UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

FELIPE ARCHANJO DA CUNHA MENDES

CLASSIFICAÇÃO DE PÁSSAROS POR MEIO DE CANTOS E CHAMADOS: UMA ABORDAGEM INTEGRADA COM ESPECTROGRAMAS EM CONJUNTO COM DESCRITORES HANDCRAFTED E TRANSFER LEARNING

CAMPO MOURÃO

2024

FELIPE ARCHANJO DA CUNHA MENDES

CLASSIFICAÇÃO DE PÁSSAROS POR MEIO DE CANTOS E CHAMADOS: UMA ABORDAGEM INTEGRADA COM ESPECTROGRAMAS EM CONJUNTO COM DESCRITORES HANDCRAFTED E TRANSFER LEARNING

Bird Classification Through Songs and Calls: An Integrated Approach with Spectrograms in Conjunction with Handcrafted Descriptors and Transfer Learning

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Juliano Henrique Foleis

CAMPO MOURÃO

2024

FELIPE ARCHANJO DA CUNHA MENDES

CLASSIFICAÇÃO DE PÁSSAROS POR MEIO DE CANTOS E CHAMADOS: UMA ABORDAGEM INTEGRADA COM ESPECTROGRAMAS EM CONJUNTO COM DESCRITORES HANDCRAFTED E TRANSFER LEARNING

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Ciência da Computação do Curso de Bacharelado em Ciência da Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 12/08/2024

Juliano Henrique Foleis
Doutorado
Universidade Tecnológica Federal do Paraná

Rodrigo Campiolo
Doutorado
Universidade Tecnológica Federal do Paraná

Rodrigo Hubner
Doutorado
Universidade Tecnológica Federal do Paraná

CAMPO MOURÃO
2024

Dedico este trabalho ao meu pai, cuja vida foi um mosaico de sacrifícios pela nossa família. Seu amor inabalável e dedicação permitiram-me chegar até aqui. Pai, cada página deste trabalho carrega um pouco da força e da inspiração que você me deu. Obrigado por tudo.

AGRADECIMENTOS

É impossível expressar minha gratidão a todas as pessoas que marcaram essa jornada significativa de minha vida nas linhas que se seguem. Por isso, peço desculpas antecipadamente àqueles que não foram mencionados diretamente aqui. Saibam que, mesmo ausentes nestas linhas, ocupam um lugar de destaque em meus pensamentos e gratidão.

Meus sinceros agradecimentos ao Prof. Dr. Juliano Henrique Foleis, meu orientador, cuja sabedoria foi essencial nesta caminhada. Sua orientação e conhecimento foram fundamentais para a execução deste trabalho.

Quero expressar minha profunda gratidão à minha família, cujo apoio incondicional foi crucial para superar os desafios enfrentados. Sem eles, esta conquista não seria possível.

Em especial, agradeço a meu pai, Antonio da Cunha Mendes, e minha mãe, Rute Archanjo, pelo investimento e sacrifícios feitos ao longo destes anos. Permitiram-me uma vida confortável no Paraná, focada apenas nos estudos, o que foi decisivo para minhas conquistas acadêmicas.

Não posso deixar de mencionar meu gato, Frajola, adotado em Campinas no mês em que fui aprovado no curso, em janeiro de 2020. Seu companheirismo foi um suporte emocional inestimável nos momentos mais difíceis dessa jornada, oferecendo-me um apoio psicológico singular.

Por fim, minha gratidão se estende a todos que, de alguma forma, contribuíram para a realização desta pesquisa.

Pode-se dizer que um homem pode ’injetar’
uma ideia na máquina, e que ela responderá
até certo ponto e depois cairá em quiescência,
como uma corda de piano atingida por um
martelo (TURING, 1950)

RESUMO

A classificação de espécies de pássaros é uma tarefa importante no monitoramento e conservação do meio ambiente. Neste estudo, o objetivo foi desenvolver um modelo de aprendizado de máquina destinado a classificar distintas espécies de pássaros com base em seus cantos e chamados. Inicialmente, espectrogramas foram gerados para cada gravação disponível no banco de dados. Posteriormente, adotou-se uma estratégia de divisão desses espectrogramas em segmentos menores, denominados *patches*, para investigar a influência dessa abordagem na performance do modelo. Diversas técnicas de extração de características foram exploradas, incluindo descritores *handcrafted*, como o *Local Binary Pattern (LBP)* e Filtros Gabor, e descritores derivados de modelos de *Transfer Learning*, tais como *VGG16*, *ResNet50*, *DenseNet121* e *MobileNet*. Esse processo foi realizado utilizando conjuntos de duas, três, cinco e dez classes, com o objetivo de analisar o comportamento do resultado conforme o número de classes aumenta. Além disso, foram conduzidos experimentos utilizando o classificador *Support Vector Machine (SVM)* para avaliar a precisão e eficácia do modelo. Os resultados revelaram que os descritores baseados em *Transfer Learning* apresentaram desempenho superior, com destaque para o *Resnet50*, que obteve uma pontuação F1 de 0,8994 na classificação entre duas espécies de pássaros, com o espectrograma dividido em três *patches*. Outros modelos, como o DenseNet121 e o MobileNet, também mostraram resultados competitivos na mesma configuração com dois pássaros: o DenseNet121 atingiu uma pontuação F1 de 0,8992 utilizando apenas um patch, enquanto o MobileNet registrou uma pontuação de 0,8989 com três patches. Esses achados sugerem que a abordagem de divisão dos espectrogramas em *patches*, combinada com técnicas extração de características baseadas em *transfer-learning*, podem ser eficazes para a classificação de espécies de pássaros baseada em seus cantos e chamados. Os resultados também mostraram que conforme o número de espécies aumenta, as taxas de classificação caem rapidamente.

Palavras-chave: classificação de espécies de pássaros; espectrogramas; aprendizagem de máquina; reconhecimento de padrões; descritores de características.

ABSTRACT

The classification of bird species is an important task in environmental monitoring and conservation. This study aimed to develop a machine-learning model to classify bird species based on their songs and calls. Initially, spectrograms were generated for each recording available in the database. Subsequently, a strategy of dividing these spectrograms into smaller segments, called patches, was adopted to investigate the influence of this approach on the model's performance. Various feature extraction techniques were explored, including handcrafted descriptors such as LBP and Gabor Filters, and descriptors derived from Transfer Learning models, such as *VGG16*, *ResNet50*, *DenseNet121*, and *MobileNet*. This process used sets with two, three, five, and ten classes to analyze the outcome behavior as this subset gradually increased. Additionally, experiments using the SVM classifier were conducted to evaluate the accuracy and effectiveness of the model. The results revealed that the descriptors based on Transfer Learning showed superior performance, with particular emphasis on *ResNet50*, which achieved an F1 score of 0.8994 in classifying two birds with the spectrogram divided into three patches. Other models, such as *DenseNet121* and *MobileNet*, also showed competitive results in the same configuration with two birds: *DenseNet121* achieved an F1 score of 0.8992 using just one patch. In contrast, *MobileNet* scored 0.8989 with three patches. These findings suggest that dividing the spectrograms into patches, combined with advanced feature extraction techniques, can effectively classify bird species based on their songs and calls. The results also show that classification performance drops sharply as the number of species increases.

Keywords: bird species classification; spectrograms; machine learning; pattern recognition; feature descriptors.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Siglas

CNN	Redes Neurais Convolucionais
DenseNet	Dense Convolutional Network
GLCM	Gray-Scale Level Co-occurrence Matrix
KDE	Kernel Density Estimation
KNN	K-Nearest Neighbors
LBP	Local Binary Pattern
LPQ	Local Phase Quantization
MFCC	Mel Frequency Cepstral Coefficients
RH	Rhythm Histogram
RLBP	Run-Length Binary Patterns
RP	Rhythm Patterns
SGD	Stochastic Gradient Descent
SSD	Statistical Spectrum Descriptor
STFT	Transformada de Fourier de Curto Prazo
SVM	Support Vector Machine
VGG	Visual Geometry Group

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, os pássaros têm um papel de destaque como indicadores da saúde dos ecossistemas, atraindo uma considerável quantidade de pesquisas científicas.

A correta identificação das espécies de pássaros representa um desafio complexo, que exige conhecimento especializado e experiência por parte dos ornitólogos. Tradicionalmente, a identificação dos pássaros se baseia em características morfológicas, tais como tamanho, formato do bico, plumagem e vocalização ou sonoras como seus cantos e chamados. 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 pássaros.

