução do algoritmo de pré-processamento dos dados, o catálogo Galaxy Zoo 2 foi segmentado em sete novos catálogos baseados nos tipos morfológicos das galáxias. A distribuição dos objetos em cada classe é a seguinte: Elípticas (E) com 96.514 objetos, Espirais do tipo Sa com 1.706 objetos, Espirais do tipo Sb com 37.818 objetos, Espirais do tipo Sc com 49.106 objetos, Espirais Barradas do tipo SBa com 307 objetos, Espirais Barradas do tipo SBb com 11.024 objetos, e Espirais Barradas do tipo SBc com 9.208 objetos. Esta segmentação detalhada e a subsequente categorização dos dados fornecem a base necessária para separar as classes de interesse das outras que serão generalizadas, ou seja, as classes que não serão investigadas serão tratadas como uma classe só.

5.3 Ranking de objetos por Redshift

A criação do ranqueamento dos objetos por redshift se justifica pela necessidade de aumentar a confiabilidade na classificação morfológica dos objetos no catálogo Galaxy Zoo 2. Galáxias com redshift elevado, por estarem mais distantes, fornecem imagens com menor quantidade de detalhes, o que dificulta a correta identificação de suas características morfológicas. Consequentemente, essas galáxias podem ter sido classificadas de forma equivocada. Ao priorizar as galáxias mais próximas, com redshift menor, garantimos que as CNNs sejam treinadas com dados de alta qualidade e mais detalhados, aumentando a acurácia do modelo e minimizando erros de classificação.

Para otimizar a performance das redes neurais convolucionais (CNNs) na classificação de galáxias, implementamos um algoritmo de ranking baseado em redshift, ordenando os objetos de forma crescente, do redshift z menor para o maior. Este procedimento é fundamental, pois permite que as CNNs observem primeiramente as galáxias mais próximas, cuja riqueza de detalhes morfológicos pode ser mais facilmente capturada, resultando em uma maior taxa de acurácia na classificação.

O algoritmo desenvolvido foi escrito em Python, utilizando bibliotecas especializadas para manipulação e análise de dados astronômicos como Astropy para manipulação dos arquivos FITS. Inicialmente, extraímos os valores de redshift de cada objeto catalogado no Galaxy Zoo 2. Em seguida, os objetos foram ordenados em ordem crescente de redshift. Este ranking facilita a seleção das galáxias mais próximas para serem utilizadas no treinamento das CNNs, permitindo que o modelo aprenda padrões morfológicos com mais precisão. Abaixo, a Figura 6 ilustra através de um pseudo-código como o algoritmo está organizado.

Figura 6 – Classificação de Arquivos FITS de Acordo com os Valores de Redshift

```
1:  function RANQUEAR_REDSHIFT(caminho_arquivos)
2:      Inicializa um dicionário vazio chamado ranking_dict
3:      Lista todos os arquivos no diretório especificado por caminho_arquivos
4:      for cada filename na lista de arquivos do
5:          Constrói o caminho completo do arquivo chamado file_path
6:          Tente:
7:              Abre o arquivo FITS em file_path
8:              Extrai o valor de redshift da estrutura de dados no arquivo FITS
9:              Adiciona o par filename-redshift no dicionário ranking_dict
10:         Captura erro (IndexError, KeyError, TypeError, OSError):
11:             Imprime uma mensagem de erro com o nome do arquivo e o erro ocorrido
12:     end for
13:     Ordena os nomes dos arquivos em ranking_dict pelo valor de redshift
14:     Retorna a lista de nomes de arquivos ordenados e o dicionário ranking_dict
15: end function
```

Fonte: Elaborado pelo autor (2024)

A aplicação do algoritmo de ranking por redshift resultou em uma distribuição ordenada dos objetos no catálogo, pronta para ser utilizada no treinamento das CNNs. Este método foi

fundamental para garantir que as redes neurais recebessem informações mais detalhadas e precisas sobre de galáxias mais próximas, permitindo um aprendizado mais preciso.

