



UNIVERSIDADE FEDERAL DO AMAZONAS - UFAM  
INSTITUTO DE COMPUTAÇÃO- ICOMP  
PROGRAMA PÓS-GRADUAÇÃO EM INFORMÁTICA - PPGI

**Explorando variações do canto do Corruíra sob ruído  
antropogênico: Uma abordagem orientada à inteligência  
artificial explicável para análise e interpretação  
bioacústica**

Larissa de Andrade Silva

Manaus - AM

Julho 2024

Larissa de Andrade Silva

Explorando variações do canto do Corruíra sob ruído  
antropogênico: Uma abordagem orientada à inteligência  
artificial explicável para análise e interpretação  
bioacústica

Dissertação submetido à avaliação, como requisito,  
para a obtenção do título de Mestre em Informática  
no Programa de Pós-Graduação em Informática,  
Instituto de Computação.

Orientador(a)

Juan Gabriel Colonna, Dr.

Manaus - AM

Julho 2024

## Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

S586e	<p>Silva, Larissa de Andrade Explorando variações do canto do Corruíra sob ruído antropogênico: Uma abordagem orientada à inteligência artificial explicável para análise e interpretação bioacústica / Larissa de Andrade Silva . 2024 88 f.: il. color; 31 cm.</p> <p>Orientador: Juan Gabriel Colonna Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.</p> <p>1. Corruíra. 2. Bioacústica. 3. Inteligência artificial explicável. 4. Aprendizagem Profunda. I. Colonna, Juan Gabriel. II. Universidade Federal do Amazonas III. Título</p>
-------	--



Ministério da Educação  
Universidade Federal do Amazonas  
Coordenação do Programa de Pós-Graduação em Informática

## FOLHA DE APROVAÇÃO

### "EXPLORANDO VARIAÇÕES DO CANTO DO CURRUÍRA SOB RUÍDO ANTROPOGÊNICO: UMA ABORDAGEM ORIENTADA POR XAI PARA ANÁLISE E INTERPRETAÇÃO BIOACÚSTICA"

**LARISSA DE ANDRADE SILVA**

### **DISSERTAÇÃO DE MESTRADO DEFENDIDA E APROVADA PELA BANCA EXAMINADORA CONSTITUÍDA PELOS PROFESSORES:**

Prof. Dr. Juan Gabriel Colonna - PRESIDENTE

Prof. Dr. Eduardo Freire Nakamura - MEMBRO INTERNO

Dr. Bernardo Bentes Gatto - MEMBRO EXTERNO

MANAUS, 24 de julho de 2024.



Documento assinado eletronicamente por **Juan Gabriel Colonna, Professor do Magistério Superior**, em 26/07/2024, às 11:14, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Eduardo Freire Nakamura, Professor do Magistério Superior**, em 30/07/2024, às 16:03, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Bernardo Bentes Gatto, Usuário Externo**, em 31/07/2024, às 10:51, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



Documento assinado eletronicamente por **Maria do Perpétuo Socorro Vasconcelos Palheta, Secretária em exercício**, em 06/08/2024, às 12:30, conforme horário oficial de Manaus, com fundamento no art. 6º, § 1º, do [Decreto nº 8.539, de 8 de outubro de 2015](#).



A autenticidade deste documento pode ser conferida no site  
[https://sei.ufam.edu.br/sei/controlador\\_externo.php?  
acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](https://sei.ufam.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **2148745** e o código CRC **55AC821B**.

Avenida General Rodrigo Octávio, 6200 - Bairro Coroado I Campus Universitário  
Senador Arthur Virgílio Filho, Setor Norte - Telefone: (92) 3305-1181 / Ramal 1193  
CEP 69080-900, Manaus/AM, coordenadorppgi@icomp.ufam.edu.br

Referência: Processo nº 23105.031435/2024-17

SEI nº 2148745

---

## AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD.

*Antes um trabalho feito, do que perfeito.*

Colonna, Juan G., 2024

# Explorando variações do canto do Corruíra sob ruído antropogênico: Uma abordagem orientada à inteligência artificial explicável para análise e interpretação bioacústica

Autor: Larissa de Andrade Silva

Orientador: Juan Gabriel Colonna, Dr.

## Resumo

Este trabalho descreve um método automatizado de IA explicável (XAI) para identificar variações no canto da Curruíra (*Troglodytes aedon*) em resposta ao ruído antropogênico. Propomos um modelo End-to-End (E2E) usando Biophony e Kapre para classificar as vocalizações dessa ave. Para fornecer explicações sobre as classificações, aplicamos o método SHAP (SHapley Additive exPlanations). O modelo E2E separa com sucesso os espécimes de Curruíra que foram expostos a ruído antropogênico, os quais alteram seus sinais acústicos durante o canto, daqueles que não foram afetados. Ao analisar os padrões de frequência temporal enfatizados pelo modelo, o SHAP revela o papel crítico das mudanças relacionadas às frequências do som no processo de tomada de decisão do classificador. Esta descoberta corrobora pesquisas recentes que sugerem que as alterações no canto desta espécie em áreas urbanas afetadas por ruído antropogênico são predominantemente caracterizadas por mudanças de frequência e amplitude. No entanto, e diferente dos trabalhos relacionados, com nossa metodologia conseguimos explicar mudanças sutis no canto de cada indivíduo. Além disso, este trabalho apresenta novas visualizações de valores SHAP no contexto da bioacústica: o espectro SHAP, o espectrograma SHAP e padrões de sinais temporais SHAP. Esta abordagem permite

uma interpretação objetiva e explicável das escolhas do modelo, beneficiando biólogos e ecologistas em suas análises e interpretações bioacústicas.

*Palavras-chave:* Bioacústica, IA Explicável, Curruíra, Aprendizado Profundo.

# Explorando variações do canto do Corruíra sob ruído antropogênico: Uma abordagem orientada à inteligência artificial explicável para análise e interpretação bioacústica

Autor: Larissa de Andrade Silva

Orientador: Juan Gabriel Colonna, Dr.

## Abstract

This work describes an automated explainable AI (XAI) method to identify variations in the song of the House Wren (*Troglodytes aedon*) in response to anthropogenic noise. We propose an End-to-End (E2E) model using Biophony and Kapre to classify the vocalizations of this bird. To provide explanations for the classifications, we applied the SHAP (SHapley Additive exPlanations) method. The E2E model successfully separates House Wren specimens that were exposed to anthropogenic noise, which alter their acoustic signals during singing, from those that were not affected. By analyzing the temporal frequency patterns emphasized by the model, SHAP reveals the critical role of frequency changes in the sound in the classifier's decision-making process. This finding corroborates recent research suggesting that alterations in the song of this species in urban areas affected by anthropogenic noise are predominantly characterized by changes in frequency and amplitude. However, unlike related works, with our methodology, we were able to explain subtle changes in the song of each individual. Additionally, this work presents new visualizations of SHAP values in the context of bioacoustics: the SHAP spectrum, the SHAP spectrogram, and SHAP temporal signal patterns. This approach allows for an objective and explainable interpretation of the

model's choices, benefiting biologists and ecologists in their bioacoustic analyses and interpretations.

