UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

FELIPE ARCHANJO DA CUNHA MENDES

CLASIFICAÇÃO DE ESPÉCIES DE AVES POR MEIO DA EXTRAÇÃO DE CARACTERÍSTICAS DE SEUS MEL-ESPECTROGRAMAS

- Erros na pg. 2 (resumo)
- Sugestão na pg. 5 (introdução)
- Sugestão na pg. 7 (referencial teórico)
- Correções na pg. 10 (referencial teórico)
- Sugestão na pg. 12 (materiais)
- Questão na pg. 13 (metodologia)
- Questão na pg. 14 (metodologia – extração de características)
- Formatação na pg. 15 (treinamento)
- Formatação de quadros nas pgs. 16, 17 e 18
- Erros nas referências

Na apresentação:

- Questões sobre o dataset (pg. 12)

CAMPO MOURÃO

2023

RESUMO

A classificação de espécies de pássaros é uma tarefa importante no monitoramento e conservação do meio ambiente. Neste estudo, o objetivo é desenvolver um modelo de aprendizado de máquina destinado a classificar distintas espécies de aves com base em seus cantos e chamados, empregando diversas técnicas de extração de características. Diversas técnicas de extração serão testadas, abrangendo a utilização de descritores *handcrafted*, como LBP e filtros Gabor, bem como descritores baseados em *transfer learning*, a exemplo de VGG16 e ResNet50, visando a identificação do método de extração mais eficaz. Além disso, serão realizados testes com o KNN e SVM com o intuito de avaliar a precisão e eficácia do modelo proposto. Espera-se que o modelo de aprendizado de máquina desenvolvido seja aplicável em sistemas de monitoramento ambiental e projetos de conservação, proporcionando uma base sólida para o desenvolvimento de ferramentas de identificação automática de aves.

Palavras-chave: aprendizagem de muina; local binary patterns; filtros gabor; vgg16; resnet50.

ABSTRACT

The classification of bird species is an important task in environmental monitoring and conservation. In this study, the goal is to develop a machine learning model aimed at classifying different bird species based on their songs and calls, employing various feature extraction techniques. Several feature extraction techniques will be tested, including the use of handcrafted descriptors such as LBP and Gabor filters, as well as transfer learning-based descriptors like VGG16 and ResNet50, to identify the most effective extraction method. Additionally, tests will be conducted with KNN and SVM to evaluate the accuracy and effectiveness of the proposed model. It is expected that the developed machine learning model will be applicable in environmental monitoring systems and conservation projects, providing a solid foundation for the development of automatic bird identification tools.

Keywords: machine learning; local binary patterns; filtros gabor; vgg16; resnet50.

SUMÁRIO

1 INTRODUÇÃO

A diversidade de espécies desempenha um papel essencial na manutenção da estabilidade dos ecossistemas em nosso planeta, sendo um dos pilares fundamentais para a sustentabilidade e equilíbrio. Esta variedade de formas de vida é fundamental para garantir serviços ecossistêmicos vitais, tais como a polinização (CARPENTER, 1978), o controle de pragas (HOLMES, 1990) e a regulação do ciclo de nutrientes (SNOW, 1971). Dentro desse amplo espectro de biodiversidade, as aves têm um papel de destaque como indicadores da saúde dos ecossistemas, atraindo uma considerável quantidade de pesquisas científicas.

A correta identificação das espécies de aves representa um desafio complexo, que exige conhecimento especializado e experiência por parte dos ornitólogos. Tradicionalmente, a identificação das aves se baseia em características morfológicas, tais como tamanho, formato do bico, plumagem e vocalização ou sonoras como o canto dessas aves. No entanto, esse método manual apresenta desvantagens, sendo demorado, suscetível a erros e dependentes do conhecimento e das habilidades dos especialistas (FIGUEIREDO *et al.*, 2018). Ademais, em consonância com o crescente foco na conservação da biodiversidade, há uma demanda em ascensão por abordagens mais eficazes e precisas na identificação das espécies de aves.

Na esfera científica, inúmeros estudos exploram diversas abordagens para a classificação de espécies de aves por meio da análise de seus cantos e chamados. Muitos desses métodos empregam tecnologias avançadas de extração de características em áudios e espectrogramas, tais como os algoritmos *handcrafted*, que se referem a técnicas de extração de características específicas, meticulosamente projetadas e ajustadas manualmente por especialistas para identificar padrões sonoros relevantes nos cantos de aves, como averiguado em (LUCIO; COSTA, 2016). Além disso, utiliza-se a abordagem de *Transfer Learning*, aplicado em (INCZE *et al.*, 2018) que consiste em aproveitar o conhecimento prévio de modelos treinados em outras tarefas para a extração de características ou a classificação de novos dados, tornando o processo mais eficiente e preciso. Essas técnicas são utilizadas para extrair características relevantes dos sinais de áudio, permitindo, posteriormente, a classificação por meio de algoritmos de aprendizagem supervisionada.

1.1 Objetivos

Nesse contexto, este trabalho tem como objetivo principal desenvolver um modelo de aprendizado de máquina que deve ser capaz de identificar diversas espécies de aves a partir das vocalizações dessas aves.

Para alcançar esse objetivo, será fundamental realizar uma avaliação abrangente de várias abordagens. Isso incluirá a utilização de descritores de textura *handcrafted* como *Local Binary Pattern* (LBP), filtros Gabor e outros semelhantes, bem como a aplicação de técnicas de *transfer learning* como VGG16 e RESNET50 para extrair as características dos espectrogramas

correspondentes às gravações. Essas características serão utilizadas na classificação por meio de diferentes modelos de aprendizado de máquina. O foco será encontrar os melhores parâmetros para otimizar o desempenho e atingir uma significativa taxa de acerto na classificação.