Em resumo, o ranking de objetos por redshift se mostrou uma etapa crucial no pré-processamento dos dados. Este processo não apenas melhorou a acurácia das CNNs, mas também garantiu que as galáxias mais próximas, com detalhes mais pronunciados, fossem priorizadas no treinamento do modelo, resultando em um avanço significativo na análise e categorização das galáxias do tipo SBb e SBc. A tabela 2 mostra as primeiras linhas da tabela gerada, para os objetos da classe SBb, mostrando uma ordenação dos objetos.

Tabela 2 – Objetos ranqueados por redshift

Arquivo	**Redshift**
spec-1015-52709-0523.fits	-0.0038485848
spec-1314-52792-0625.fits	-0.0033263185
spec-0449-51900-0497.fits	-0.00012443581
spec-0283-51584-0002.fits	0.0
spec-0526-52312-0441.fits	0.000033842738
spec-0518-52282-0177.fits	0.00018633419
spec-0302-51616-0224.fits	0.0024251305
spec-0967-52636-0578.fits	0.0024843556

Fonte: Elaborado pelo autor (2024)

As tabelas com os objetos ranqueados foram geradas para cada uma das classes. Desse modo, podemos usar para a criação e treinamento do modelo apenas os objetos próximos, ou seja, apenas objetos com alta confiança.

5.4 Arquitetura implementada

Após testes de diferentes arquiteturas testadas e configuração de hiperparâmetros, o modelo descrito foi identificado como a configuração mais eficiente para o treinamento da rede neural convolucional (CNN). Essa arquitetura, apresentada na tabela 3, foi selecionada com base em experimentos conduzidos para otimizar a capacidade de generalização e reduzir a ocorrência de overfitting aos dados de treinamento.

A estrutura do modelo proposto foi projetada para equilibrar profundidade e complexidade computacional, utilizando dois blocos de camadas convolucionais, cada um seguido de normalização em lotes (*Batch Normalization*), função de ativação ReLU, operações de agrupamento máximo (*MaxPooling*) e camadas de *Dropout* com taxa de 20%. Após as camadas convolucionais, o modelo realiza a achatamento (*Flatten*) dos mapas de características e utiliza duas camadas densas totalmente conectadas com 64 neurônios cada, ambas com função de ativação ReLU e camadas de *Dropout* com taxa de 50% para mitigar o risco de *overfitting*. O uso do regularizador *L2* nas camadas convolucionais promove uma penalização adicional sobre os pesos do modelo.

Tabela 3 – Arquitetura da Rede Neural Convolucional (CNN)

Layer	Type	Parameters
1	Conv2D	Filters: 32, Kernel: (3, 3)
2	BatchNormalization	-
3	ReLU	-
4	MaxPooling2D	Pool size: (2, 2)
5	Conv2D	Filters: 64, Kernel: (3, 3)
6	BatchNormalization	-
7	ReLU	-
8	MaxPooling2D	Pool size: (2, 2)
9	Conv2D	Filters: 128, Kernel: (3, 3)
10	BatchNormalization	-
11	ReLU	-
12	MaxPooling2D	Pool size: (2, 2)
13	Flatten	-
14	Dense	Units: 128
15	BatchNormalization	-
16	ReLU	-
17	Dropout	Rate: 0.5
18	Dense	Units: 128
19	BatchNormalization	-
20	ReLU	-
21	Dropout	Rate: 0.5
22	Dense	Units: 128
23	BatchNormalization	-
24	ReLU	-
25	Dropout	Rate: 0.5
26	Dense	Units: 1, Activation: Sigmoid

Fonte: Elaborado pelo autor (2024)

Por fim, a camada de saída utiliza a função de ativação sigmoide para realizar a classificação binária entre galáxias gerais e galáxias barradas.

Este modelo foi implementado e usado para todo o processo de treinamento e validação, sendo considerado o mais adequado para a tarefa proposta neste trabalho.

5.4.1 Treinamento do modelo

O treinamento do modelo foi conduzido com a utilização de várias técnicas e parâmetros para garantir uma otimização eficaz e evitar o sobreajuste. A seguir, são descritos os principais aspectos envolvidos no processo de treinamento.