*Keywords:* Bioacoustics, XAI, House Wren, Deep Learning.

---

# LISTA DE ILUSTRAÇÕES

Figura 1 – (a) Individuo Corruíra macho (imagem de copyright livre) (b) Espectrograma do canto do Corruíra mostrando o trinado e as medições de frequência mínima e máxima utilizadas para análise. (c) Espectro de potência usado para definir a frequência máxima e mínima no canto. Linhas tracejadas mostram os limites de frequência máxima e mínima no espectrograma e no espectro de potência por SANDOVAL; REDONDO; BARRANTES . . . . .	24
Figura 2 – Etapas do processamento sonoro bioacústico. . . . .	27
Figura 3 – Banco de Filtros Mel. . . . .	30
Figura 4 – Estrutura básica de uma CNN (LECUN et al., 1998) . . . . .	31
Figura 5 – Exemplo de convolução de uma matriz 3x3 e um <i>kernel</i> 2x2 . . . . .	32
Figura 6 – Arquitetura do Biophony . . . . .	39
Figura 7 – Matriz de confusão de um problema de classificação binário . . . . .	44
Figura 8 – Processamento das amostras de áudio do banco de dados "curruíra".	55
Figura 9 – Modelo E2E composto por camadas Kapre e o modelo Biophony. Os valores SHAP são obtidos a partir do espectrograma, antes de convertê-lo para a escala Mel. . . . .	56
Figura 10 – Um trecho de código que retorna a camada de espectrograma Mel .	57
Figura 11 – Um trecho de código com a) KAPRE e b) Biophony . . . . .	57

Figura 12 – Amostra de vocalização de corruíra e os valores SHAP para a amostra.	60
Esta representação mostra a distribuição dos valores SHAP em dois quadros ao lado do espectrograma; a intensidade da cor é definida pelos valores de contribuição, quanto mais intenso o vermelho, maior a contribuição positiva, e quanto mais intenso o azul, maior a contribuição negativa; na parte de baixo da figura, uma régua revelando a escala de valores de contribuição para essa amostra . . . . .	60
Figura 13 – a) Matriz de confusão e b) Curvas de aprendizagem: Acurácia; c) Loss	61
Figura 14 – Interpretação SHAP para a) Canto do corruíra em ambiente não-impactado. b) Canto do corruíra em ambiente impactado. . . . .	63
Figura 15 – SHAP espectrograma a) de uma amostra de vocalização de uma região impactada (TP); b) de uma amostra de vocalização de uma região não impactada (TN); c) de uma amostra de vocalização de uma região não-impactada, mas classificada como impactada (FP); d) Uma amostra de vocalização de uma região não impactada, mas classificada como impactada (FN) . . . . .	65
Figura 16 – SHAP Spectrum e SHAP time de (a) uma amostra de vocalização de uma região impactada (TP); (b) uma amostra de vocalização de uma região não impactada (TN) . . . . .	67
Figura 17 – SHAP Spectrum e SHAP time de (a) uma amostra de vocalização de uma região não-impactada classificada como impactada (FP); (b) uma amostra de vocalização de uma região impactada classificada como não-impactada (FN) . . . . .	68
Figura 18 – SHAP Time-Amplitude de a) de uma amostra de vocalização de uma região impactada (TP); b) de uma amostra de vocalização de uma região não impactada (TN); c) de uma amostra de vocalização de uma região não-impactada, mas classificada como impactada (FP); d) Uma amostra de vocalização de uma região não impactada, mas classificada como impactada (FN) . . . . .	71

Figura 19 – SHAP Envoltório de: a) de uma amostra de vocalização de uma região impactada (TP); b) de uma amostra de vocalização de uma região não impactada (TN); c) de uma amostra de vocalização de uma região não-impactada, mas classificada como impactada (FP); d) Uma amostra de vocalização de uma região não impactada, mas classificada como impactada (FN) . . . . .	72
Figura 20 – Espectrograma Global SHAP . . . . .	73
Figura 21 – (a) SHAP Frequênci a Global (b) SHAP frequênci a global em valores absolutos . . . . .	73

---

## LISTA DE TABELAS

Tabela 1 – Categorização da transferência de aprendizagem profunda por TAN et al. . . . .	36
Tabela 2 – Parâmetros da função <i>get_melspectrogram_layer</i> . . . . .	56
Tabela 3 – Métricas: Precisão, Revocação e F1-Score . . . . .	62

---

## LISTA DE ABREVIATURAS E SIGLAS

---

# SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>16</b>
<b>1.1</b>	<b>Contextualização do problema</b>	<b>17</b>
<b>1.2</b>	<b>Objetivos</b>	<b>18</b>
1.2.1	Objetivos específicos	19
<b>1.3</b>	<b>Justificativa</b>	<b>19</b>
<b>1.4</b>	<b>Contribuições esperadas</b>	<b>20</b>
<b>1.5</b>	<b>Organização desta Dissertação</b>	<b>21</b>
<b>2</b>	<b>FUNDAMENTOS</b>	<b>23</b>
<b>2.1</b>	<b>Corruíra e suas Características Sonoras</b>	<b>23</b>
2.1.1	Características do Canto	24
2.1.2	Desenvolvimento das Vocalizações	24
2.1.3	Funções do Canto e variações	25
<b>2.2</b>	<b>Bioacústica</b>	<b>25</b>
<b>2.3</b>	<b>Processamento de Sinais Acústicos</b>	<b>27</b>
2.3.1	Amostragem	27
2.3.2	Filtragem	28
2.3.3	Janela de Hann	28
2.3.4	Transformada Rápida de Fourier (FFT)	29
2.3.5	Espectrograma Mel	29
<b>2.4</b>	<b>Redes Neurais Convolucionais (CNNs)</b>	<b>30</b>
2.4.1	Estrutura das CNNs	31
2.4.2	Aplicações em Bioacústica	33
2.4.3	Vantagens das CNNs	34
<b>2.5</b>	<b>Transferência de Aprendizagem</b>	<b>34</b>