Acho sempre interessante colocar os Trabalhos Relacionados ao final do Capítulo de REFERENCIAL TEÓRICO. Desta forma, o leitor irá ler sobre os trabalhos já com o entendimento de vários conceitos apresentados ao longo do Capítulo.

2 REFERENCIAL TEÓRICO

2.1 Trabalhos Relacionados

É fundamental ressaltar que pesquisas similares a esta não são novidade, visto que existe um considerável acervo de estudos recentes dedicados à classificação de aves com base em suas vocalizações. Estes estudos empregam uma variada gama de técnicas de extração de características e treinamento de modelos, cada um adotando abordagens específicas para alcançar seus objetivos e explorando metodologias diversas na classificação das aves, além de utilizarem bases de dados variadas.

Em um artigo (LUCIO; COSTA, 2016), é proposta uma abordagem para a classificação automatizada de espécies de aves. Esta abordagem combina características acústicas, tais como o *Rhythm Histogram* (RH), *Rhythm Patterns* (RP) e *Statistical Spectrum Descriptor* (SSD), com características visuais, incluindo *Local Binary Patterns* (LBP), *Local Phase Quantization* (LPQ), *Run-Length Binary Patterns (RLBP)*, *Gray-Scale Level Co-occurrence Matrix* (GLCM) e Filtros de Gabor. Em seguida, a classificação é executada utilizando o SVM por meio da biblioteca LIBSVM, utilizando um conjunto de dados para treinamento e outro para teste. Para garantir resultados mais consistentes, emprega-se a técnica de validação cruzada. Esta abordagem utiliza um conjunto de dados com 2814 amostras de áudio de 46 espécies de aves, sendo ela um subconjunto da base de dados de cantos e chamados disponibilizada pelo site Xeno-Canto. Com isso, obteve um Recall de 91,08%, Precision de 94,02%, e F-Measure de 92,34

Em outro estudo (RAI *et al.*, 2016), é apresentado um método para a identificação automática de espécies de aves com base em seus cantos. O estudo se concentra em quatro espécies comuns do norte da Índia: o tordo-preto, o pato, o papagaio e o corvo doméstico, utilizando um subconjunto de 156 amostras de áudio da base de dados do Xeno-canto. O método de extração de características adotado é o *mel frequency cepstral coefficients* (MFCC), que descreve o espectro de um registro de áudio de forma concisa e informativa. A classificação é realizada por meio de um algoritmo baseado em *support vector machines* (SVM). Os resultados mais promissores foram obtidos para as espécies de tordo-preto e corvo doméstico, com uma acurácia de 73% e 89%, respectivamente, e uma acurácia geral de 64%, baseada na média entre todas as classes.

Um estudo recente sobre a identificação de espécies de aves a partir de dados de áudio (WU; KOSURU; TIPPAREDDY, 2023) adotou um conjunto de dados obtido do Xeno-canto, um site que reúne uma vasta coleção de sons da vida selvagem de todo o mundo. Os autores extraíram diversas características acústicas das gravações de áudio, incluindo a Transformada de Fourier de Curto Prazo (STFT), Coeficientes Cepstrais de Frequência Mel (MFCC), Energia RMS, Centróide Espectral, Largura de Banda Espectral, Roll-off Espectral e Taxa de Cruzamento de Zero. Estas características foram então utilizadas para alimentar vários modelos de aprendizado de máquina, como K-vizinhos mais próximos (KNN), Descida do Gradiente Esto-

cástica, SVM, Árvores de Decisão, Floresta Aleatória e Classificador de Bayes Ingênuo Gaussiano, para classificar as espécies de aves. Os melhores resultados foram alcançados utilizando a Descida do Gradiente Estocástica, quando treinada com um subconjunto de três espécies de aves, obtendo uma acurácia de 90% e um F1-score de 89%. Esses resultados ilustram o potencial do uso de dados de áudio e modelos de aprendizado de máquina na identificação e monitoramento de espécies de aves.

Além disso, em um estudo conduzido por (INCZE *et al.*, 2018), é apresentado um sistema de classificação de sons de aves baseado em redes neurais convolucionais (CNNs). O método envolve o ajuste de um modelo pré-treinado de CNN, o MobileNet, utilizando um conjunto de dados proveniente do portal de compartilhamento de sons de aves, o Xeno-canto. Espectrogramas gerados a partir dos dados baixados foram usados como entrada para a rede neural. Diversas configurações e hiperparâmetros foram avaliados em experimentos, incluindo o número de classes (espécies de aves) e o esquema de cores dos espectrogramas. Para 2 classes, o modelo treinado com imagens na escala de cores "jet"e em escala de cinza alcançou uma acurácia de 81% e 79%, respectivamente. Ao expandir o conjunto de classes para 10, a precisão diminuiu para 39% e 30%. Finalmente, ao classificar 50 classes distintas, o desempenho do modelo foi reduzido para uma acurácia de 20% e 10%. Os resultados indicam que a escolha de um esquema de cores compatível com as imagens com as quais a rede neural foi pré-treinada proporciona vantagens mensuráveis. Os resultados mais satisfatórios foram obtidos quando um número limitado de classes foi considerado, sugerindo que o sistema é mais eficaz quando aplicado a um pequeno conjunto de espécies de aves.