Primeiramente, o tamanho do lote (*batch size*) foi configurado para 32. Este valor determina quantas amostras serão processadas por vez durante o treinamento. Um *batch size*

moderado permite um bom equilíbrio entre o uso de memória e a eficiência computacional, possibilitando a atualização dos pesos do modelo de forma estável.

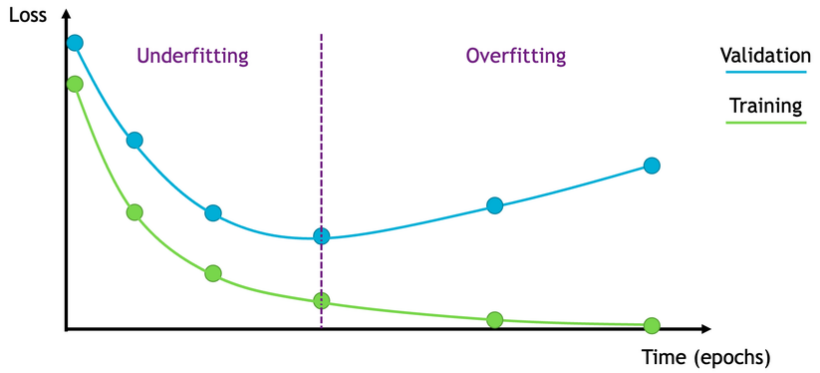
Os dados foram organizados em três conjuntos: treinamento, validação e teste. Utilizou-se a técnica de validação cruzada com *k-folds*, onde o conjunto de dados é dividido em k subconjuntos, e o treinamento é repetido k vezes, com cada subconjunto utilizado como conjunto de validação uma vez. Essa abordagem garante que o modelo seja avaliado de maneira imparcial, utilizando diferentes subconjuntos dos dados para treinamento e validação.

O número máximo de épocas foi definido como 1000. Isso significa que o modelo poderia passar por até 1000 iterações completas sobre o conjunto de dados de treinamento. No entanto, para evitar o sobreajuste e garantir uma melhor generalização, foi implementado o mecanismo de *Early Stopping*.

O *Early Stopping* monitora a acurácia de validação durante o treinamento. Caso o desempenho no conjunto de validação não melhore após um número pré-determinado de épocas, o treinamento é interrompido para evitar que o modelo continue aprendendo ruídos dos dados de treinamento, o que poderia resultar em sobreajuste. Neste caso, a paciência foi configurada para 10 épocas, ou seja, o treinamento será interrompido se não houver melhoria na acurácia de validação por 10 épocas consecutivas.

A Figura 7 ilustra o comportamento do *Early Stopping*, mostrando como o treinamento é interrompido no momento em que a acurácia de validação deixa de melhorar, evitando o ajuste excessivo ao conjunto de treinamento.

Figura 7 – Exemplo de *Early Stopping* indicando onde o treinamento deve parar para evitar sobreajuste.



Fonte: (Vakalopoulou et al., 2023)

Além disso, foi utilizado um agendador de taxa de aprendizado (*learning rate scheduler*) para ajustar dinamicamente a taxa de aprendizado ao longo das épocas. A taxa de aprendizado inicial foi configurada como 0.001, e a cada 10 épocas, ela é reduzida em um fator de 0.1. Esse ajuste gradual da taxa de aprendizado permite que o modelo faça ajustes mais finos nos estágios finais do treinamento, favorecendo a convergência sem grandes oscilações nos pesos.

A geração de dados foi realizada com a utilização do *ImageDataGenerator* do Keras, que aplicou técnicas de aumento de dados, como rotação, deslocamento e inversão horizontal, para

aumentar a diversidade das imagens de entrada e, assim, melhorar a capacidade de generalização do modelo. Essas transformações são realizadas em tempo real durante o treinamento, garantindo que o modelo seja exposto a uma variedade maior de exemplos.

O modelo foi construído utilizando uma arquitetura de rede neural convolucional (CNN) com duas camadas convolucionais seguidas de camadas de normalização em lote, max-pooling e *dropout* para reduzir o risco de sobreajuste. A camada de saída foi projetada com uma única unidade ativada por uma função sigmoide, adequada para problemas de classificação binária.