Na esfera científica, inúmeros estudos exploram diversas abordagens para resolver esse problema, ao classificar de forma automatica essas especies utilizando algoritmos de aprendizagem de máquina aplicados a gravações dos cantos e chamados de pássaros. 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 pássaros, 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

O objetivo principal deste trabalho foi desenvolver um modelo de aprendizado de máquina que seja capaz de classificar diversas espécies de pássaros com base em seus cantos e chamados.

Para alcançar o objetivo geral, foi fundamental realizar uma avaliação abrangente das técnicas de extração de características. Esta avaliação incluiu tanto descritores de textura *handcrafted*, como *LBP* e Filtros Gabor, quanto técnicas modernas de *Transfer Learning*, tais como

a extração de características dos modelos *VGG16*, *RESNET50*, *DENSENET* e *MobileNet*. Com isso, as características extraídas foram aplicadas na classificação através do *SVM*. Ajustar os parâmetros desse modelo foi essencial para otimizar o desempenho do sistema e alcançar uma taxa de acerto significativa na classificação das espécies.

Outro aspecto avaliado foi o impacto de subdividir os espectrogramas em N patches sobre o desempenho do modelo, investigando se tal abordagem resultaria em melhorias ou deteriorações na sua eficácia.

Além disso, examinou-se o efeito de variar o número de classes provenientes de uma região geográfica específica no comportamento do modelo, especialmente em relação às mudanças observadas com o aumento do número de classes.

Essas etapas serão integradas para garantir que o modelo desenvolvido não apenas funcione com eficácia, mas também contribua para avanços no campo da bioacústica, permitindo uma identificação precisa e automatizada de espécies de pássaros a partir de seus cantos característicos.

1.2 Justificativa

Este trabalho justifica-se pela oportunidade de aplicar conhecimentos teóricos e práticos adquiridos ao longo do curso em um problema real, demonstrando a viabilidade de soluções tecnológicas avançadas em questões de relevância ambiental e científica. Além disso, a implementação e análise de modelos de aprendizado de máquina neste contexto específico possibilita a exploração de dados complexos e a produção de conhecimento aplicado, contribuindo para o aprimoramento das técnicas de monitoramento da biodiversidade.

2 REFERENCIAL TEÓRICO

O conceito subjacente 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.

Os trabalhos nesta área empregam diversas abordagens e técnicas. Isso ocorre devido ao fato de que em seus metodos, 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.

2.1 Descritores *Handcrafted*

Os descritores *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). Neste contexto, é importante destacar que a principal vantagem reside na interpretabilidade e na capacidade de incorporar conhecimento especializado no modelo. Desta forma, para alcançar resultados satisfatórios, geralmente é necessário desenvolver descritores específicos, adaptados ao problema em questão. Entretanto, existem descritores *handcrafted* que demonstram eficácia em uma variedade de problemas.

***LBP*:** O *LBP* é um descritor de textura utilizado em processamento de imagens (OJALA; PIETIKAINEN; MAENPAA, 2002). 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 1. O resultado do *LBP* é um mapa de textura da imagem, onde cada pixel é substituído pelo seu valor *LBP* correspondente. Para gerar um vetor de características útil para problemas posteriores de classificacação, o próximo passo envolve a criação de um histograma a partir dos padrões LBP obtidos. Especificamente, este histograma é construído contando a frequência de cada padrão binário local encontrado na imagem. Na variação mais comum, focada em padrões uniformes de LBP, um padrão é considerado uniforme se contiver

no máximo duas transições de 0 para 1 ou de 1 para 0 em sua sequência binária. Esses padrões uniformes são particularmente importantes porque tendem a cobrir a maior parte das texturas comuns em imagens naturais, reduzindo a dimensão do histograma enquanto mantêm a maioria das informações úteis.

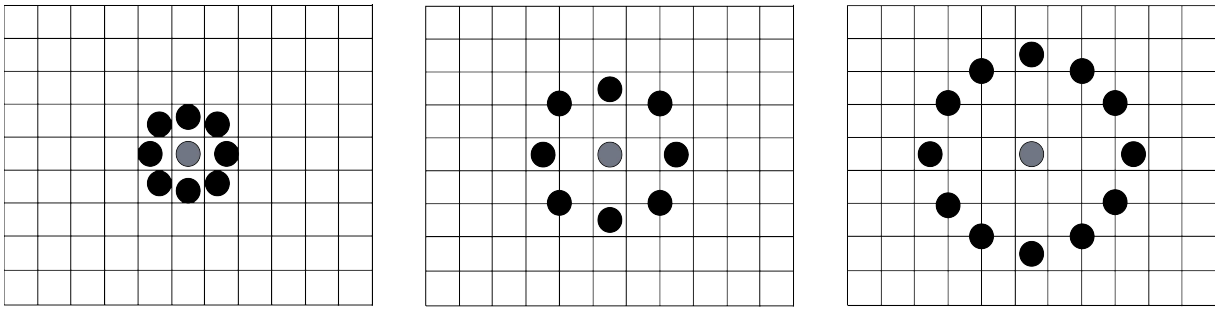


Figura 1 – *LBP* com diferentes valores de P e R

Filtros de Gabor: Os filtros de Gabor, criados por Dennis Gabor, são amplamente empregados em processamento de imagens e visão computacional (ANGELO, 2001) para tarefas como análise de texturas, detecção de bordas e extração de características locais, graças à sua capacidade de analisar propriedades espaciais de imagens, como orientação e escala. Esses filtros combinam uma função gaussiana, que suaviza a imagem e reduz ruídos, com uma onda sinusoidal, facilitando a detecção de padrões específicos. A eficácia dos filtros de Gabor reside na sua habilidade para capturar características distintas em diversas escalas e orientações, ajustáveis por meio de parâmetros como a orientação e frequência da onda sinusoidal, a largura da gaussiana (sigma), e a fase da onda, permitindo a detecção precisa de estruturas e bordas em direções específicas.

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 Descritores *Transfer Learning*

A transferência de aprendizado (LIANG; FU; YI, 2019) constitui um arcabouço de técnicas empregadas na aprendizagem profunda, caracterizando-se pelo aproveitamento de modelos previamente treinados como ponto de partida para a elaboração de novos modelos. Em geral, esses modelos pré-treinados são desenvolvidos com base em vastos conjuntos de dados, sendo a base de dados *ImageNet* um exemplo proeminente devido à sua ampla utilização. A *ImageNet* é notável por conter uma grande variedade de classes, especificamente 1000 classes, e milhões de imagens, o que proporciona uma capacidade de identificar e extrair características relevantes de imagens. Essas características, extraídas a partir dos modelos treinados com *ImageNet*, são aproveitáveis em um leque diversificado de tarefas. Uma das formas mais

comuns de aplicar a transferência de aprendizado é utilizar esses modelos para a extração de características, que são posteriormente classificadas por um modelo de aprendizagem supervisionada. Esta técnica, que está sendo utilizada no presente trabalho, permite que os benefícios de um modelo treinado em um domínio rico e diversificado possam ser transferidos para um novo problema com relativa facilidade.

Entre os exemplos mais notáveis de modelos que se beneficiam dessa técnica, incluem-se:

***VGG16*:** O *VGG16* (TAO *et al.*, 2021) é uma arquitetura de rede neural convolucional desenvolvida pelos pesquisadores do grupo *Visual Geometry Group (VGG)* da Universidade de Oxford. Essa arquitetura é composta por 16 camadas, sendo 13 delas convolucionais e as três restantes totalmente conectadas. O *VGG16* possui um total de 138 milhões de parâmetros. Quando empregado como um extrator de características pré-treinado, o *VGG16* é capaz de extrair um vetor de 512 características se a última camada for removida, ou um vetor de 1000 características se a última camada for mantida. Devido à sua capacidade de capturar características essenciais para a classificação em diversos contextos, o *VGG16* é amplamente utilizado para transferência de aprendizado, destacando-se pela sua eficiência apesar de sua aparente simplicidade.