2.5.1	Transferência de Aprendizagem . . . . .	34
2.5.1.1	Categorias de Transferência de Aprendizagem . . . . .	35
2.5.1.2	Transferência de Aprendizagem Profunda . . . . .	36
<b>2.6</b>	<b>Modelos Ponta-a-Ponta (E2E)</b> . . . . .	<b>37</b>
2.6.1	Kapre: Pré-processamento em Tempo Real usando Camadas do Modelo . . . . .	38
2.6.2	Biophony: Modelo Convolucional para Bioacústica . . . . .	39
<b>2.7</b>	<b>Inteligência Artificial Explicável</b> . . . . .	<b>40</b>
2.7.1	SHAP . . . . .	40
2.7.2	DeepSHAP . . . . .	42
<b>2.8</b>	<b>Métricas</b> . . . . .	<b>43</b>
2.8.1	Métricas Comuns . . . . .	43
<b>2.9</b>	<b>Considerações Finais</b> . . . . .	<b>44</b>
<b>3</b>	<b>TRABALHOS RELACIONADOS</b> . . . . .	<b>46</b>
<b>3.1</b>	<b>Aprendizagem profunda em Bioacústica</b> . . . . .	<b>47</b>
<b>3.2</b>	<b>Explicabilidade na bioacústica</b> . . . . .	<b>49</b>
<b>3.3</b>	<b>Efeito de ruído antropogênico em <i>Troglodytes Aedon</i></b> . . . . .	<b>51</b>
<b>3.4</b>	<b>Considerações Finais</b> . . . . .	<b>52</b>
<b>4</b>	<b>METODOLOGIA</b> . . . . .	<b>54</b>
<b>4.1</b>	<b>Conjunto de Dados Corruíra</b> . . . . .	<b>54</b>
<b>4.2</b>	<b>Processamento dos áudios</b> . . . . .	<b>55</b>
<b>4.3</b>	<b>Modelo E2E (Biophony e Kapre)</b> . . . . .	<b>56</b>
4.3.1	Verificando se há viés em relação ao ruído de fundo . . . . .	59
<b>4.4</b>	<b>SHAP</b> . . . . .	<b>60</b>
<b>5</b>	<b>RESULTADOS</b> . . . . .	<b>61</b>
<b>5.1</b>	<b>O Modelo</b> . . . . .	<b>61</b>
<b>5.2</b>	<b>Explicabilidade do Modelo</b> . . . . .	<b>62</b>
5.2.1	Espectrograma SHAP . . . . .	64
5.2.2	Visualizações SHAP . . . . .	66

5.2.2.1	SHAP Spectrum . . . . .	67
5.2.2.2	SHAP Time . . . . .	69
5.2.3	SHAP Time-Amplitude . . . . .	70
5.2.3.1	Envoltório-Amplitude SHAP . . . . .	70
5.2.4	Contribuições de frequência global (SHAP Espectrograma global)	72
5.2.4.1	SHAP Frequênci a global . . . . .	73
<b>5.3</b>	<b>Considerações Finais</b> . . . . .	<b>74</b>
<b>6</b>	<b>CONCLUSÕES</b> . . . . .	<b>76</b>
<b>6.1</b>	<b>Extensões futuras</b> . . . . .	<b>77</b>
<b>Referências</b>	. . . . .	<b>78</b>

# 1

---

## INTRODUÇÃO

O monitoramento bioacústico é uma ferramenta não invasiva e de baixo custo para a observação animal, oferecendo perspectivas relevantes sobre biodiversidade, habitat e ecossistema.

As características sonoras apresentadas por esses animais podem ser indicadores para determinar o bem-estar e a relação de um grupo de indivíduos com o ambiente que eles habitam. O monitoramento bioacústico não só permite reconhecer espécies, mas também prever mudanças na ecologia, observando fenômenos correlacionados, como alterações no ar, solo e clima, com a vantagem de ser não invasivo ([SOUZA; GATTO; FUKUI, 2019](#)). Esses indicadores também mostram informações sobre uma variedade de populações como pássaros, sapos, morcegos, peixes e mosquitos ([EENS et al., 1999; FERNANDES; CORDEIRO; RECAMONDE-MENDOZA, 2020; CULLINAN; MATZNER; DUBERSTEIN, 2015; MALFANTE et al., 2016](#)).

O monitoramento de aves tem se destacado como uma das áreas mais promissoras para o uso de indicadores bioacústicos, especialmente devido à riqueza e complexidade das vocalizações das aves. Essas vocalizações podem refletir as condições ambientais e as pressões ecológicas às quais os pássaros estão sujeitos, pois o canto influencia na reprodução, escolha de parceiros, defesa territorial e reconhecimento individual. No entanto, a presença de ruído antropogênico, como o tráfego urbano, pode forçar as aves a ajustarem seus cantos, o que pode comprometer sua comunicação e, consequentemente, sua sobrevivência e reprodução. ([SEMENTILI-CARDOSO; DONATELLI, 2021](#))

A espécie de pássaro corruíra (*Troglodytes aedon*) serve como um excelente modelo para estudos bioacústicos devido à sua ampla distribuição e à diversidade de suas vocalizações. Compreender como essa espécie ajusta suas vocalizações em resposta ao ruído antropogênico é decisivo para desenvolver estratégias eficazes de conservação. Além disso, a aplicação de técnicas de aprendizado profundo (Deep Learning) no estudo das vocalizações de corruíra oferece novas possibilidades para analisar e interpretar diferentes padrões de maneira eficiente e compreensiva.

Este trabalho propõe uma abordagem E2E (End-to-End) para classificar vocalizações do corruíra em ambientes de impacto antropogênico e explicar de forma transparente as previsões do modelo usando os valores de SHAP (SHapley Additive exPlanations). Ao integrar modelos de Deep Learning (DL) baseados em Redes Neurais Convolucionais (CNNs) com técnicas de interpretabilidade, este estudo visa não apenas melhorar a precisão da classificação bioacústica, mas também fornecer percepções valiosas sobre os padrões acústicos mais afetados no canto da ave pelo ruído antropogênico. Assim, espera-se contribuir significativamente para a compreensão e preservação da espécie em ambientes impactados pela atividade humana.

## 1.1 Contextualização do problema

A bioacústica aviária é uma das mais utilizadas para obtenção de indicadores ecológicos. O canto das aves, além de auxiliar na escolha de parceiros, é parte da reprodução, reconhecimento individual, auxilia na fuga de predadores e defesa territorial, e reflete as mudanças no ambiente em quais esses animais estão inseridos ([SEMENTILI-CARDOSO; DONATELLI, 2021](#)).

Qualquer interferência na transmissão dos cantos pode comprometer a subsistência dessas espécies, dificultando a reprodução e a proteção de seu território. O ruído antropogênico, em particular, compele várias aves a ajustar seus sinais acústicos para minimizar o seu impacto na comunicação.

Ao se adaptar, aves mudam várias características sutis, mas fundamentais em sua estrutura sonora e definir quais foram as características e fatores que mudaram

nessas vocalizações se tornam desafiantes tanto no ponto de vista do especialista como da bioacústica computacional. Quando consideramos os classificadores computacionais utilizados, o que os influenciam, e as escolhas feitas, e a natureza de "caixa preta" dos métodos de aprendizagem profunda, há a necessidade de transparência e confiança para os estudos bioacústicos.