O treinamento foi realizado com o uso de um otimizador *Adam*, que combina as vantagens de outros métodos de otimização, como o *Momentum* e *RMSprop*, ajustando a taxa de aprendizado de maneira adaptativa para cada parâmetro. A função de perda utilizada foi a *binary crossentropy*, que é comum em problemas de classificação binária.

Com esses parâmetros e técnicas, o modelo foi treinado, utilizando validação cruzada e *Early Stopping* para evitar o sobreajuste e garantir uma boa capacidade de generalização para dados não vistos.

5.4.2 Divisão dos Dados e Validação Cruzada

Os dados foram divididos e avaliados por meio da técnica de validação cruzada *K-Fold*, com 5 dobras ($K = 5$). Essa abordagem foi escolhida para garantir uma avaliação imparcial da performance da rede neural convolucional (CNN), pois cada amostra do conjunto de dados é utilizada tanto para treinamento quanto para validação ao longo das iterações.

Na técnica K-Fold, os dados foram divididos em 5 subconjuntos aproximadamente iguais (*folds*). Em cada iteração, 4 *folds* (80%) foram utilizados para o treinamento e 1 *fold* (20%) foi usado para validação. Esse processo foi repetido 5 vezes, permitindo que cada *fold* servisse como conjunto de validação uma vez e como treinamento nas outras quatro iterações. Dessa forma, o modelo foi avaliado em todos os dados disponíveis.

A Tabela 4 ilustra a distribuição total dos dados entre as classes, considerando que a técnica K-Fold utiliza a totalidade dos dados de forma alternada para treinamento e validação.

Tabela 4 – Distribuição total dos dados entre as classes para a validação cruzada K-Fold.

Conjunto	**Classe 1**	**Classe 2**	**Total**
Dados totais	9.957	9.957	19.914

Fonte: Elaborado pelo autor (2024)

A utilização do *K-Fold Cross-Validation* permitiu avaliar o modelo de maneira mais precisa e generalizável, pois mitigou o risco de viés na divisão dos dados e garantiu que todas as amostras contribuíssem para o treinamento e validação. Além disso, essa abordagem possibilitou a obtenção de métricas médias (como acurácia, precisão, F1-score e AUC) e seus respectivos desvios padrão, fornecendo uma visão da performance geral do modelo.

6 RESULTADOS E DISCUSSÕES

Neste capítulo, são apresentados em detalhes os resultados obtidos a partir da aplicação do método proposto, que utilizou arquiteturas de Redes Neurais Convolucionais (*CNN*s) para a análise das galáxias. Além disso, são discutidas as características específicas das arquiteturas empregadas e como elas contribuíram para o desempenho do modelo. Esses resultados oferecem uma visão abrangente do impacto e da eficácia das *CNN*s na classificação e estudo de galáxias gerais e barradas.

6.1 Resultados da Validação

A avaliação do modelo foi realizada utilizando *k-fold cross-validation* com $k = 5$. A Tabela 5 apresenta os principais resultados obtidos na validação do modelo durante a execução da última iteração (*Fold 5*).

Tabela 5 – Resultados da Validação no Fold 5.

Métrica	**Classe General**	**Classe SBB_SBC**	**Média Geral**
Precision	0.80	0.80	0.80
Recall	0.80	0.79	0.80
F1-Score	0.80	0.79	0.80
Acurácia	0.80		
Validation Loss	0.597		

Fonte: Elaborado pelo autor (2024)

Os resultados mostram uma acurácia média de 80% durante a validação, com valores similares para as métricas *precision*, *recall* e *F1-score*, tanto para a classe **general** quanto para a classe **SBB_SBC**. Observa-se que o modelo conseguiu equilibrar o desempenho entre ambas as classes, sem apresentar grandes variações entre elas.

O valor da *Validation Loss* foi de aproximadamente 0.597, o que indica que a rede neural convolucional apresentou uma convergência estável durante o treinamento.

Esses resultados reforçam a capacidade do modelo de generalizar bem para novos dados e confirmam a eficácia da arquitetura proposta para a tarefa de classificação das imagens.