O artigo (LUCIO; MALDONADO; COSTA, 2015) apresenta um sistema para classificação automática de espécies de aves com base em características extraídas de imagens de espectrogramas. Os autores utilizaram três operadores de textura comuns para extrair características de textura e um classificador SVM para a classificação. O experimento foi realizado em um conjunto de dados composto por 46 classes, e a melhor taxa de precisão obtida foi de cerca de 77,65%. O método proposto neste trabalho envolveu a divisão da base de dados em *folds*, geração dos espectrogramas, extração das características e treinamento do classificador SVM com otimização de parâmetros. Para a extração de características, foram utilizados três operadores de textura: Local Binary Patterns (LBP), Gray-Level Co-occurrence Matrix (GLCM) e Local Phase Quantization (LPQ). O classificador SVM foi treinado com um kernel linear e otimizado com validação cruzada. Os resultados obtidos mostram que o sistema proposto é capaz de classificar com precisão várias espécies de aves com base em suas vocalizações, o que é relevante para a área de processamento de sinais e reconhecimento de padrões.

Estes estudos ilustram o amplo espectro de abordagens e técnicas empregadas na classificação de aves com base em suas vocalizações, revelando o progresso significativo alcançado na automação deste processo e destacando o potencial do uso de dados de áudio e modelos de aprendizado de máquina nesse campo.

2.2 Descritores de Características

A ideia por trás da utilização de descritores de características é a necessidade de mapear os áudios em uma representação de menor dimensionalidade do que a gravação original, mantendo informações relevantes para a classificação. Como mencionado anteriormente, os trabalhos nesta área empregam um amplo espectro de abordagens e técnicas. Isso ocorre devido ao fato de que em suas metodologias, a extração de características de uma fonte de dados específica, seja ela de áudio ou imagens, é uma prática comum. O que normalmente diferencia um trabalho do outro é a utilização de diferentes descritores de características, possibilitando uma variedade de combinações e tipos de descritores. Nesta seção, vamos explorar dois tipos distintos de descritores: aqueles elaborados manualmente com base no conhecimento do domínio (*Handcrafted*), e aqueles obtidos por meio da transferência de aprendizado (*Transfer Learning*).

2.2.1 Descritores Handcrafted

Os descritores de características *Handcrafted*, são técnicas de extração de características que foram projetadas para capturar informações específicas de uma imagem ou áudio, como padrões de cor, texturas ou formas, tornando-os interpretáveis e adaptados a tarefas específicas (NANNI; GHIDONI; BRAHNAM, 2017). Nesse sentido, a extração dessas características é normalmente considerada eficiente. Neste contexto, é importante destacar que a principal vantagem reside na interpretabilidade e na capacidade de incorporar conhecimento especializado no modelo. No entanto, para alcançar resultados satisfatórios, geralmente é necessário desenvolver descritores específicos, adaptados ao problema em questão. Por outro lado, existem descritores *handcrafted* que demonstram eficácia em uma variedade de problemas, como o LBP.

Dois exemplos comuns desses descritores são Local Binary Pattern (LBP) e os Filtros de Gabor:

Local Binary Pattern (LBP): O LBP é um descritor de textura utilizado em processamento de imagens (SILVA, 2017). Ele funciona comparando cada pixel central em uma imagem com seus pixels vizinhos e gerando um padrão binário local que representa a textura da região em torno do pixel central. O processo de cálculo do LBP começa com a escolha de um pixel central na imagem. Em seguida, é definido um limiar de intensidade para esse pixel. Para cada um dos pixels vizinhos, é comparada a sua intensidade com o limiar. Se a intensidade do pixel vizinho for maior ou igual ao limiar, é atribuído o valor binário 1. Caso contrário, é atribuído o valor binário 0. Esses valores binários são então concatenados em uma sequência binária, que representa o padrão local em torno do pixel central. O tamanho da vizinhança é definido pelo raio e pelo número de pontos. O raio define o tamanho da vizinhança circular em torno do pixel central, enquanto o número de pontos define o número de pixels vizinhos a serem considerados.

Por exemplo, um LBP com raio 1 e 8 pontos considera os 8 pixels vizinhos imediatos ao pixel central, como na figura 3. O resultado do LBP é um mapa de textura da imagem, onde cada pixel é substituído pelo seu valor LBP correspondente. Esse mapa pode ser utilizado como um descritor para reconhecimento de objetos ou detecção de anomalias em imagens.

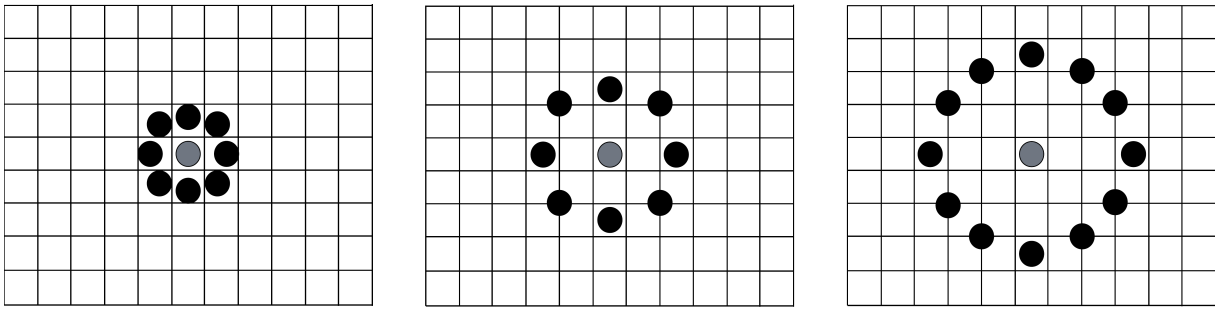

Figura 1 – Local Binary Pattern com diferentes valores de P e R

É estranho apresentar algo "útil" sem antes ter explicado pq é útil...