Além disso, com a bioacústica computacional se beneficiando dos avanços e da robustez dos modelos de DL, o monitoramento contínuo de larga escala gera desafios com o acúmulo de dados de áudio, a falta de tempo e falta de profissionais especializados para inspecionar os dados, aumentando a importância de métodos que podem automatizar maior parte ou todo o fluxo de trabalho ([BROWN; RIEDE, 2017](#); [ROCH et al., 2017](#); [STOWELL, 2022](#)).

Uma espécie de interesse é o pássaro corruíra (*Troglodytes aedon*), um pássaro amplamente difundido nas Américas e conhecido por suas vocalizações. As variações nos cantos do corruíra têm em resposta a vários fatores ambientais, incluindo a presença de ruído antropogênico, pode gerar mudanças significativas na comunicação acústica e na sobrevivência da espécie.

Quando expostas a ruídos antropogênicos, essa espécie altera sua vocalização e frequência vocal para se ajustarem ao ruído ambiente. Essas mudanças na vocalização do Corruíra foi observada por meio de análises estatísticas e acústicas tradicionais ([SEMENTILI-CARDOSO; DONATELLI, 2021](#); [DINIZ; DUCA, 2021](#); [CYR KIMBERLEY WETTEN; KOPER, 2021](#)). No entanto, essas mudanças também podem ser observadas por meio de análises computacionais, que incluem técnicas de aprendizado profundo, como serão expostas nesse trabalho.

## 1.2 Objetivos

Desenvolver uma abordagem E2E capaz de classificar quando uma vocalização sofreu mudanças devido à ação do ruído antropogênico, tirando vantagem de técnicas de transferência de aprendizado aplicadas a bases pequenas para melhorar a eficiência da classificação automatizada. Posteriormente, identificar com uso de XAI os padrões

subjacentes às mudanças observadas no repertório vocal do corruíra, contribuindo assim para uma compreensão mais abrangente da ecologia acústica da espécie e sua adaptação a ambientes degradados.

### 1.2.1 Objetivos específicos

- Criar e treinar o modelo Biophony e sua integração com Kapre num modelo E2E;
- Utilizar técnicas de transferência de aprendizagem para melhorar o desempenho do modelo em bases de dados bioacústicos com poucas amostras;
- Demonstrar como as características identificadas pelo SHAP são utilizadas pelo classificador para diferenciar as vocalizações em ambientes com e sem ruído antropogênico;
- Criar novas formas de visualização para os padrões observados pelo modelo E2E, comparáveis àquelas utilizadas em processamento digital de sinais.

## 1.3 Justificativa

O monitoramento bioacústico auxilia no processo de proteção e conhecimento ambiental, mas apresenta desafios constantes, como, por exemplo, grande volume de dados gerado que excede a quantidade de especialistas disponíveis para rotular, poucos dados para espécies raras ou difíceis de encontrar, ou capturar, quantidade pouco significativa de exemplo de áudios; eventos sonoros que exijam especialistas para rotulagem e vários outros. A bioacústica computacional propõe estratégias que ajudem a facilitar a vida desses profissionais.

Nesse trabalho, é proposta a abordagem E2E (ponta-a-ponta), que integra o modelo Biophony com Kapre, com o potencial de melhorar significativamente o pipeline que envolve o processamento de dados acústicos, permitindo uma identificação e classificação dos indivíduos da espécie *Troglodytes aedon*, mesmo com vocalizações modificadas pelos efeitos do ruído antropogênico.

Ao fornecer uma análise e interpretação acústica mais objetiva e rápida, esta pesquisa oferece aos biólogos e pesquisadores uma ferramenta poderosa para estudar e proteger a biodiversidade. Além disso, introduz uma nova maneira de analisar os resultados de modelos de classificação bioacústica, permitindo uma visualização clara dos padrões de tempo-frequência aprendidos pelo modelo CNN e justificando as decisões do modelo para o sucesso da classificação.

## 1.4 Contribuições esperadas

Este estudo oferece contribuições significativas para a bioacústica computacional e a ecologia informática:

- Uma abordagem E2E capaz de classificar indivíduos sujeitos a ruídos antropogênicos, e perceberem as variações nas vocalizações do corruíra. A integração do modelo Biophony com Kapre em uma abordagem E2E demonstra a capacidade de classificar indivíduos da mesma espécie, mesmo quando suas vocalizações são afetadas por mudanças ambientais, validando o impacto de técnicas de transferência de aprendizagem e aumento artificial de dados acústicos. Isso é particularmente importante para esforços de conservação e pesquisa, fornecendo aos biólogos e pesquisadores ferramentas mais precisas e rápidas para estudar e proteger a biodiversidade.
- Adicionalmente, ao utilizar técnicas de XAI (Explainable AI) como o SHAP, este estudo desvenda os padrões intrincados e os fatores causais subjacentes às mudanças no repertório vocal do corruíra, contribuindo para uma compreensão mais abrangente da ecologia acústica da espécie e sua adaptação a ambientes em rápida mudança.
- Ao criar novas formas de visualização para os padrões observados pelo modelo, esta pesquisa oferece uma maneira inovadora de analisar modelos de classificação bioacústica, permitindo uma visualização clara dos padrões de tempo-frequência

aprendidos pelo modelo CNN e justificando as decisões do modelo, contribuindo assim para o avanço da bioacústica computacional.

- Por fim, produzir 2 (dois) artigos, um já publicado como coautor para Simpósio Brasileiro de Computação aplicada a Saúde (SBCAS 2023) intitulado "Classification of Tropical Disease-carrying Mosquitoes Using Deep Learning and SHAP" e outro submetido como autor para a Conference on Graphics, Patterns and Images (SIBGRAPI 2024) intitulado "An XAI-Driven Approach for Bioacoustics Classification and Interpretation".

## 1.5 Organização desta Dissertação

Este trabalho está organizado em seis capítulos, cada um abordando diferentes aspectos do estudo. O **Capítulo 2 - Fundamentação Teórica** revisa os principais conceitos e estudos relacionados à bioacústica, processamento sonoro, aprendizagem profunda, técnicas de transferência de aprendizagem e aumento de dados, abordagem E2E (Biophony e Kapre) e explicabilidade em inteligência artificial. Por fim falamos um pouco sobre o ator principal dessa pesquisa: A corruíra (*Troglodytes aedon musculus*).

No **Capítulo 3 - Trabalhos Relacionados**, são discutidos três pontos principais: a bioacústica computacional, as características e mudanças das vocalizações da espécie *Troglodytes aedon* sob os impactos do ruído antropogênico, e as técnicas de XAI para bioacústica. Este capítulo fornece uma visão geral das pesquisas já realizadas nessas áreas, destacando os avanços, desafios e lacunas que o presente estudo visa abordar.