Os resultados obtidos a partir do processo de *k-fold cross-validation* para a classificação das imagens de galáxias indicam uma performance consistente do modelo em cada uma das divisões dos dados. A Tabela 6 apresenta as métricas de perda, acurácia e F1-Score de validação alcançadas em cada um dos cinco *folds*.

Tabela 6 – Resultados das métricas de validação para as *5 folds* do processo de K-Fold Cross Validation.

Fold	Validation Loss	Validation Accuracy	Validation F1-Score
Fold 1	0.5978	0.7826	0.7818
Fold 2	0.5769	0.7901	0.7887
Fold 3	0.5656	0.7994	0.7985
Fold 4	0.5891	0.7765	0.7764
Fold 5	0.5971	0.7972	0.7972

Fonte: Elaborado pelo autor (2024)

A acurácia alcançada em cada *fold* variou de 77.65% a 79.94%, demonstrando uma boa estabilidade e confiabilidade do modelo. A média da acurácia em todos os *folds* foi de aproximadamente 78.72%, o que indica que o modelo possui um bom desempenho geral na tarefa de classificação de galáxias.

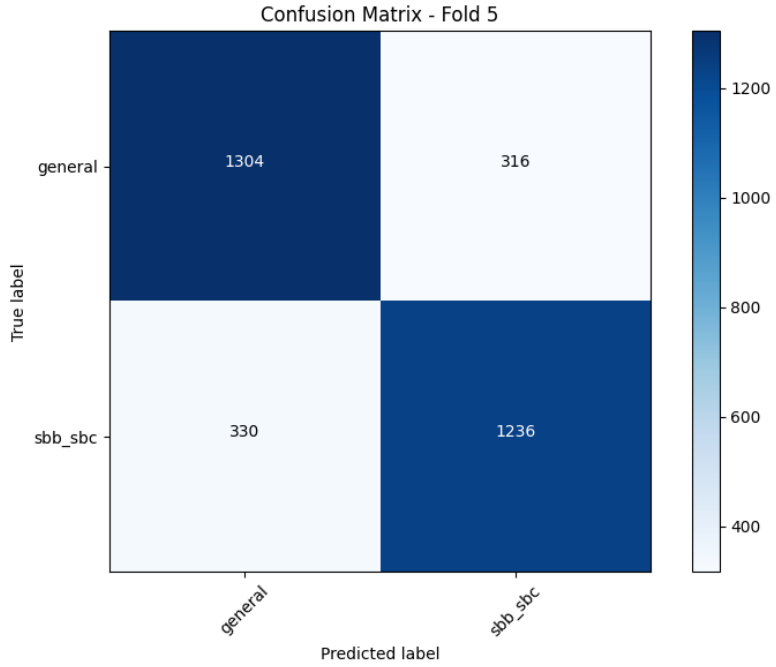
Os valores de acurácia mostram que o modelo é capaz de distinguir bem entre as classes, com resultados muito próximos em cada iteração, o que sugere que ele generaliza bem em dados não vistos durante o treinamento. A leve variação observada entre os *folds* pode ser atribuída às diferenças nas divisões dos dados, mas a performance consistente ao longo dos *k-folds* reforça a robustez do modelo.

Esses resultados são indicativos de que a abordagem de rede neural convolucional (*CNN*) utilizada é eficaz para a classificação morfológica das galáxias, e que o uso de técnicas de validação cruzada contribui para uma avaliação sólida da capacidade do modelo em diferentes cenários.

6.2 Análise da Matriz de confusão

Um dos métodos mais utilizados para avaliar o desempenho de modelos de aprendizado de máquina em tarefas de classificação é a matriz de confusão. Neste trabalho, a matriz de confusão apresentada foi gerada para ilustrar a performance do modelo na classificação de galáxias gerais e barradas. A análise detalhada dessa matriz nos permite identificar não apenas a taxa de acertos, mas também os padrões de erros cometidos pelo modelo. A Figura 8, apresentada abaixo, ilustra a matriz de confusão obtida durante o treinamento utilizando validação cruzada com 5 dobras (k-fold 5).