Filtros de Gabor: Os filtros de Gabor são uma família de filtros gaussianos usados principalmente para a extração de características de textura (ANGELO, 2001). Eles têm a propriedade útil de serem sensíveis à orientação, o que os torna muito úteis para muitas tarefas de visão computacional. Um filtro de Gabor é essencialmente uma função sinusoidal modulada por uma gaussiana. Eles são usados para capturar informações sobre a estrutura e orientação local da textura em uma imagem.

Esses descritores são chamados de “handcrafted” porque foram projetados manualmente para capturar certas propriedades das imagens. Eles contrastam com os descritores aprendidos automaticamente que são usados em aprendizado profundo, onde as características são aprendidas diretamente dos dados durante o treinamento de um modelo.

2.2.2 *Transfer Learning*

A transferência de aprendizado é um conjunto de técnicas usadas em aprendizagem profunda onde um modelo pré-treinado é usado como ponto de partida para a criação de um novo modelo. Esses modelos pré-treinados são geralmente treinados em grandes conjuntos de dados, como o ImageNet, e podem extrair características úteis de imagens que podem ser usadas para uma variedade de tarefas diferentes. Aqui estão alguns exemplos desses modelos:

VGG16: O VGG16 é uma arquitetura de rede neural convolucional (TAO *et al.*, 2021) que foi proposta pelos pesquisadores da Visual Geometry Group (VGG) da Universidade de Oxford. É composto por 16 camadas, incluindo 13 camadas convolucionais, seguidas por três camadas totalmente conectadas. O VGG16 é amplamente utilizado para transferência de aprendizado porque, apesar de sua simplicidade, ele pode capturar características com informações relevantes para classificação em vários contextos diferentes.

ResNet50: A ResNet50 (HANNE *et al.*, 2022) é uma variante da arquitetura ResNet (Residual Network) que tem 50 camadas. A ideia chave por trás das ResNets é a introdução de

“conexões residuais” que permitem o treinamento efetivo de redes muito profundas. A ResNet50 é frequentemente usada para transferência de aprendizado porque pode capturar uma grande quantidade de características visuais diferentes.

DenseNet: A DenseNet (Dense Convolutional Network) é uma rede que conecta (ZHANG *et al.*, 2019) cada camada a todas as outras camadas de uma maneira feed-forward. Isso resulta em redes com conexões densas, daí o nome DenseNet. A DenseNet-169 é uma variante dessa arquitetura que tem 169 camadas.

MobileNet: As MobileNets são redes neurais convolucionais eficientes (SINHA; EL-SHARKAWY, 2019) projetadas para dispositivos móveis e aplicações embarcadas. Elas são pequenas, com baixa latência e baixo consumo de energia, permitindo a implantação em uma variedade de plataformas.

Neste trabalho, a aplicação de *transfer learning* envolve a passagem das imagens através de uma rede neural, permitindo a captura das ativações em uma camada apropriada, geralmente localizada mais próxima ao final da rede. As características resultantes desse processo podem, então, ser empregadas para alimentar um modelo de aprendizagem convencional, como, por exemplo, um classificador KNN ou SVM.

Em todos esses casos, a ideia básica por trás da transferência de aprendizado é a seguinte: em vez de iniciar o treinamento do zero, parte-se de padrões previamente adquiridos pelo modelo, os quais são ajustados para a nova tarefa. Isso permite a utilização do conhecimento prévio do modelo, geralmente resultando em melhorias e maior rapidez nos resultados.

3 PROPOSTA

3.1 Materiais

Suponho que o dataset também faça parte dos materiais.

A linguagem de programação Python 3.11 foi selecionada para o desenvolvimento dos algoritmos necessários neste trabalho. O Python 3.11 foi escolhido devido a sua versatilidade e capacidade de atender as exigências do projeto.

Adicionalmente, diversas bibliotecas foram empregadas para suportar diferentes aspectos do projeto. A biblioteca librosa é utilizada para a geração de mel-espectrogramas, enquanto o matplotlib é usado na geração de imagens e visualizações gráficas. O numpy é empregado para cálculos matemáticos e manipulação eficiente de matrizes e vetores.

No contexto do processamento de imagens, o OpenCV é a ferramenta principal para leitura e escrita de imagens. Para lidar com dados estruturados, a biblioteca pandas é usada na criação de dataframes. Além disso, a scikit-image é empregada para implementar algoritmos de descritores *handcrafted*, como o LBP e Filtros Gabor.

Para a construção e avaliação de modelos de aprendizagem de máquina, o scikit-learn (sklearn) oferece múltiplos algoritmos e ferramentas para análise preditiva. Por fim, o TensorFlow é utilizado para implementar descritores de *transfer learning*, como o VGG16, permitindo a utilização de modelos de redes neurais pré-treinadas para extração de características.

3.2 Dataset

Não está claro (no texto e na apresentação) se cada áudio possui o canto de somente uma espécie. Também seria interessante dizer o tempo médio de um canto e/ou chamado. E ainda se existem diferenças significativas para a classificação entre um e outro.