No **Capítulo 4 - Metodologia**, são detalhados os métodos e procedimentos utilizados no desenvolvimento da pesquisa.

No **Capítulo 5 - Resultados e Discussão**, são apresentados e discutidos os resultados dos experimentos realizados. São avaliadas a precisão e robustez do modelo E2E na classificação das vocalizações, a eficácia das técnicas de aumento de dados e transferência de aprendizagem, e a interpretação das previsões utilizando DeepSHAP alinhada as novas formas de visualizações propostas para esse trabalho.

Finalmente, o **Capítulo 6 - Conclusões e Trabalhos Futuros** resume as principais

descobertas da pesquisa, destacando as contribuições para a área de bioacústica e a relevância das técnicas de explicabilidade em modelos de aprendizagem profunda. Além disso, são discutidas as limitações do estudo e propostas direções para pesquisas futuras, visando o aprimoramento das abordagens apresentadas.

## 2

---

# FUNDAMENTOS

Neste capítulo serão apresentados os conceitos fundamentais para compreensão e desenvolvimento do trabalho. Esse embasamento teórico será dividido em oito seções.

A seção 2.1 apresenta o principal alvo deste estudo: a Corruíra (*Troglodytes aedon*). A seção 2.2 apresenta o conceito de bioacústica, destacando seu uso para monitoramento ambiental e as contribuições proporcionadas pelo uso de ferramentas computacionais. Em seguida, na seção 2.3, é fornecido um breve resumo sobre processamento de sinais acústicos. Na seção 2.4, são definidas as redes neurais convolucionais (CNNs) e sua aplicação em bioacústica. A seção 2.5 explora técnicas para melhorar o desempenho dos modelos de aprendizagem profunda, como transferência de aprendizagem. Na seção 2.6, é detalhada a abordagem ponta-a-ponta utilizada neste trabalho, incluindo a descrição das componentes Biophony e do framework Kapre. A seção 2.7 aborda conceitos de IA explicável, com ênfase em SHAP e DeepSHAP. Em seguida, a seção 2.8 apresenta as métricas utilizadas neste estudo. Finalmente, na seção 2.9, são apresentadas as considerações finais deste capítulo.

## 2.1 Corruíra e suas Características Sonoras

A Corruíra (*Troglodytes musculus*) são pequenos pássaros com plumagem marrom, asas e cauda com barras, conhecida por suas vocalizações complexas e variadas, cujas características sonoras têm sido amplamente estudadas. A estrutura vocal da Corruíra varia significativamente dependendo do ambiente, e podem ser encontrados em toda a

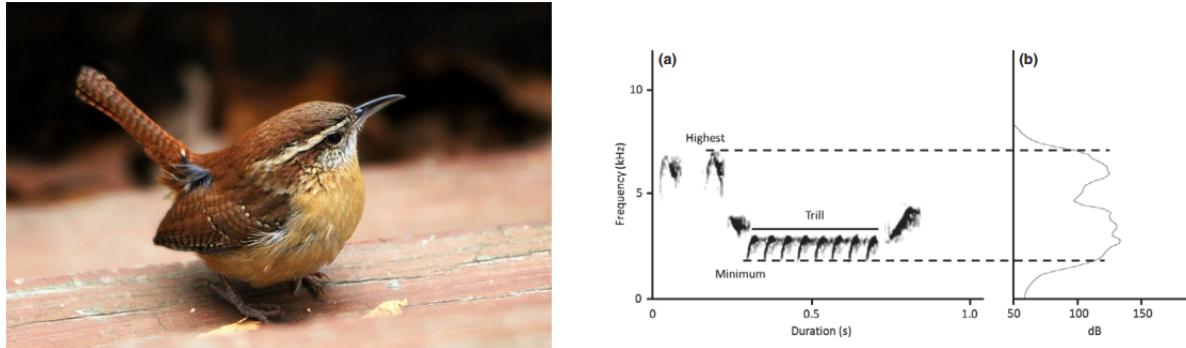


Figura 1 – (a) Individuo Corruíra macho (imagem de copyright livre) (b) Espectrograma do canto do Corruíra mostrando o trinado e as medições de frequência mínima e máxima utilizadas para análise. (c) Espectro de potência usado para definir a frequência máxima e mínima no canto. Linhas tracejadas mostram os limites de frequência máxima e mínima no espectrograma e no espectro de potência por [SANDOVAL; REDONDO; BARRANTES](#)

extensão da América. ([SANDOVAL; REDONDO; BARRANTES, 2013; SKUTCH, 1953](#)).

### 2.1.1 Características do Canto

As corruíras machos utilizam cantos espontâneos de alto volume, principalmente para atrair parceiras antes do acasalamento e após perder uma parceira. Esses cantos também são direcionados a intrusos e machos vizinhos para defender territórios. A diversidade de tipos de sílabas e a maior quantidade de repertórios de cantos nos machos de corruíra estão associadas ao emparelhamento mais precoce e à maior produção de ovos pelas fêmeas. No entanto, esses machos exibem menor diversidade imediata de tipos de canto. Os Machos possuem cantos mais complexos e regiões de controle de canto maiores no cérebro em comparação com as fêmeas. Nas fêmeas, o aumento da complexidade do canto está ligado a volumes maiores e mais células em regiões cerebrais específicas ([E. Cramer, 2013a; E. Cramer, 2013b](#)).

### 2.1.2 Desenvolvimento das Vocalizações

O desenvolvimento das vocalizações nas corruíras começa cedo na vida. Os filhotes produzem chamadas que evoluem para chamadas semelhantes às dos adultos, que

servem como sinais de alerta. Os jovens também produzem sub cantos, que se assemelham às notas dos cantos dos machos adultos. O desenvolvimento dessas características vocais não é linear e pode apresentar mudanças abruptas, possivelmente devido a novas funções sociais à medida que as aves amadurecem ([Mala H. Sawhney; M. C. Baker; Bradley R. Bisbee, 2006](#)).

### 2.1.3 Funções do Canto e variações

Os cantos dos machos corruíras têm múltiplas funções ao longo do ciclo reprodutivo. Cantos espontâneos de alto volume são usados principalmente para atração de parceiras antes do acasalamento e após a perda de uma parceira. Uma vez acasalados, os machos usam cantos para se comunicar com suas parceiras, muitas vezes para sinalizar a ausência de ameaças imediatas, permitindo que a fêmea se mova com segurança de e para o ninho. Os cantos também são direcionados a intrusos e machos vizinhos para defender territórios. ([L. Kermott; L. Scott Johnson, 1991](#))

A estrutura do canto das corruíras varia entre populações. Por exemplo, em duas populações no leste da Argentina, os cantos diferem no número de tipos de sílabas e no comprimento geral. Essa variação sugere um potencial para o aprendizado vocal imitativo, evidenciado pelo compartilhamento de cantos intraespecíficos e pela presença de tipos de sílabas variáveis entre os indivíduos ([P. Tubaro, 1990](#)). Essas diferenças expressam como as características vocais das corruíras são influenciadas por fatores ambientais, estágios de desenvolvimento e contextos sociais. Seus cantos e chamadas desempenham funções críticas na comunicação, atração de parceiros e defesa de territórios, com variações observadas em diferentes populações e ambientes.