Figura 8 – Matriz de confusão



Fonte: Elaborado pelo autor (2024)

A matriz é composta por quatro quadrantes que representam as previsões do modelo em relação aos rótulos reais:

- Verdadeiros Positivos (VP): O número de galáxias barradas (SBb e SBc) corretamente classificadas pelo modelo. Nesse caso, o valor é 1236, indicando que uma grande parte das galáxias barradas foi corretamente identificada.
- Falsos Negativos (FN): Galáxias barradas (SBb e SBc) que foram incorretamente classificadas como galáxias gerais. O valor neste quadrante é 330, sugerindo que o modelo teve alguma dificuldade em identificar parte das galáxias barradas.
- Falsos Positivos (FP): Galáxias gerais que foram erroneamente classificadas como galáxias barradas (SBb e SBc). O valor neste quadrante é 316, indicando uma taxa de confusão relativamente baixa para esta categoria.
- Verdadeiros Negativos (VN): Galáxias gerais corretamente classificadas pelo modelo. O valor neste quadrante é 1304, demonstrando que a *CNN* conseguiu identificar corretamente uma quantidade significativa de galáxias gerais.

A análise da matriz de confusão revela que o modelo alcançou um desempenho consistente na classificação das galáxias barradas (SBb e SBc) e das galáxias gerais. Observa-se que o modelo classificou corretamente 1236 galáxias barradas (Verdadeiros Positivos) e 1304 galáxias gerais (Verdadeiros Negativos).

Por outro lado, ocorreram 316 Falsos Positivos, em que galáxias gerais foram classificadas como barradas, e 330 Falsos Negativos, em que galáxias barradas foram identificadas como gerais. Esse comportamento sugere que, embora a arquitetura da *CNN* tenha sido eficaz em capturar as características predominantes das galáxias barradas e gerais, ainda há desafios na distinção entre as classes.

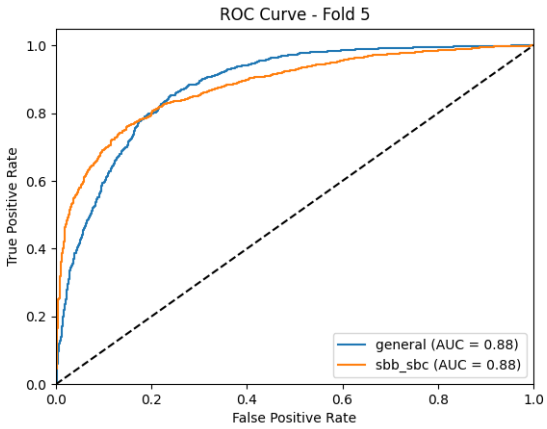
Uma possível explicação para os erros observados é a existência de características compartilhadas entre as galáxias gerais e barradas, como formas estruturais similares ou variações intrínsecas dentro das próprias classes. Além disso, as imagens utilizadas, que estão no espectro visível, podem limitar a detecção de características mais relevantes. Imagens em outros comprimentos de onda, como no ultravioleta ou infravermelho, poderiam revelar detalhes adicionais que ajudariam a melhorar o desempenho do modelo.

6.2.1 Análise da Curva ROC

A área sob a curva ROC (AUC - Area Under the Curve) é frequentemente utilizada como um indicador quantitativo do desempenho do modelo. Um valor de *AUC* igual a 1 representa um classificador perfeito, enquanto um valor de AUC igual a 0,5 indica um modelo sem capacidade discriminativa, equivalente a uma classificação aleatória.

Na Figura 9, observa-se a curva *ROC* para as duas classes do problema: "general"e "sbb_sbc", utilizando validação cruzada com $k\text{-}fold = 5$. Ambas as classes apresentam uma AUC de 0,88, o que demonstra que o modelo possui uma boa capacidade discriminativa para separar as galáxias gerais das galáxias barradas. A proximidade das duas curvas e seus valores semelhantes de AUC indicam que o desempenho do modelo é balanceado entre as duas classes, sem favorecer excessivamente uma em detrimento da outra.

Figura 9 – Curva ROC



Fonte: Elaborado pelo autor (2024)

Além disso, a forma da curva ROC revela que o modelo apresenta alta sensibilidade em baixos valores de FPR, com um rápido crescimento da taxa de verdadeiros positivos à medida que a taxa de falsos positivos aumenta. Isso sugere que o modelo consegue realizar boas predições em

condições onde é importante minimizar falsos positivos, o que é crucial em tarefas que exigem alta precisão.