***ResNet50*:** A *ResNet50* (HANNE *et al.*, 2022), é uma variante com 50 camadas da arquitetura *ResNet* (Redes Residuais). A principal inovação das *ResNets* é a adoção de "conexões residuais" que facilitam o treinamento de redes extremamente profundas, superando o problema de desvanecimento do gradiente que ocorre em redes mais simples. A *ResNet50* é amplamente empregada em técnicas de transferência de aprendizado, devido à sua capacidade de capturar uma ampla variedade de características visuais. Esta arquitetura possui cerca de 25.6 milhões de parâmetros. Além disso, sua camada totalmente conectada pode extrair 2048 características distintas de uma imagem, tornando-a particularmente útil para tarefas de visão computacional.

***DenseNet121*:** A *Dense Convolutional Network (DenseNet)* (ZHANG *et al.*, 2019) representa uma abordagem inovadora na arquitetura de redes neurais para visão computacional, especificamente em tarefas de classificação de imagens. O princípio fundamental da DenseNet é conectar cada camada a todas as outras camadas subsequentes de maneira *feed-forward*. Isso significa que cada camada recebe como entrada todas as saídas das camadas anteriores, promovendo a reutilização de características e reduzindo o problema de desvanecimento de gradientes em redes profundas. Esta estrutura resulta em redes com conexões extremamente densas, o que explica o nome *DenseNet*. Uma das principais vantagens desse tipo de arquitetura é a eficiência na utilização dos parâmetros. A *DenseNet121*, especificamente, possui cerca de 8 milhões de parâmetros, o que é relativamente baixo quando comparado com outras arquiteturas de redes profundas. Essa eficiência se traduz em uma redução significativa do risco de overfitting, mesmo com um número menor de parâmetros. Além disso, a DenseNet121 é capaz de extrair 1024 características ao final de sua última camada densa, o que lhe permite captar uma ampla gama de padrões e detalhes nas imagens. Isso é particularmente útil em tarefas

de classificação de imagens, onde detalhes finos podem ser cruciais para o desempenho do modelo. A *DenseNet-169* é uma variante maior dessa arquitetura, contendo 169 camadas, oferecendo uma capacidade ainda maior de modelagem e extração de características complexas, adequada para conjuntos de dados mais desafiadores e extensos.

***MobileNet*:** As *MobileNets* (SINHA; EL-SHARKAWY, 2019) são redes neurais convolucionais eficientes projetadas para dispositivos móveis e aplicações embarcadas. Estas redes são notáveis por serem pequenas, com baixa latência e baixo consumo de energia, características que permitem sua implantação em uma variedade de plataformas. Uma MobileNet típica tem aproximadamente 4,2 milhões de parâmetros e é capaz de extrair um vetor de até 1024 características, facilitando o uso em tarefas complexas de visão computacional apesar de seu tamanho reduzido.

Em todos os cenários mencionados, a premissa fundamental da transferência de aprendizado reside na abordagem de não iniciar o treinamento a partir do zero, mas sim utilizar como base os padrões que o modelo já adquiriu previamente. Esses padrões são posteriormente ajustados para se adequarem à nova tarefa em questão. Tal metodologia possibilita a aplicação do conhecimento pré-existente do modelo, o que frequentemente contribui para a obtenção de resultados aprimorados e mais ágeis (LIANG; FU; YI, 2019).

3 TRABALHOS RELACIONADOS

Estudos anteriores que visam a classificação de espécies de pássaros através de seus cantos e chamados empregam uma variada gama de técnicas de extração de características e classificação, cada um adotando abordagens específicas para alcançar seus objetivos e explorando metodologias diversas na classificação das pássaros, além de utilizarem bases de dados variadas.

Em (LUCIO; COSTA, 2016), é proposto uma abordagem para a classificação automatizada de especies de pássaros utilizando caracteristicas visuais e acusticas de sinais de audio. Esta abordagem utiliza um conjunto de dados com 2814 amostras de áudio divididas entre 46 espécies de pássaros, sendo ela um subconjunto da base de dados de cantos e chamados disponibilizada pelo site Xeno-Canto. A metodologia combina características acústicas, tais como o *Rhythm Histogram (RH)*, *Rhythm Patterns (RP)* e *Statistical Spectrum Descriptor (SSD)*, com características visuais, incluindo *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*. Para alcançar uma maior confiabilidade nos resultados, adotou-se o método de validação cruzada. Nesse método, cada um dos *folds* formados é alternadamente usado como conjunto de teste, enquanto os restantes servem para treinamento. Esse ciclo é continuado até que todos os *folds* tenham sido utilizados na função de teste. Por fim, os melhores resultados foram obtidos pela combinação do *SSD* com o *RLBP*, que apresentou as medidas de *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 classificação de quatro espécies de pássaros utilizando descritores de audio em gravações dos cantos desses pássaros. 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 do *SVM*. Os resultados mais promissores foram obtidos para as espécies de tordo-preto e corvo doméstico, com uma taxa de acerto 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 pássaros 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, sendo que, para esse trabalho, foi escolhida apenas 30 das 153 diferentes espécies de pássaros disponíveis. Os autores extraíram diversas características acústicas das gravações de áudio, incluindo a *Transformada de Fourier de Curto Prazo (STFT)*, *MFCC*, Energia RMS, Centróide Espectral, Largura de Banda Espectral, Roll-off Espectral e Taxa de Cruzamento de Zero. Estas caracte-

rísticas foram então utilizadas para alimentar vários modelos de aprendizado de máquina, como o *K-Nearest Neighbors (KNN)*, *SVM*, *Decision Trees*, *Random Forests*, *Niave Bayes* e *Stochastic Gradient Descent (SGD)* para classificar as espécies de pássaros. Os melhores resultados foram alcançados utilizando a *SGD*, quando treinada com um subconjunto de três espécies de pássaros, obtendo uma acurácia de 88% e um *F1-score* de 87%.

Um estudo conduzido por (INCZE *et al.*, 2018) é apresentado um sistema de classificação de sons de pássaros baseado em redes *Redes Neurais Convolucionais (CNN)*. 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 pássaros, 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 pássaros) e o esquema de cores dos espectrogramas (*grayscale* ou *jet*). 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 dez, 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%.

O artigo (LUCIO; MALDONADO; COSTA, 2015) apresenta um sistema para classificação de espécies de pássaros com base em descritores de textura. A base de dados é composta por um subconjunto da base de dados de cantos e chamados disponibilizada pelo site *Xeno-Canto*, com 2814 amostras de audio divididas entre 46 espécies distintas. Para a extração de características, foram utilizados três operadores de textura: *LBP*, *LPQ* e Filtros Gabor. O classificador *SVM* foi treinado com otimização de parametros e validação cruzada. A melhor acurácia obtida foi de cerca de 77,65% utilizando os Filtros Gabor.

Estes estudos ilustram o amplo espectro de abordagens e técnicas empregadas na classificação de pássaros com base em seus cantos ou chamado, revelando o progresso significativo alcançado na automação deste processo e destacando o potencial do uso de gravações de áudio e modelos de aprendizado de máquina nesse campo.

4 MATERIAIS E MÉTODOS

4.1 Materiais

A linguagem de programação *Python 3.11* foi selecionada para o desenvolvimento dos algoritmos necessários neste trabalho. Essa linguagem foi escolhida devido a sua facilidade, versatilidade e capacidade de atender as exigências do projeto, uma vez que possui toda uma estrutura de bibliotecas que auxilia o trabalho com cálculos numéricos, processamento de imagens e aprendizagem de máquina.

Diversas bibliotecas foram empregadas para suportar diferentes aspectos do projeto. A biblioteca *librosa* é utilizada para a geração de 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* ou Filtros Gabor.

Para a construção e avaliação de modelos de aprendizagem de máquina, o *scikit-learn* oferece múltiplos algoritmos e ferramentas para análise preditiva. Por fim, o *TensorFlow* é utilizado para implementar a extração de descritores via *transfer learning*, como o *VGG16*, *Resnet50*, *DenseNet121* e *MobileNet*, fornecendo modelos de redes neurais pré-treinadas.

4.1.1 Base de Dados

A base de dados adotada neste estudo é proveniente de (KLINCK *et al.*, 2023), constituída por gravações de cantos e chamados de aves de diversas partes do globo, cobrindo um amplo espectro de espécies em variados ecossistemas. Este repositório contém informações detalhadas de aproximadamente 16.900 gravações de áudio, abrangendo 264 espécies.

Destaca-se pela rica coleção de cantos e chamados de aves disponíveis para cada espécie incluída. Contudo, apesar de sua abrangência, a distribuição dos exemplos por classe e a duração das gravações variam, o que pode levar a um desbalanceamento, como demonstrado na Figura 2.