## 2.2 Bioacústica

Bioacústica é o estudo interdisciplinar da produção, transmissão e recepção de sons em animais, incluindo humanos. Este campo envolve a análise dos mecanismos biológicos e físicos que possibilitam a produção e percepção dos sons, bem como as

funções ecológicas e evolutivas desses sinais acústicos (SOUZA; GATTO; FUKUI, 2019; ERBE, 2016). A comunicação sonora animal é um processo biológico que serve como ferramenta importante para o estudo da filogenia, do estado fisiológico, da ecologia comportamental e das relações evolutivas entre os animais. Esses sons são produzidos para diversas finalidades, como defesa territorial, interações em grupo, atração de parceiros e orientação. Através da análise desses sons, é possível obter informações sobre a biodiversidade local. (EENS et al., 1999; FERNANDES; CORDEIRO; RECAMONDE-MENDOZA, 2020; CULLINAN; MATZNER; DUBERSTEIN, 2015; MALFANTE et al., 2016) Assim, a bioacústica tem uso na identificação e monitoramento de espécies, na conservação da biodiversidade, no manejo de populações naturais e na investigação dos impactos dos sons naturais e antropogênicos sobre os comportamentos e o bem-estar dos organismos. O monitoramento bioacústico é um método não invasivo para avaliar parâmetros ecológicos relevantes por envolver a gravação passiva dos sons animais e fornece perspectivas valiosas sobre questões de integridade ambiental, como biodiversidade e a presença ou ausência de espécies.

A bioacústica computacional, por sua vez, utiliza métodos de aprendizado profundo para analisar vocalizações animais e paisagens sonoras naturais, oferecendo evidências valiosas sobre comportamentos animais, populações e ecossistemas (STOWELL, 2022; MUTANU et al., 2022a; KVSN et al., 2020). Com avanços tecnológicos, como gravadores eficientes e computadores com maior poder de processamento, a bioacústica computacional acelerou nas últimas décadas, aliada aos progressos apresentados em áreas da computação como processamento de sinais e aprendizagem de máquina. Apesar dos avanços, as demandas e desafios gerados pelos dados bioacústicos são muitas vezes diferentes dos já estabelecidos nas áreas de processamento de linguagem natural. Assim, permanecem várias preocupações e problemas ainda não resolvidos, o que torna esse campo promissor (STOWELL, 2022).

## 2.3 Processamento de Sinais Acústicos

Processamento acústico refere-se ao conjunto de técnicas e métodos aplicados para analisar e interpretar sinais sonoros. É uma etapa essencial na bioacústica computacional por envolver a transformação de sinais acústicos em representações que podem ser analisadas e interpretadas por métodos de aprendizado de máquina. Este processo compreende várias etapas desde a amostragem, filtragem, conversão do sinal usando transformações matemáticas para a preparação dos dados para análises subsequentes([STOWELL, 2022](#)).

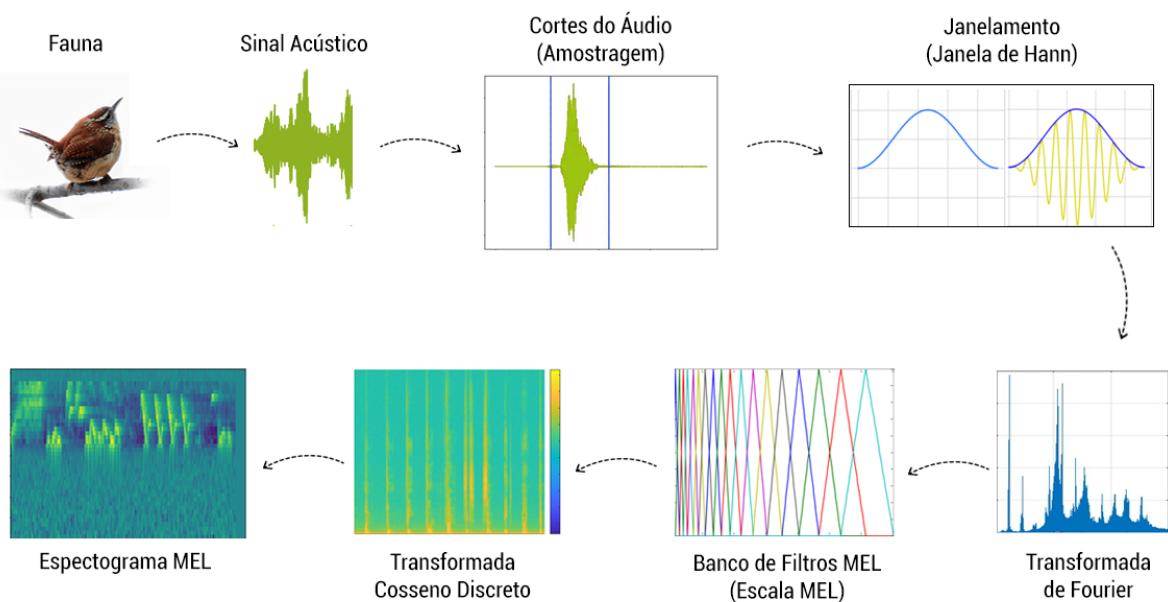


Figura 2 – Etapas do processamento sonoro bioacústico.

### 2.3.1 Amostragem

Os sinais acústicos são contínuos no tempo e precisam ser digitalizados para processamento computacional. Esse processo envolve amostragem, onde o sinal contínuo é convertido em uma sequência discreta de pontos. A taxa de amostragem deve ser alta o suficiente para capturar todas as informações relevantes do sinal, conforme o teorema de Nyquist-Shannon, que estabelece que a frequência de amostragem  $f_s$  deve ser pelo menos o dobro da frequência máxima  $f_{max}$  contida no sinal ([SONG et al., 2012](#)), i.e.  $f_{max} = \frac{f_s}{2}$ .

### 2.3.2 Filtragem

A filtragem é usada para remover ruídos indesejados e interferências do sinal. Filtros passa-baixa, passa-alta e passa-banda são os filtros mais comuns no processamento de áudio na bioacústica ([MADHUSUDHANA et al., 2022](#)). O passa-banda é o filtro que permite a passagem de frequências em um intervalo específico, enquanto atenua ou rejeita as frequências que estão fora desse intervalo. Um filtro passa-banda pode ser representado pela função de transferência  $H(z)$ :

$$H(z) = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{1 + a_1 z^{-1} + a_2 z^{-2}} \quad (2.1)$$

onde:

- $b_0, b_1, b_2$  são os coeficientes do numerador que determinam como o sinal de entrada é ponderado para produzir o sinal de saída desejado.
- $a_1, a_2$  são os coeficientes do denominador que determinam como os valores passados do sinal de saída e do sinal de entrada são ponderados para produzir o próximo valor do sinal de saída.