O dataset do desafio BirdCLEF 2023 (BIRDCLEF, 2023) consiste em gravações de paisagens sonoras do Quênia, abrangendo uma vasta diversidade de ecossistemas, e contém registros de cantos e chamados de aves. Este conjunto de dados é composto por uma variedade de espécies de aves, com um total de mais de 200 classes de espécies diferentes. As gravações de áudio abrangem diversas faixas sonoras que capturam os sons dessas aves em seus habitats naturais. No entanto, apesar de ser um banco de dados robusto, a distribuição de exemplos em cada classe é desbalanceada, como mostra a figura 2.

3.3 Metodo Proposto

No desenvolvimento deste estudo, o método seguirá uma abordagem composta pelos seguintes passos: geração dos espectrogramas, subdivisão em *patches*, extração de características e treinamento do modelo, como mostrado na Figura 3.

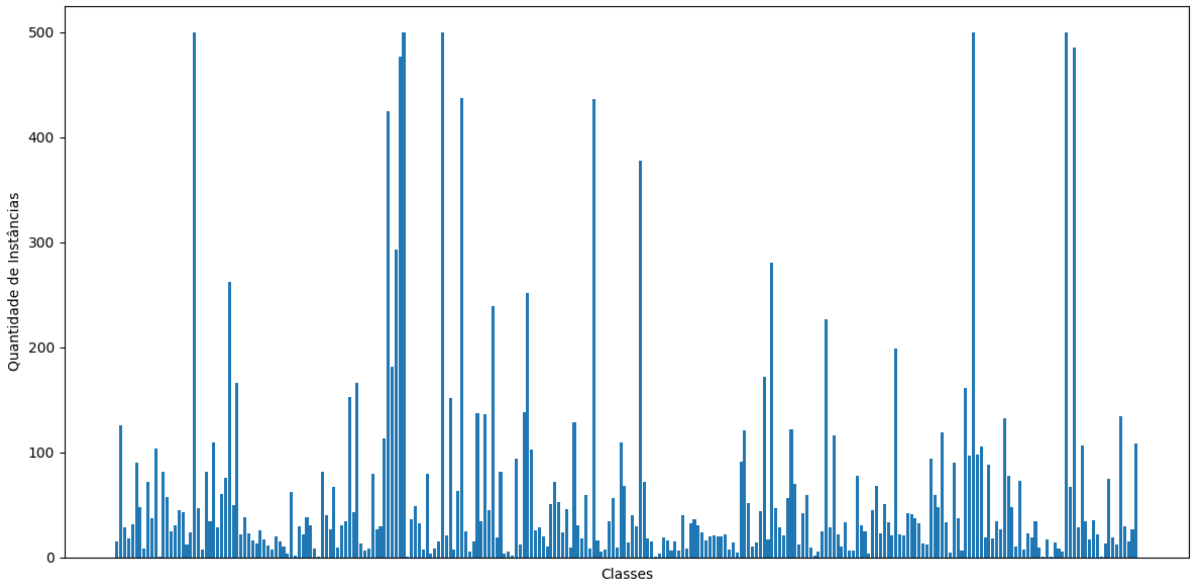

Figura 2 – Distribuição de exemplos em cada classe

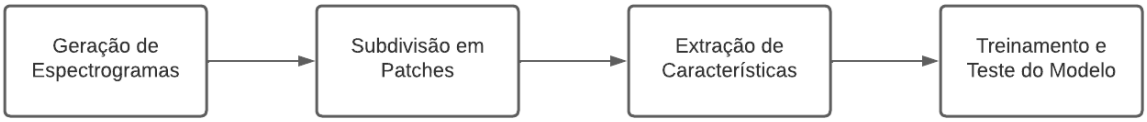

Figura 3 – Método proposto

3.3.1 Geração dos Espectrogramas

O processo de geração dos espectrogramas em escala mel se inicia com o calculo da STFT do sinal de áudio. Para isso, 2048 amostras serão usadas para calcular cada transformada de Fourier que compõe a STFT. Além disso, para obter um espectrograma mais detalhado nas variações de amplitude em frequência, há uma sobreposição de 50% das amostras entre janelas subsequentes.

Depois de realizar o cálculo do espectrograma, um conjunto de filtros é empregado para converter o espectrograma resultante em uma representação na escala Mel. Essa transformação tem como finalidade destacar as faixas de frequência que correspondem à sensibilidade auditiva humana, de modo a possibilitar que o espectrograma reproduza, em certa medida, a percepção auditiva humana. Inúmeras pesquisas anteriores na área de classificação de áudio demonstram consistentemente que os resultados tendem a ser superiores quando se opta por utilizar os Mel-espectrogramas.

Por último, uma imagem do espectrograma é gerada, na qual o espectrograma Mel é representado em cores, como apresentado na Figura 4.

Qual é a necessidade de corresponder à sensibilidade auditiva humana (em termos de utilização de ML)?