Além disso, um arquivo de metadados acompanha cada gravação, fornecendo detalhes como o nome da espécie, o nome do arquivo de áudio, o autor da gravação e informações geográficas, incluindo latitude e longitude do local de registro.

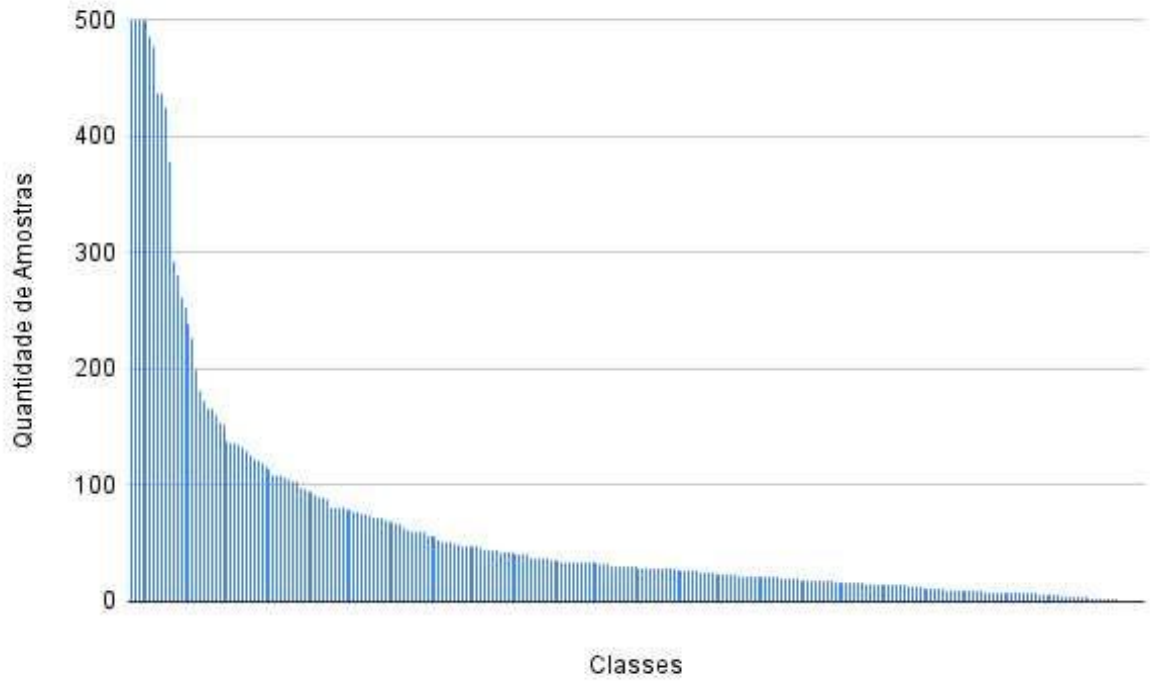


Figura 2 – Distribuição de amostras em cada classe

4.2 Métodos

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

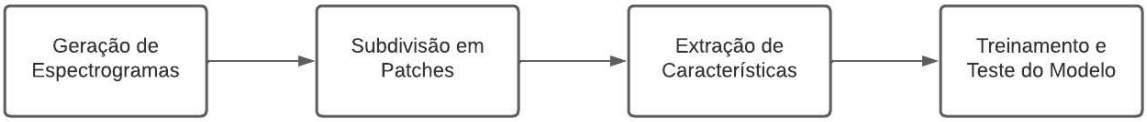


Figura 3 – Método proposto

4.2.1 Geração dos Espectrogramas

A criação de espectrogramas na escala Mel começa com a aplicação da Transformada de Fourier de Tempo Curto (STFT) ao sinal de áudio que se deseja analisar. Este cálculo utiliza blocos de 2048 amostras do sinal de áudio para realizar cada Transformada de Fourier individual que, juntas, compõem a STFT. Esse tamanho de bloco é escolhido para equilibrar a resolução temporal e de frequência do espectrograma resultante. A fim de capturar detalhes mais refinados das variações de amplitude ao longo do tempo, implementa-se uma técnica de sobreposição de 50% entre as janelas de amostras consecutivas. Este método permite que a

análise capture melhor as transições entre sons ao longo do tempo, aumentando a continuidade e a qualidade da informação de frequência.

Após a obtenção do espectrograma através da STFT, o processo continua com a aplicação de um banco de filtros ponderados pela escala Mel sobre o espectrograma. Esses filtros têm a função de mapear as frequências originais do espectrograma para uma escala que imita a percepção auditiva humana, a escala Mel. Essa escala é logarítmica e foi projetada para refletir como o ouvido humano percebe as diferenças de frequência, dando mais ênfase às variações nas frequências mais baixas e menos às altas frequências, de modo que o espectrograma final ressoe mais intimamente com a forma como nós percebemos o som. Este refinamento visa não apenas a análise técnica, mas também aplicações práticas que requerem uma representação mais fiel da audição humana, como o reconhecimento de voz e a música.

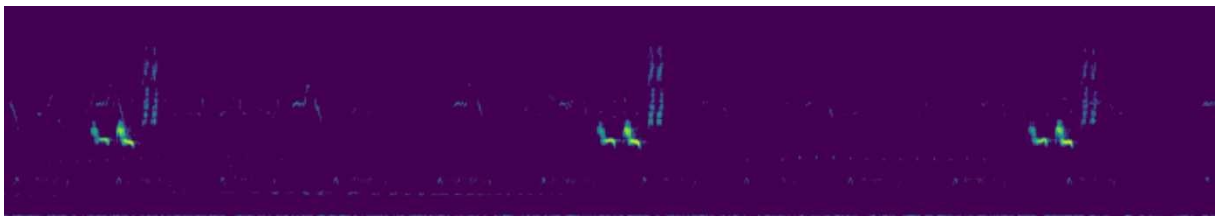


Figura 4 – Espectrograma em escala Mel da faixa de audio XC128013.ogg

4.2.2 Subdivisão em Patches

Nesta fase, o principal objetivo é processar os espectrogramas segmentando-os em N patches menores na dimensão do tempo. É crucial notar que o tamanho dos espectrogramas varia de acordo com o comprimento dos áudios, resultando em uma característica intrinsecamente não uniforme devido à natureza variada dos sons capturados em ambientes naturais, que ocorrem em momentos distintos.

Para abordar esta variação sonora ao longo dos áudios, é essencial extrair características de diferentes janelas temporais para maximizar a informação obtida. Essa técnica, ilustrada na Figura 5, permite uma análise da ampla gama de sons em diferentes momentos das gravações.

4.2.3 Extração de Caracteristicas

Nesse momento é importante definir metodicamente as espécies de pássaros selecionadas para a extração de características, visando o desenvolvimento do modelo de classificação. A precisão na escolha dessas espécies é crucial para assegurar a relevância e a eficácia do modelo em cenários reais.