No entanto, é importante observar que, embora o AUC de 0,88 seja um resultado positivo, ainda existem áreas de melhoria, especialmente na redução de erros de classificação entre as duas classes. Tais melhorias podem ser alcançadas com ajustes finos no modelo, como o uso de técnicas de regularização ou o refinamento dos dados de entrada.

Em resumo, a análise da curva ROC e do valor de AUC confirma que a arquitetura do modelo é robusta e capaz de aprender características relevantes que diferenciam as classes "general"e "sbb_sbc", resultando em um desempenho satisfatório para o problema de classificação.

7 CONCLUSÃO

7.1 Conclusão e Trabalhos Futuros

Este trabalho apresentou uma abordagem baseada em redes neurais convolucionais para classificar galáxias em gerais e barradas do tipo *SBb* e *SBc*. Apesar disso, limitações relacionadas às bandas espectrais utilizadas e à precisão do modelo apontam caminhos para melhorias.

Para perspectiva de trabalhos futuros, levantamos as seguintes temáticas e experimentos: o uso de imagens em outras bandas do espectro, como infravermelho e ultravioleta, que podem revelar detalhes adicionais, e o uso de algoritmos alternativos, como *Random Forest*, um método de aprendizado baseado em árvores de decisão. Além disso, explorar diferentes pré-processamentos e ajustes na arquitetura da rede neural convolucional pode contribuir para melhorar a performance do modelo, possibilitando uma extração mais eficiente das características relevantes das imagens.

REFERÊNCIAS

BARCHI, P. H. **Machine and deep learning applied to galaxy morphology**. 2020. 25

BINNEY, J.; TREMAINE, S. **Galactic dynamics**. [S.l.]: Princeton university press, 2011. v. 13. 12

HUBBLE, E. Extragalactic nebulae. **The Astrophysical Journal**, University of Chicago Press, v. 64, p. 321, 1926. 11

JUNGE, M. R.; DETTORI, J. R. Roc solid: Receiver operator characteristic (roc) curves as a foundation for better diagnostic tests. **Global Spine Journal**, SAGE Publications Sage CA: Los Angeles, CA, v. 8, n. 4, p. 424–429, 2018. 24

KENNICUTT, R. C. **STAR FORMATION IN GALAXIES ALONG THE HUBBLE SEQUENCE**. 1998. 12

KHALIFA, N. E. M. et al. **Deep Galaxy: Classification of Galaxies based on Deep Convolutional Neural Networks**. 2017. Disponível em: <https://arxiv.org/abs/1709.02245>. Acessado em 25 de abril de 2024. 24, 25

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015. 19

POWERS, D. M. **Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation**. 2020. Disponível em: <https://arxiv.org/abs/2010.16061>. Acessado em 14 de maio de 2024. 23

SCHäFER, J. et al. Measuring particle size distributions in multiphase flows using a convolutional neural network. **Chemie Ingenieur Technik**, v. 91, n. 11, p. 1688–1695, 2019. 21

STEHMAN, S. V. Selecting and interpreting measures of thematic classification accuracy. **Remote sensing of Environment**, Elsevier, v. 62, n. 1, p. 77–89, 1997. 22

TAHA, A. A.; HANBURY, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. **BMC medical imaging**, Springer, v. 15, p. 1–28, 2015. 23

VAKALOPOULOU, M. et al. **Deep learning: basics and convolutional neural networks (CNN)**. [S.l.: s.n.], 2023. 33

WILLETT, K. W. et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. **Monthly Notices of the Royal Astronomical Society**, Oxford University Press, v. 435, n. 4, p. 2835–2860, 2013. 17, 27

XU, Y. et al. What does the milky way look like? **The Astrophysical Journal**, IOP Publishing, v. 947, n. 2, p. 54, 2023. 12

YEGNANARAYANA, B. **Artificial neural networks**. [S.l.]: PHI Learning Pvt. Ltd., 2009. 19