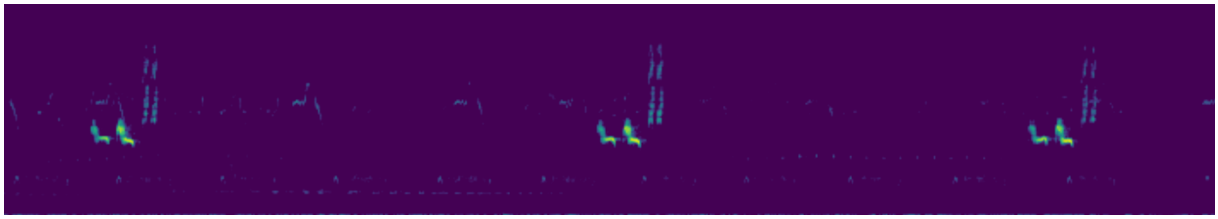

Figura 4 – Espectrograma da faixa de audio XC128013.ogg

3.3.2 Subdivisão em Patches

Neste estágio, o objetivo principal é processar os espectrogramas, dividindo-os em unidades menores denominadas "*patches*"com uma largura de N colunas. É importante ressaltar que os espectrogramas podem apresentar tamanhos variados, uma vez que estão diretamente relacionados ao comprimento do áudio. Essa variação decorre da natureza não uniforme do conteúdo do áudio, onde diversos sons ocorrem em diferentes momentos.

Para lidar com essa heterogeneidade sonora ao longo do áudio, torna-se necessário extrair características de diversas janelas de tempo. Consequentemente, a divisão dos *patches* pode resultar em um último *patch* contendo menos de N colunas. A solução implementada consiste na adição de preenchimento de zeros a esse último *patch*, assegurando que todos os *patches* compartilhem dimensões idênticas. Essa abordagem, apresentada na figura 5, é crucial para facilitar etapas subsequentes da análise, possibilitando a consideração da diversidade de sons presentes em intervalos temporais distintos do áudio.

3.3.3 Extração de Caracteristicas

Nesta etapa do trabalho, iremos realizar o processamento de cada um dos *patches* a fim de extrair suas características. Além disso, determinaremos o número de classes a serem utilizadas com base na quantidade de amostras disponíveis para este experimento.

O nosso principal objetivo consiste em extrair características das classes que possuem as maiores quantidades de amostras. Isso nos permitirá avaliar o desempenho futuro no treinamento de forma mais eficaz, uma vez que essa abordagem em dividir o dataset em subconjuntos com um número menor de classes, busca simular um ambiente mais próximo da realidade. Isso é particularmente relevante, pois ao lidar com a análise de um espaço geográfico específico, é pouco provável que exista muitas espécies coexistindo naquele ambiente. Portanto, essa estratégia de seleção das classes é essencial para tornar a nossa análise mais robusta e aplicável a cenários reais.

No que diz respeito aos descritores utilizados neste experimento, incluímos o LBP, Gabor, VGG16, RESNET50, DenseNet e MobileNet, que foram apresentados na seção anterior do trabalho. Esta variedade de descritores foi escolhida pois são amplamente utilizados em classificação de áudio, obtendo bons resultados em vários problemas.

Existe alguma informação relevante para extrair um subconjunto de classes que coexistem em um mesmo cenário real? Se não, este seria um estudo relevante?