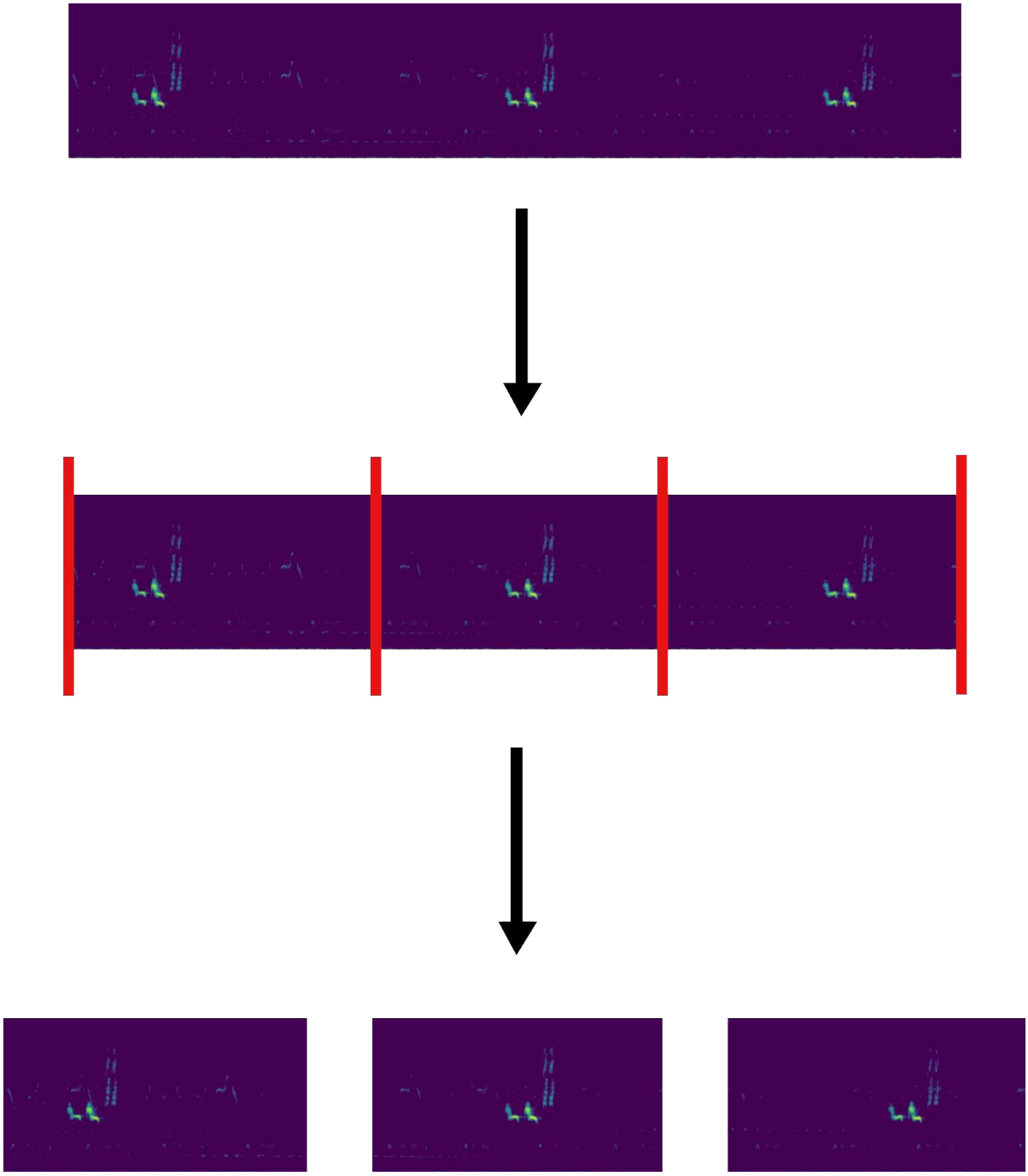


Figura 5 – Geração de *patches* com 300 colunas da faixa de audio XC128013.ogg

4.2.3.1 Subdivisão da Base de Dados

Nesta etapa do trabalho, o propósito central foi identificar as K espécies mais próximas a um ponto de referência específico, localizado em um determinado espaço geográfico destinado à análise. Para alcançar este objetivo, a base de dados foi dividida em subgrupos. Cada subgrupo é constituído por um número restrito de espécies, selecionadas com base em sua proximidade geográfica. Esta estratégia visa refletir, de maneira mais fidedigna, as condições encontradas em ambientes naturais reais. Tal abordagem se mostra fundamental, considerando

que pesquisas de campo em áreas geográficas específicas frequentemente revelam uma coexistência de um número limitado de espécies em uma mesma localidade. Por consequência, a construção de um modelo que tente classificar um grande número de espécies globalmente distribuídas seria impraticável e distante da realidade prática. Portanto, a escolha cuidadosa das espécies a serem analisadas surge como um pilar crucial para o enriquecimento da precisão e da aplicabilidade dos resultados em contextos reais.

A fim de determinar quais são as K espécies mais próximas ao ponto de referência designado, adotou-se a técnica de *KDE*. O *KDE* é um método não paramétrico que permite estimar a função de densidade de probabilidade de um conjunto de dados. Essa ferramenta é útil para identificar regiões com maior agregação de pontos (Figura 6a), o que facilita a determinação do local de maior densidade dentro de um conjunto de coordenadas geográficas.

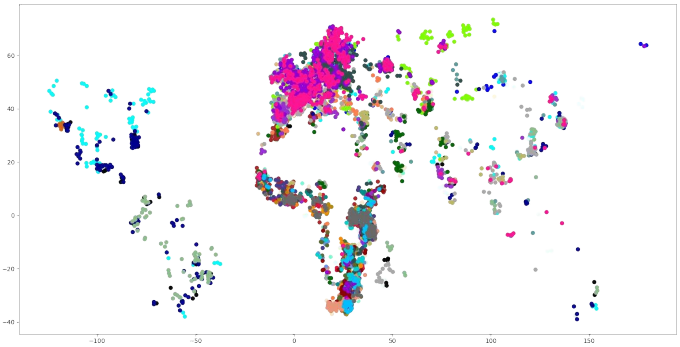
No contexto deste estudo, focado em espécies de pássaros, utilizou-se o *KDE* para calcular a coordenada de maior densidade de cada espécie (Figura 6b) com base na base de dados, que inclui todas as coordenadas das amostras disponíveis. É importante destacar que um filtro foi aplicado para excluir quaisquer espécies com menos de 100 amostras, visando assegurar a confiabilidade do processo de validação cruzada na fase de treinamento do modelo.

Após a determinação das coordenadas de maior densidade para cada espécie, o *KDE* foi novamente aplicado para identificar um ponto de referência no mapa (Figura 6c), caracterizado pela sua significativa densidade de espécies. Esse ponto serve como foco para a análise subsequente.

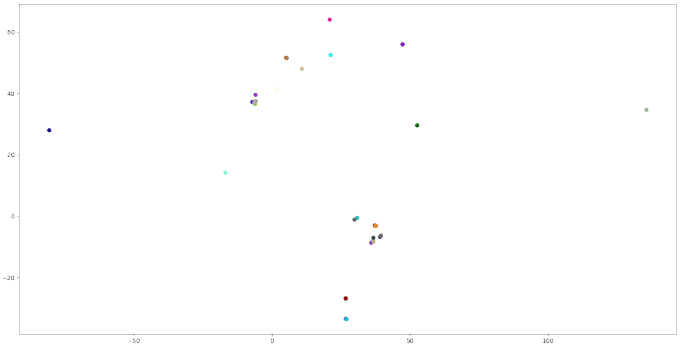
Por fim, calculou-se a distância entre o ponto de referência identificado e as coordenadas de maior densidade de cada espécie. A ordenação dessas distâncias, do menor para o maior valor, facilitou a compilação de uma lista contendo as K espécies geograficamente mais próximas ao ponto de referência. Para este trabalho, o algoritmo foi aplicado para as duas, três, cinco e dez espécies mais próximas desse ponto de referência. A distribuição das espécies nesses subconjuntos é exibida na Tabela 1, enquanto os detalhes dessas espécies podem ser encontrados na Tabela 2.

Nome Cientifico	2 Classes	3 Classes	5 Classes	10 Classes
Hirundo rustica	-	-	X	X
Milvus migrans	X	X	X	X
Bubulcus ibis	-	-	X	X
Delichon urbicum	-	-	-	X
Motacilla flava	-	-	-	X
Merops apiaster	-	X	X	X
Spatula querquedula	-	-	-	X
Upupa epops	-	-	-	X
Egretta garzetta	X	X	-	X
Cecropis daurica	-	-	-	X

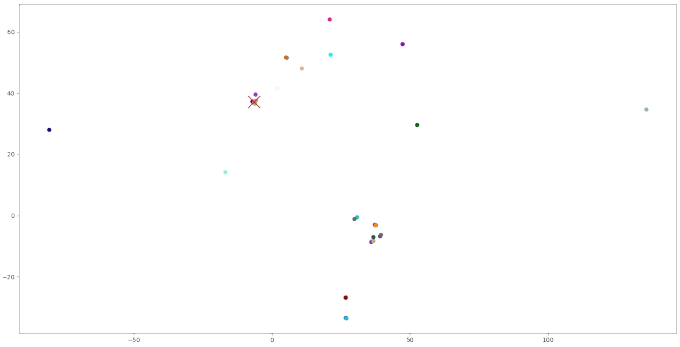
Tabela 1 – Distribuição das dez Espécies mais próximas em cada subconjunto de classe