Esta função de transferência descreve o comportamento de um filtro passa-banda digital, onde  $z^{-1}$  e  $z^{-2}$  representam os atrasos de uma amostra no domínio discreto.

### 2.3.3 Janela de Hann

Como resultado da discretização dos dados, as extremidades do conjunto do finito de pontos, ao passarem pela transformação de Fourier, uma função periódica, geram pontos de descontinuidade que se apresentam como frequências inexistentes no espectro do sinal. As técnicas de janelamento servem para reduzir os problemas gerados pela discretização dos dados, suavizando esses pontos de descontinuidade ([CORP., 2021](#)). Uma das funções de janelamento mais usadas no processamento sonoro é a janela de Hann ([CHACHADA; KUO, 2014](#)). A função de Hann  $w(n)$  é definida pela equação:

$$w(n) = 0.5 \left( 1 - \cos \left( \frac{2\pi n}{N-1} \right) \right) \quad \text{para } 0 \leq n \leq N-1, \quad (2.2)$$

onde  $N$  é o número total de pontos na janela.

### 2.3.4 Transformada Rápida de Fourier (FFT)

Para poder representar um som pelas suas frequências é necessário aplicar a transformada discreta de Fourier ([COOLEY; TUKEY, 1965](#)). Essa transformação é usada para converter o sinal de áudio do domínio do tempo para o domínio da frequência. No entanto, devido à complexidade computacional da FT, a Transformada Rápida de Fourier (FFT) é utilizada para tornar este processo mais eficiente utilizando um método numérico baseado em divisão e conquista. A fórmula da FFT é dada por:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-i\frac{2\pi}{N}kn} \quad (2.3)$$

Nesta fórmula,  $X[k]$  representa o valor da transformada no índice  $k$ , enquanto  $x[n]$  é o valor da sequência de entrada no índice  $n$ . A variável  $N$  denota o número total de pontos na DFT. O índice  $k$  varia de 0 a  $N - 1$  e corresponde à frequência, enquanto o índice  $n$  também varia de 0 a  $N - 1$  e corresponde ao tempo. A constante  $e$  é a base do logaritmo natural, e  $i$  é a unidade imaginária, onde  $i^2 = -1$ . O fator de normalização  $\frac{2\pi}{N}$  relaciona os índices de tempo e frequência.

A FFT computa essa transformação em tempo  $O(N \log N)$ , onde  $N$  é o número de pontos no sinal, tornando-a adequada para processamento em tempo real.

### 2.3.5 Espectrograma Mel

O espectrograma Mel é uma representação visual do espectro de frequências Mel com magnitude medida em decibéis ([BROWN; RIEDE, 2017](#)). É utilizado para analisar como a frequência de um sinal varia ao longo do tempo, fornecendo uma visão detalhada das características acústicas dos sinais. O espectrograma é obtido aplicando a Transformada de Fourier a segmentos do sinal, multiplicados por uma janela deslizante. A magnitude do espectrograma  $S(t, f)$  é dada por:

$$S(t, f) = \left| \int_{-\infty}^{\infty} x(\tau)w(\tau - t)e^{-j2\pi f\tau} d\tau \right| \quad (2.4)$$

Para transformar as frequências para a escala Mel, a fórmula usada é:

$$M(f) = 2595 \log_{10} \left( 1 + \frac{f}{700} \right) \quad (2.5)$$

onde  $f$  é a frequência em Hertz e  $M(f)$  é a frequência na escala Mel. Essa relação apresenta um comportamento linear abaixo da frequência 1 kHz e logarítmica acima de 1 kHz. Na prática, é utilizado um banco de filtros de resposta triangular, com espaçamento e largura determinada pela constante de frequência Mel para fazer essa conversão de sinal para a escala Mel.

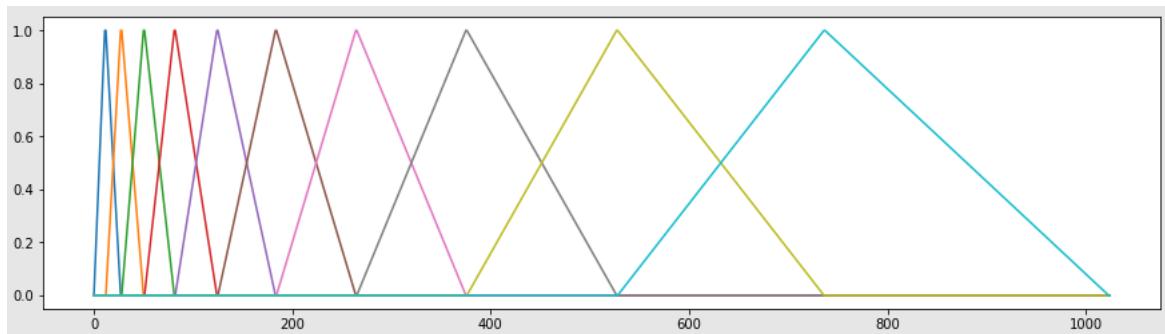


Figura 3 – Banco de Filtros Mel.

O spectrograma Mel é bastante popular na área de reconhecimento de fala e bioacústica pelo seu uso nos classificadores de aprendizagem profunda ([STOWELL, 2022](#); [KVSN et al., 2020](#))

## 2.4 Redes Neurais Convolucionais (CNNs)

Redes Neurais Convolucionais (CNNs) são uma classe de redes neurais artificiais que usam camadas de convolução para extrair características dos dados de entrada e camadas densas para classificação. Propostas por Yann LeCun na década de 1980, as CNNs são populares em diversas áreas, incluindo classificação de imagens, detecção de objetos, reconhecimento de ação, processamento de fala e linguagem natural, e bioacústica computacional ([GU et al., 2018a](#)).