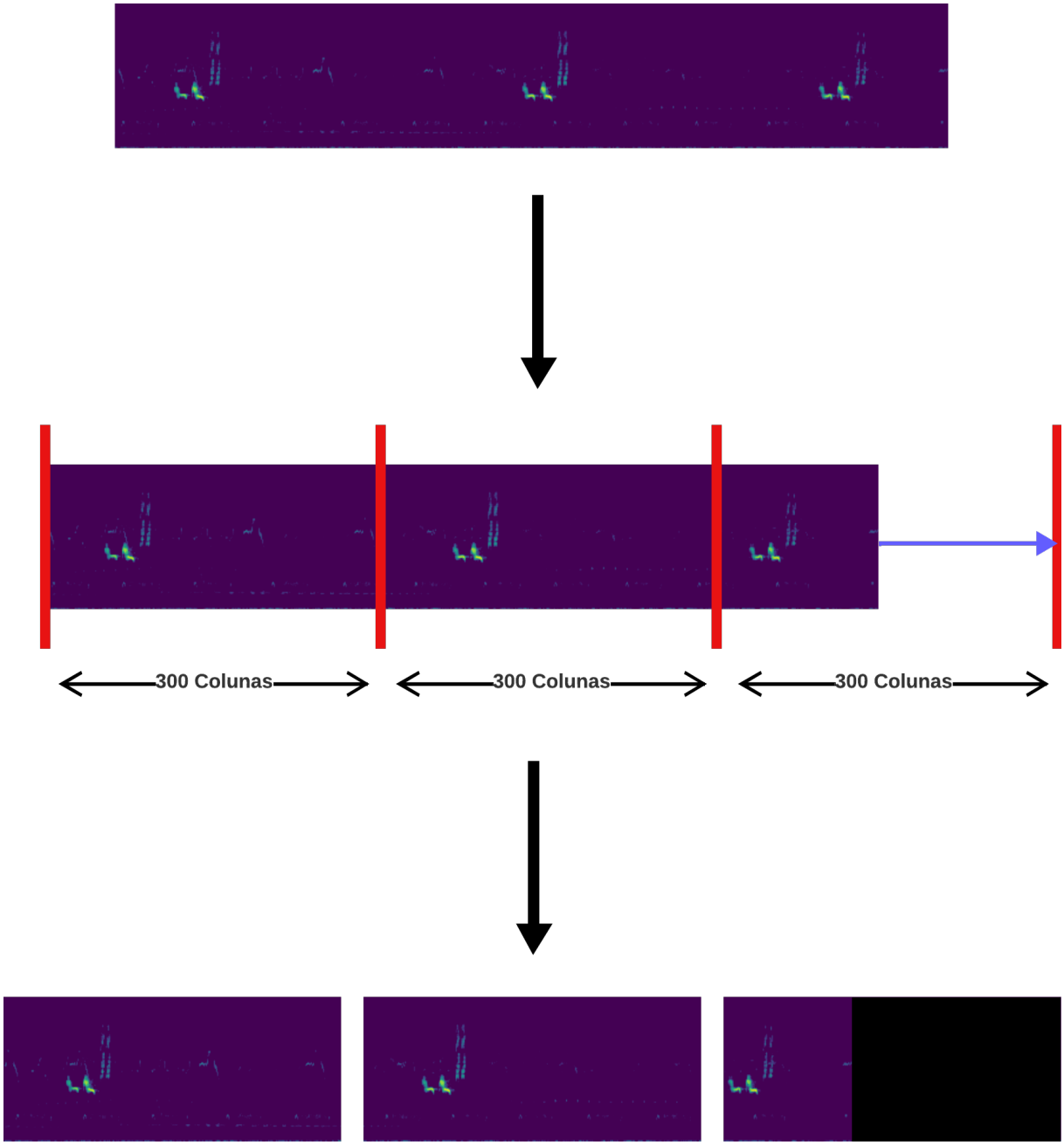

Figura 5 – Geração de *patches* com 300 colunas da faixa de audio XC128013.ogg

3.3.4 Treinamento

Nesta etapa, realiza-se o treinamento e o teste visando avaliar a qualidade do modelo. Nesse sentido, é necessário dividir quais arquivos serão alocados para treinamento e quais serão destinados aos testes. Essa abordagem utiliza a separação em folds aplicada a um vetor que contém os nomes de todos os áudios de entrada e suas classes, com 10 *folds*.

Dessa forma, obtém-se o nome dos arquivos separados para os conjuntos de treinamento e teste para cada *split*. Com esses nomes, são identificados todos os *patches* gerados a partir dos arquivos alocados para treinamento e teste. Isso é fundamental para garantir que todos os *patches* de um arquivo específico sejam tratados exclusivamente como dados de trei-

namento ou teste, evitando qualquer compartilhamento que possa comprometer a validação do modelo.

Em seguida, os dados são normalizados e submetidos ao treinamento, tanto com o algoritmo KNN quanto com o SVM, utilizando a técnica de busca exaustiva para ajustar hiperparametros do modelo. Posteriormente, um novo modelo é treinado com os melhores parâmetros encontrados e testado com as amostras de teste. Durante esse processo, é realizada uma votação em relação aos *patches* de cada arquivo, a fim de determinar a classe que foi predominantemente atribuída a cada espectrograma. A votação é uma técnica que aumenta a confiabilidade dos resultados, uma vez que permite obter uma decisão consensual com base nas previsões dos *patches* individuais.

Esse processo é feito para cada um dos *split* e, neste caso, as taxas de acerto do modelo são calculadas a partir da média dos resultados nos *folds* de teste.

3.4 Resultados Preliminares

Testes preliminares foram realizados para avaliar a viabilidade inicial do estudo. Nesse contexto, foram gerados espectrogramas na escala Mel, conforme definido no método proposto, e criados *patches* contendo 300 colunas, com a aplicação de preenchimento quando necessário.

Posteriormente, procedeu-se à extração de características das classes com maior número de amostras, especificamente das duas, três, cinco e dez classes mais representativas.As proporções dessas classes em relação ao total de gravações do conjunto de dados podem ser observadas na Tabela 1. Para realizar essa extração, foi utilizado o descritor LBP com uma configuração de 8 pontos e um raio de 2. O modelo foi treinado com o algoritmo KNN com a otimização dos hiperparâmetros abrangendo os valores de K vizinhos mais próximos de 1 a 20, com incremento de 2 unidades. Nos testes preliminares foi usado apenas um split, com 80% dos dados para treino e 20% para teste. A tabela 2 mostra os resultados obtidos com KNN para as 2,3,5 e 10 classes com mais exemplos.

	2 Classes	3 Classes	5 Classes	10 Classes
Relação	5,90%	8,85%	14,76%	28,10%

Tabela 1 – Relação entre a quantidade de gravações nos subconjuntos avaliados e a quantidade total de áudios na base de dados

Em seguida, o mesmo procedimento foi repetido, mas utilizando a rede neural VGG16 para a extração de características. Novamente, o modelo foi treinado com o algoritmo KNN, obtendo os resultados mostrados na tabela 2.

De maneira inicial, observa-se que à medida que mais classes são envolvidas no processo de treinamento, a pontuação F1 tende a diminuir. Isso pode ser atribuído ao desbalanceamento das classes e à presença de ruídos semelhantes nas faixas de áudio, que tornam a tarefa

de classificação mais desafiadora. Também é relevante observar que, à primeira vista, a VGG16 parece apresentar uma ligeira vantagem em relação ao método LBP. No entanto, é importante ressaltar que a significância estatística dessa diferença só poderá ser devidamente avaliada por meio da realização de validação cruzada em k folds. Este procedimento será minuciosamente abordado e conduzido durante o desenvolvimento do TCC2, permitindo uma análise estatística robusta e conclusiva das comparações entre os dois métodos.

Por fim, é válido incluir que os resultados preliminares sugerem que o método em questão apresenta potencial para alcançar resultados satisfatórios. Essas observações iniciais motivam a continuação deste estudo com a perspectiva de obter uma compreensão mais aprofundada e sólida acerca de seu desempenho e aplicabilidade.