(a) Distribuição das Amostras



(b) Pontos de Referência de Cada Espécie



(c) Ponto de Referência em análise

Figura 6 – Processo de estimação do *KDE*

4.2.3.2 Processamento dos *Patches*

Após a seleção das K classes de interesse, procede-se à etapa de extração de características. Esta fase envolve o processamento dos *patches* associados a cada classe escolhida. Para isso, os patches são analisados e processados utilizando diversos descritores de características. Esses descritores são divididos em duas categorias principais: os descritores manuais

Nome Cientifico	Nome Comum	Localização	Nº amostras
Hirundo rustica	Barn Swallow	Sul da Espanha	500
Milvus migrans	Black Kite	Sul da Espanha	262
Bubulcus ibis	Cattle Egret	Sul da Espanha	166
Delichon urbicum	Common House-Martin	Sul da Espanha	425
Motacilla flava	Western Yellow Wagtail	Sul da Espanha	500
Merops apiaster	European Bee-eater	Sul da Espanha	437
Spatula querquedula	Garganey	Nordeste da Espanha	136
Upupa epops	Eurasian Hoopoe	Sul da Espanha	436
Egretta garzetta	Little Egret	Sul da Espanha	378
Cecropis daurica	Red-rumped Swallow	Centro-oeste da Espanha	227

Tabela 2 – Detalhes das dez Classes mais próximas

(*handcrafted*), como o *LBP* e os Filtros Gabor, e os descritores baseados em técnicas de aprendizado por transferência, incluindo modelos de *deep learning* como *VGG16*, *Resnet50*, *Densenet121* e *Mobilenet*. A escolha desses descritores é fundamentada em estudos anteriores que exploraram o uso isolado de descritores manuais ou aplicaram estratégias de aprendizado por transferência. A integração dessas duas abordagens visa não apenas uma análise comparativa detalhada, mas também uma avaliação abrangente do desempenho e eficácia dos resultados alcançados com a aplicação dessas tecnologias.

Para cada descritor utilizado, os parâmetros utilizados foram:

- *LBP*: Utilizado para descrever texturas e padrões locais, é configurado com um número de pontos p=8 e um raio r=2, usando o método nri_uniform para promover invariância a rotação.

- Filtros Gabor: Projetados para capturar as frequências e orientações específicas dentro das imagens, operam com frequências definidas em [0.1, 0.3, 0.5] e ângulos [0, π/4, π/2, 3π/4], visando uma análise detalhada das características espaciais.

- *VGG16*, *Resnet50*, *Densenet121*, e *Mobilenet*: Todos esses modelos são configurados com pesos pré-treinados no conjunto de dados *imagenet*, não incluem a camada superior (include_top=False), utilizam uma operação de *pooling* médio (pooling='avg') e esperam imagens de entrada no formato 224×224×3. Esses parâmetros permitem que os modelos concentrem-se na extração de características genéricas das imagens, que são posteriormente utilizadas em problemas de classificação.

A partir dessas configurações, para cada descritor, é gerado um *dataframe* contendo as características extraídas de cada *patch*. Este *dataframe* inclui também informações relevantes como a classe a que o *patch* pertence e o arquivo de áudio correspondente. Com o intuito de melhorar a precisão do classificador e facilitar a identificação individual das amostras, dados de localização geográfica, especificamente latitude e longitude de cada gravação, são adicionados ao *dataframe*. Essa abordagem multidimensional não só enriquece o conjunto de dados, como também promove uma análise mais detalhada e contextualizada das amostras.

4.2.4 Treinamento

Nesta fase do trabalho, é realizado o treinamento e avaliação do modelo de *machine learning* para determinar sua eficiência. O processo inicia-se com a preparação do conjunto de dados, composto por *patches* extraídos de arquivos de áudio. Cada *patch* contém as características extraídas e está vinculado a um arquivo de áudio particular. A etapa inicial envolve a seleção dos arquivos de áudio que serão designados para os conjuntos de treinamento ou teste. Utiliza-se a estratégia de validação cruzada para essa distribuição, na qual os arquivos de áudio são divididos em dez grupos distintos para treino e teste, garantindo uma distribuição proporcional e equitativa.

Para cada subdivisão (ou *split*), os arquivos de áudio são categorizados como de treinamento ou teste, e seus *patches* correspondentes no *dataframe* são identificados e separados de acordo com essa classificação. Tal procedimento assegura que, se um arquivo é selecionado para treinamento em uma divisão específica, todos os seus *patches* são exclusivamente usados para treinamento, evitando o risco de que o modelo simplesmente reconheça os *patches* de teste devido à sua semelhança com os de treino, o que não seria indicativo de uma boa generalização. Seguidamente, os dados são normalizados utilizando o algoritmo *StandardScaler* para garantir a uniformidade na escala das variáveis. Esse método de normalização subtrai a média e divide pelo desvio padrão de cada variável, assegurando que os dados estejam centrados em zero e com variância unitária.

O próximo passo consiste no treinamento do modelo, utilizando o SVM acompanhado de uma busca exaustiva para a otimização dos hiperparâmetros. Este processo tem como objetivo refinar o modelo com a configuração mais eficaz de hiperparâmetros. Durante esta etapa, foram avaliados diferentes combinações de hiperparâmetros cujos valores são especificados na tabela 3. Posteriormente, o modelo ajustado é avaliado utilizando os dados de teste.

Hiperparâmetros	Valores	Contagem
C	0.01, 0.1, 1, 10	4
Kernel	linear, poly, rbf, sigmoid	4
Gamma	0.0625, 0.125, 0.25, 0.5, 1, 2, scale, auto	8
Total de Variações		128

Tabela 3 – Combinação dos valores avaliados para os hiperparâmetros do SVM

Uma característica importante deste modelo é a classificação baseada em votação. Como cada patch de um mesmo áudio é previamente classificado em uma categoria, realiza-se uma agregação dos resultados por arquivo de áudio, determinando a classe predominante para cada um com base na maioria das predições dos patches correspondentes. Ao finalizar a classificação dos arquivos de áudio, calcula-se a pontuação F1 para cada conjunto de teste.

Este procedimento é repetido para cada uma das dez divisões da validação cruzada. A eficácia do modelo é então avaliada pela média das pontuações F1 alcançadas em todos os testes, fornecendo uma estimativa de desempenho do modelo.

5 RESULTADOS E DISCUSSÃO

Neste trabalho foram empregados os descritores *LBP* e filtros Gabor, classificados como descritores manuais (*handcrafted*), além dos descritores *VGG16*, *ResNet50*, *DenseNet121* e *MobileNet*, que são derivados de técnicas de Transferência de Aprendizagem (*Transfer Learning*). A metodologia de análise consistiu na extração de características a partir de um conjunto de um, três e cinco *patches* (segmentos) de cada imagem. Para cada uma dessas configurações de *patches*, as características foram extraídas de um subconjunto contendo duas, três, cinco e dez categorias diferentes de pássaros.

Os resultados dos experimentos são sistematicamente apresentados na Figura 7. A Figura 7a apresenta os valores da métrica F1 obtidos na extração de características utilizando um único *patch*, equivalente ao espectrograma completo. Já a Figura 7b oferece uma visão detalhada dos resultados alcançados ao dividir os espectrogramas em três *patches*, enquanto a Figura 7c aborda os resultados derivados do uso de cinco *patches*. O objetivo principal dessas tabelas é demonstrar como a taxa de acerto de cada descritor varia ao incrementar o número de classes analisadas, para cada configuração de *patch*, permitindo identificar quais descritores apresentam as melhores taxas de acerto e entender seu comportamento conforme se expande o número de classes.

Conforme ilustrado na Figura 7a, observou-se que, para um único *patch*, as redes neurais convolucionais exibiram desempenho superior, destacando-se a *DenseNet121*, que alcançou as maiores pontuações de F1 em quase todos os subconjuntos de classes. Por exemplo, a *DenseNet121* registrou um F1 de 0,8992 para duas classes e de 0,5580 para dez classes. Em seguida, a *ResNet50* e a *MobileNet* mostraram resultados comparáveis, com F1 de 0,8976 e 0,8888 para duas classes, respectivamente, e de 0,5202 e 0,5246 para dez classes. Em contraste, os piores desempenhos para um único *patch* foram consistentemente obtidos com os Filtros Gabor, particularmente notáveis no subconjunto de dez classes, onde o F1 foi apenas 0,4182.