KNN				
Método	2 Classes	3 Classes	5 Classes	10 Classes
LBP	82%	68%	58%	43%
VGG16	86%	69%	63%	47%

Tabela 2 – Pontuação F1 obtida com os descritores LBP e VGG16

3.5 Cronograma

A Tabela 3 apresenta o cronograma de atividades programadas para o ano de 2024 neste projeto. Em linhas gerais, os dois primeiros meses serão dedicados tanto ao estudo dos descritores ainda não implementados nesta fase inicial, quanto à geração de espectrogramas e extração de características, utilizando os métodos discutidos neste trabalho. Essas etapas visam à obtenção de todos os dados iniciais necessários para o desenvolvimento dos testes. Já nos dois próximos meses serão alocados para o treinamento dos diversos modelos utilizando o algoritmo KNN e SVM, o qual serão empregados na análise deste estudo. A parte da escrita será feita durante todo o processo de desenvolvimento desse trabalho e, por fim, a defesa será feita no inicio de julho.

2024					
Atividades	MAR	ABR	MAI	JUN	JUL
Estudo sobre outros descritores de características	x	x			
Geração dos espectrogramas	x				
Geração dos patches	x				
Extração de Caracteristicas com o LBP	x				
Extração de Caracteristicas com os filtros Gabor	x				
Extração de Caracteristicas com o VGG16		x			
Extração de Caracteristicas com o RESNET50		x			
Extração de Caracteristicas com o DenseNet		x			
Extração de Caracteristicas com o MobileNet		x			
Treinamento e Teste do modelo usando SVM	x	x	x	x	
Escrita do TCC	x	x	x	x	x
Defesa do TCC					x

Tabela 3 – Cronograma

Quadro

REFERÊNCIAS

ANGELO, N. P. **INVESTIGAçãO COM RESPEITO A APLICAçãO DOS FILTROS DE GABOR NA CLASSIFICAçãO SUPERVISIONADA DE IMAGENS DIGITAIS**. 2001. Tese (Doutorado) — Universidade Federal do Rio Grande do Sul, 2001.

BIRDCLEF. 2023. Cornell Lab of Ornithology. Disponível em: https://www.kaggle.com/competitions/birdclef-2023/data. Acesso em: 07 out. 2023.

CARPENTER, F. L. A spectrum of nectar-eater communities. **American Zoologist**, Oxford University Press, v. 18, n. 4, p. 809–819, 1978. ISSN 00031569. Disponível em: http://www.jstor.org/stable/3882538.

FIGUEIREDO, N. *et al.* A comparative study on filtering and classification of bird songs. *In*: **Sound and Music Computing Conference - SMC**. [*S.l.*]: SMC, 2018.

HANNE, L. *et al.* Perception of flaws in steel using resnet50 algorithm. *In*: **2022 International Conference on Knowledge Engineering and Communication Systems (ICKES)**. [*S.l.*: *s.n.*], 2022. p. 1–5.

HOLMES, R. T. Ecological and evolutionary impacts of bird predation on forest insects an overview. **Studies in Avian Biology**, v. 13, p. 6–13, 1990.

INCZE, A. *et al.* Bird sound recognition using a convolutional neural network. *In*: **2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY)**. [*S.l.*: *s.n.*], 2018. p. 000295–000300.

LUCIO, D. R.; COSTA, Y. M. e Gomes da. Bird species classification using visual and acoustic features extracted from audio signal. *In*: **2016 35th International Conference of the Chilean Computer Science Society (SCCC)**. [*S.l.*: *s.n.*], 2016. p. 1–12.

LUCIO, D. R.; MALDONADO, Y.; COSTA, G. da. Bird species classification using spectrograms. *In*: **2015 Latin American Computing Conference (CLEI)**. [*S.l.*: *s.n.*], 2015. p. 1–11.

NANNI, L.; GHIDONI, S.; BRAHNAM, S. Handcrafted vs. non-handcrafted features for computer vision classification. **Pattern Recognition**, v. 71, p. 158–172, 2017. ISSN 0031-3203. Disponível em: https://www.sciencedirect.com/science/article/pii/S0031320317302224.

RAI, P. *et al.* An automatic classification of bird species using audio feature extraction and support vector machines. *In*: **2016 International Conference on Inventive Computation Technologies (ICICT)**. [*S.l.*: *s.n.*], 2016. v. 1, p. 1–5.

SILVA, J. A. da. **Detecção de Imagens Manipuladas utilizando Descritores Locais**. 2017. Tese (Doutorado) — Universidade Federal de Pernambuco, 2017. Disponível em: https://www.cin.ufpe.br/~tg/2017-1/jas4_tg.pdf.

SINHA, D.; EL-SHARKAWY, M. Thin mobilenet: An enhanced mobilenet architecture. *In*: **2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)**. [*S.l.*: *s.n.*], 2019. p. 0280–0285.

SNOW, D. W. Evolutionary aspects of fruit-eating by birds. **Ibis**, v. 113, p. 194–202, 1971.

TAO, J. *et al.* Research on vgg16 convolutional neural network feature classification algorithm based on transfer learning. *In*: **2021 2nd China International SAR Symposium (CISS)**. [*S.l.*: *s.n.*], 2021. p. 1–3.

WU, C. S.; KOSURU, S.; TIPPAREDDY, S. Bird species identification from audio data. *In*: **2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService)**. [*S.l.*: *s.n.*], 2023. p. 58–62.

ZHANG, K. *et al.* Multiple feature reweight densenet for image classification. **IEEE Access**, v. 7, p. 9872–9880, 2019.