Conforme apresentado na Figura 7b, o aumento no número de *patches* de 1 para 3 resultou em uma melhoria geral nas pontuações F1 para a maioria dos descritores e redes, demonstrando a eficácia de analisar os espectrogramas em múltiplas seções. Esse incremento foi particularmente notável nos modelos de redes neurais, como o *VGG16*, cuja pontuação F1 aumentou de 0,8580 para 0,8764 em duas classes. Além disso, destacaram-se a *ResNet50* e a *MobileNet*, que alcançaram pontuações muito similares de 0,8994 e 0,8989 para duas classes, respectivamente, e de 0,5408 e 0,5760 para dez classes.

A transição de 3 para 5 *patches* não seguiu uma tendência de melhoria uniforme, como apresentado na Figura 7c; em alguns casos, observou-se uma pequena piora. Por exemplo, a pontuação F1 do *MobileNet* para duas classes diminuiu ligeiramente de 0,8989 para 0,8875. Isso sugere que, embora dividir os espectrogramas em múltiplos *patches* possa inicialmente beneficiar a classificação ao destacar características únicas em diferentes seções do espectro-

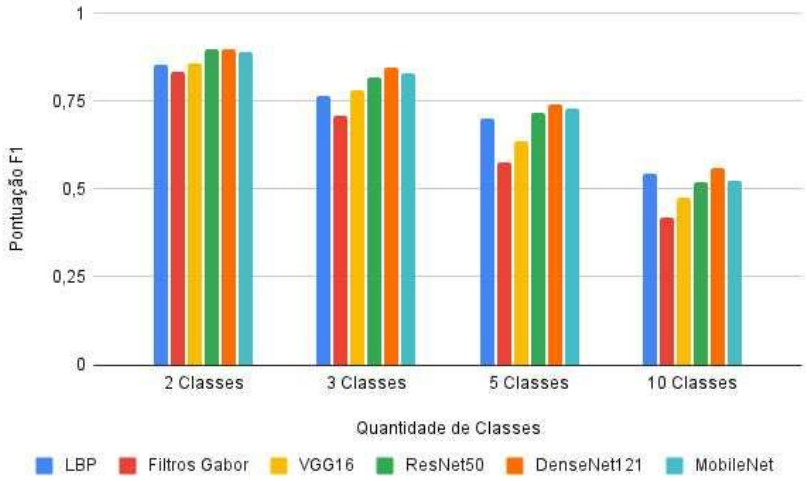
grama, pode haver um limite a partir do qual essa estratégia deixa de ser eficaz. Esse declínio pode ocorrer devido ao aumento de redundância, onde informações semelhantes são capturadas repetidamente em vários *patches*, ou pelo acréscimo de complexidade no processo de treinamento, tornando mais difícil para o modelo discernir padrões úteis entre um volume maior de dados parcialmente sobrepostos.

Esses resultados sublinham a importância de calibrar cuidadosamente o número de *patches* na análise de espectrogramas para classificação de canto de pássaros. A subdivisão em três *patches* parece ser o ponto ótimo, equilibrando o detalhamento da análise com a eficácia da classificação, especialmente quando se utilizam redes neurais avançadas como *ResNet50* e *DenseNet121*, que consistentemente superaram outras abordagens em termos de precisão, independentemente do número de classes. Assim, embora a divisão em *patches* represente um passo positivo em direção a uma análise mais granular e potencialmente mais informativa dos dados, a moderação é chave; incrementos além de um certo ponto podem não apenas deixar de oferecer melhorias significativas, mas também podem complicar desnecessariamente o modelo sem ganhos claros em desempenho.

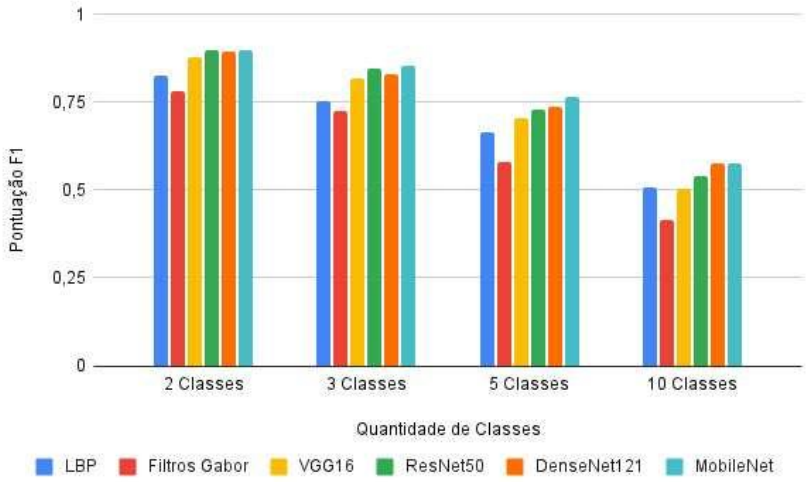
Analisando a partir de outro ponto de vista, a Figura 8 destina-se a mostrar como a taxa de acerto varia para cada configuração de *patch* com o aumento do número de classes. Esta tabela visa destacar quais configurações de *patches* proporcionam as maiores taxas de acerto, além de examinar as tendências de desempenho dos descritores à medida que se aumenta o número de classes. Esse nível de detalhamento é crucial para compreender integralmente as capacidades e limitações dos descritores utilizados, bem como para orientar a escolha dos métodos de extração de características mais eficazes para futuras pesquisas no campo de classificação de imagens de pássaros.

Analisando os resultados obtidos, observa-se uma tendência: à medida que aumenta o número de classes, a pontuação F1 de todos os descritores diminui. Esta diminuição segue um padrão similar tanto para descritores de redes neurais convolucionais, como *DenseNet121*, *ResNet50* e *MobileNet*, quanto para descritores tradicionais como *LBP* e Filtros Gabor. Apesar das diferenças iniciais nas pontuações F1 entre os descritores avançados e tradicionais, a tendência de queda à medida que o número de classes aumenta é consistentemente observada em todos eles. Essa observação sugere que, independentemente da tecnologia ou método empregado, a complexidade adicional introduzida pelo aumento no número de classes afeta de maneira similar a performance de todos os descritores.

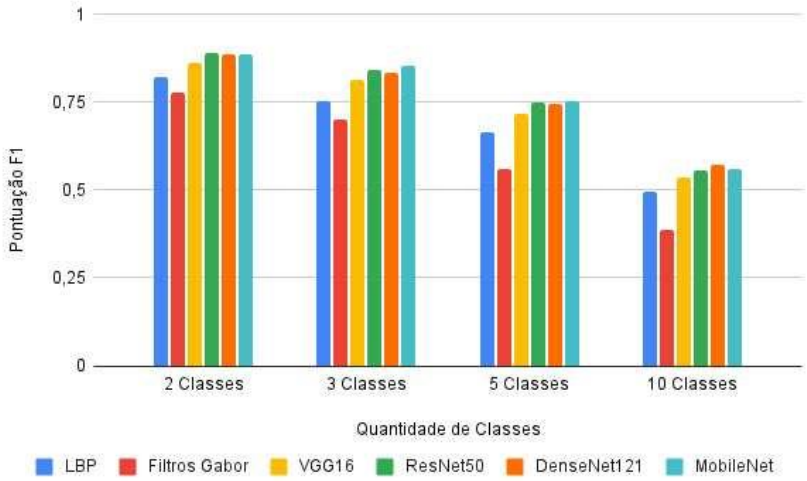
Dos descritores manuais, os Filtros Gabor apresentam uma redução mais pronunciada no desempenho conforme cresce o número de classes. O declínio é particularmente acentuado nos Filtros Gabor, que caem de um F1 de 0,8342 com duas classes para apenas 0,4182 com dez classes. Isso sugere uma menor capacidade de generalização desse método em contextos de maior diversidade de categorias. Essa análise ressalta a eficácia superior das redes neurais convolucionais, que não só mantêm desempenhos mais elevados, mas também mostram maior estabilidade na performance ao lidar com conjuntos de dados mais complexos e variados.



(a) 1 Patch

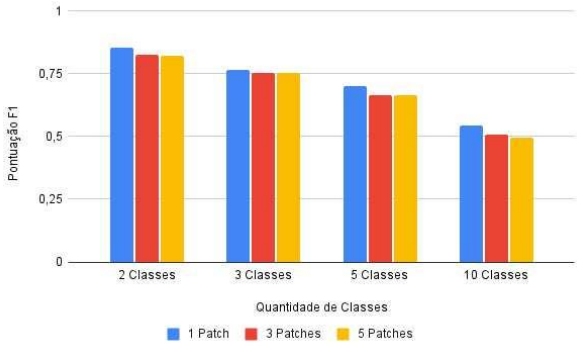


(b) 3 Patches

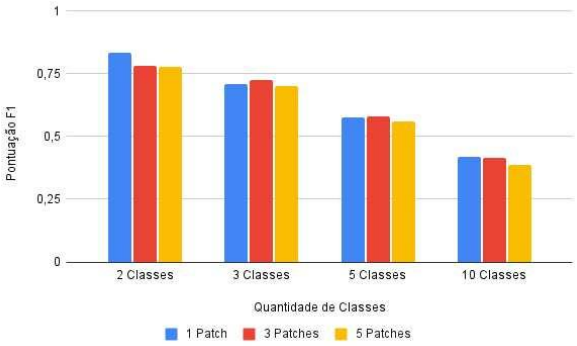


(c) 5 Patches

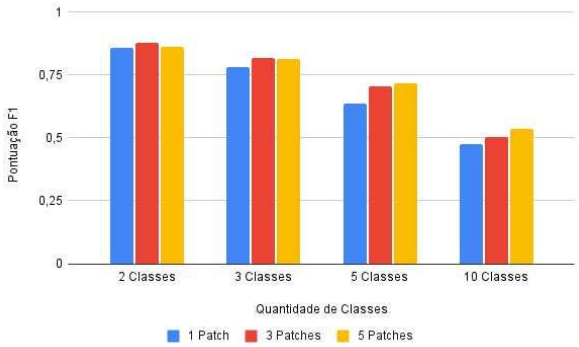
Figura 7 – Pontuação F1 em relação a quantidadede classes para cada patch



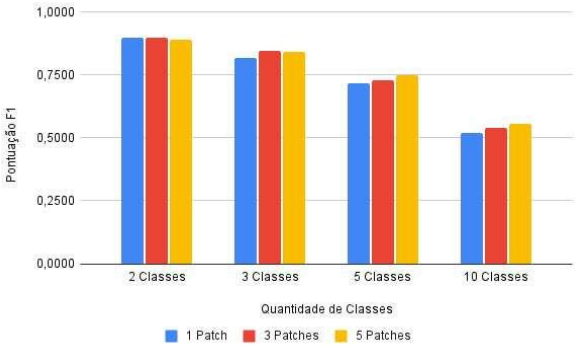
(a) LBP



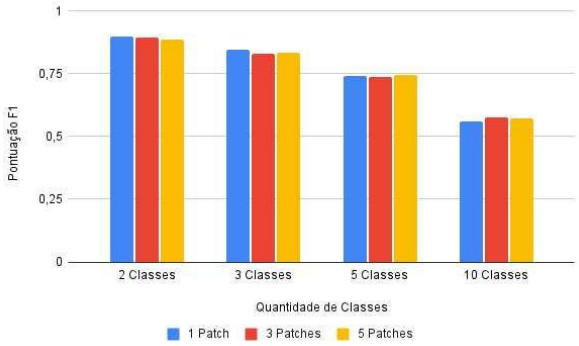
(b) Filtros Gabor



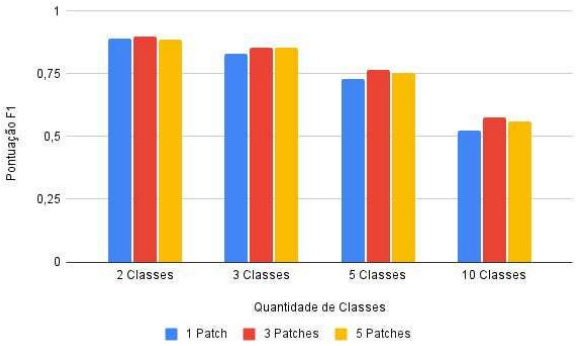
(c) VGG16



(d) ResNet50



(e) DenseNet121



(f) MobileNet

Figura 8 – Pontuação F1 em relação a quantidade de classes para cada descritor

6 CONCLUSÃO

A análise dos resultados obtidos na classificação de canto de pássaros, com base na subdivisão dos espectrogramas em *patches* e na aplicação de diferentes descritores, revela tendências e sugere direcionamentos estratégicos para futuras pesquisas nessa área. A performance dos modelos variou significativamente conforme a quantidade de classes e a metodologia empregada, evidenciando que tanto a natureza do descritor quanto a complexidade da tarefa de classificação em relação ao número de classes influenciam nos resultados.

Os resultados deste estudo demonstraram que o número de *patches* por áudio exerce uma influência significativa no desempenho dos classificadores utilizados na identificação do canto de pássaros. De acordo com os experimentos conduzidos, o uso de três *patches* por áudio mostrou-se mais eficaz na obtenção de melhores resultados de classificação para a maioria dos descritores analisados. Além disso, esta configuração de três *patches* parece apresentar um equilíbrio ideal entre o custo computacional para treinar o classificador e o desempenho obtido.

As redes neurais convolucionais, especialmente *ResNet50*, *MobileNet* e *DenseNet121*, exibiram desempenhos superiores em comparação com os descritores manuais, como *LBP* e filtros Gabor, especialmente quando o número de classes aumenta. Apesar o *LBP* seguir um padrão similar de diminuição da pontuação conforme se aumenta o número de classes, os Filtros Gabor obtiveram um declínio muito maior, obtendo os piores resultados. Este fenômeno ressalta a robustez das técnicas de transferência de aprendizado em contextos de maior complexidade e diversidade de dados. Assim, os resultados obtidos alinham-se aos objetivos iniciais de explorar e comparar diferentes metodologias de classificação e extração de características, fornecendo *insights* sobre a aplicabilidade das técnicas investigadas no contexto de classificação de pássaros por meio de gravações de seus cantos e chamados.

Com base nos resultados e conclusões apresentados neste estudo, delineiam-se várias direções promissoras para pesquisas futuras. Inicialmente, recomenda-se aprimorar o processamento dos áudios e dos espectrogramas. Isso pode incluir a aplicação de filtros para melhoria da qualidade do áudio, como a redução de ruídos, e técnicas de processamento de imagens que intensifiquem as características distintivas dos espectrogramas, como o uso de técnicas de realce. Outro aspecto relevante seria explorar a combinação de descritores manuais com redes neurais convolucionais, visando a uma sinergia entre os métodos que potencialize a eficácia da classificação. Por último, seria proveitoso investigar modelos que incorporem, além das características extraídas dos espectrogramas, dados geográficos como latitude e longitude, para avaliar como essas informações influenciam os resultados da classificação.

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.

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**. São Paulo, SP: 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)**. Chickballapur, India: IEEE Computer Society Press., 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)**. Subotica, Serbia: IEEE Computer Society Press., 2018. p. 000295–000300.

KLINCK, H. *et al.* **BirdCLEF 2023**. Kaggle, 2023. Disponível em: https://kaggle.com/competitions/birdclef-2023.

LIANG, H.; FU, W.; YI, F. A survey of recent advances in transfer learning. *In*: **2019 IEEE 19th International Conference on Communication Technology (ICCT)**. Xi'an, China: IEEE Computer Society Press., 2019. p. 1516–1523.

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)**. Valparaiso, Chile: IEEE Computer Society Press., 2016. p. 1–12.

LUCIO, D. R.; MALDONADO, Y.; COSTA, G. da. Bird species classification using spectrograms. *In*: **2015 Latin American Computing Conference (CLEI)**. Arequipa, Peru: IEEE Computer Society Press., 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.

OJALA, T.; PIETIKAINEN, M.; MAENPAA, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 7, p. 971–987, 2002.

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)**. Coimbatore, India: IEEE Computer Society Press., 2016. v. 1, p. 1–5.

SINHA, D.; EL-SHARKAWY, M. Thin mobilenet: An enhanced mobilenet architecture. *In*: **2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication**

Conference (UEMCON). New York, NY, USA: IEEE Computer Society Press., 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)**. Macao, SAR of China and online: IEEE Computer Society Press., 2021. p. 1–3.

TURING, A. M. I.—COMPUTING MACHINERY AND INTELLIGENCE. **Mind**, LIX, n. 236, p. 433–460, 10 1950. ISSN 0026-4423. Disponível em: https://doi.org/10.1093/mind/LIX.236.433.

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)**. Athens, Greece: IEEE Computer Society Press., 2023. p. 58–62.

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