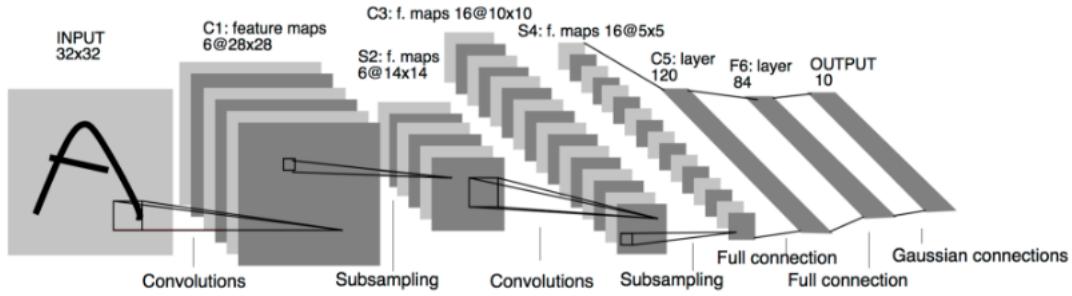


Figura 4 – Estrutura básica de uma CNN ([LECUN et al., 1998](#))

#### 2.4.1 Estrutura das CNNs

A arquitetura de uma CNN típica é composta por múltiplos grupos com quatro componentes principais: um banco de filtros (filter bank) ou kernels, camadas de convolução, camadas de pooling e uma função de ativação não-linear ([NARANJO-TORRES et al., 2020](#)).

- **Camada de Banco de filtros:** Tem a função de detectar determinada característica em cada local de entrada e manter essa característica sem alterações para transferi-la para a saída; Para que a detecção ocorra, existe um banco de  $m$  filtros em cada camada convolucional a saída  $Y_i^{(l)}$  na  $l$ -ésima camada que consiste em  $m_1^{(l)}$  mapas de características (*Feature Maps*) de tamanho  $m_2^{(l)} \times m_3^{(l)}$ . O  $i$ -ésimo mapa de características é dado pela equação:

$$Y_i^{(l)} = B_i^{(l)} + \sum_{j=1}^{m_1^{(l-1)}} k_{ij}^{(l)} * Y_j^{(l-1)} \quad (2.6)$$

onde  $B_i^{(l)}$  denota o viés da matriz de parâmetros treináveis,  $k_{ij}^{(l)}$  é o filtro com dimensões  $(2h_1^{(l+1)} \times 2h_2^{(l+1)})$ , o qual conecta o  $j$ -ésimo mapa de características da camada  $l - 1$  com o  $i$ -ésimo mapa de características da camada  $l$  e  $*$  é o operador de convolução 2D discreta.

- **Camadas Convolucionais:** Camadas em quais ocorrem as operações de convolução; nessa camada existe uma matriz 2D representando uma imagem ( $I$ ) que aplicada a uma matriz menor ( $K$ ) de kernel 2D, a equação:

$$S_{i,j} = (I * K)_{i,j} = \sum_m \sum_n I_{i,j} K_{i-m,j-n} \quad (2.7)$$

mostra a operação de convolução, aplicada entre as duas matrizes (HEATON, 2018).

No processo de convolução, um pequeno filtro deslizante percorre a imagem da esquerda para a direita e de cima à baixo. A figura 5 exemplifica com uma matriz  $3 \times 3$  e um *kernel*  $2 \times 2$ , obtendo uma matriz convolucionada na saída. Em cada posição, é feita a soma dos produtos dos elementos do *kernel* como elementos correspondentes na matriz de entrada. Esse processo é repetido com diferentes *kernels* para gerar diferentes mapas de características. (DUMOULIN; VISIN, 2016)

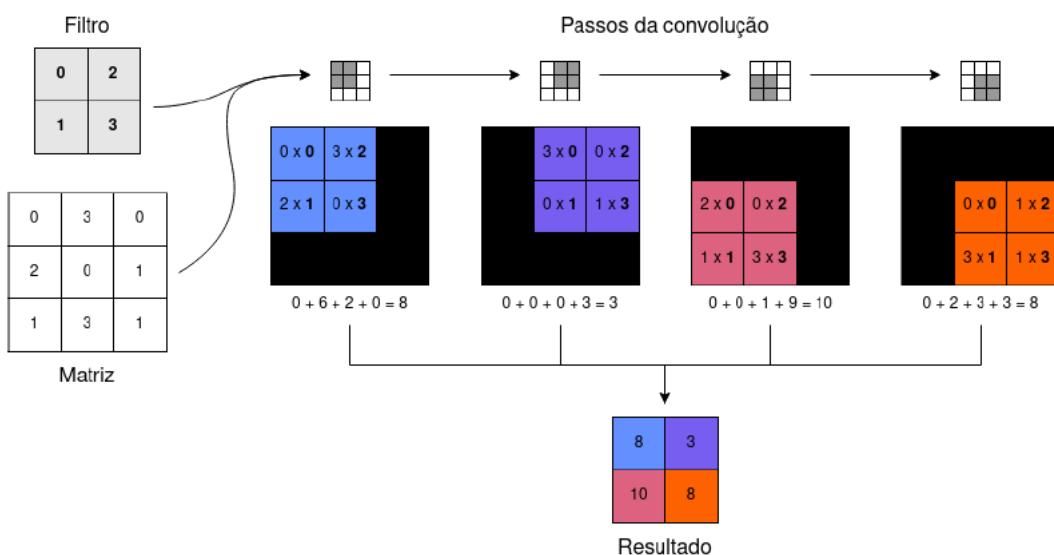


Figura 5 – Exemplo de convolução de uma matriz  $3 \times 3$  e um *kernel*  $2 \times 2$

Esse procedimento reduz o tamanho do mapas de características de saída, mas é possível manter esse tamanho usando a técnica de *padding*, inserindo zeros ao redor da imagem. (DUMOULIN; VISIN, 2016; YAMASHITA et al., 2018) O termo *stride* denota o deslocamento entre duas posições sucessivas do *kernel*. Normalmente utiliza-se *stride* 1, porém *strides* maiores podem ser usadas para reduzir a resolução dos mapas de características, técnica conhecida como subamostragem (*subsampling*).

- **Camadas de Pooling:** Reduzem a dimensionalidade dos mapas de características, preservando as informações mais importantes e tornando os resultados menos sensíveis a pequenas variações na entrada (DUMOULIN; VISIN, 2016; JORDAN;

MITCHELL, 2015; GU et al., 2018b). As principais funções de pooling são Max Pooling e Average Pooling:

- **Max Pooling:** Seleciona o valor máximo de cada região de pooling (SCHERER; MÜLLER; BEHNKE, 2010; LEE; GALLAGHER; TU, 2016).
- **Average Pooling:** Calcula a média de cada região de pooling (SCHERER; MÜLLER; BEHNKE, 2010; LEE; GALLAGHER; TU, 2016).
- **Funções de Ativação:** Introduzem não-linearidades no modelo. A função de ativação ReLU (Rectified Linear Unit) é comumente usada:

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

Essa função é aplicada após a passagem do banco de filtros para produzir os mapas de características, onde apenas as características ativadas seguem para a próxima camada.

- **Camadas Totalmente Conectadas:** Conectam cada neurônio de uma camada à totalidade dos neurônios da camada seguinte, sendo utilizadas nas etapas finais da CNN para realizar a classificação com base nas características extraídas.

O funcionamento de uma CNN envolve a propagação do sinal de entrada através das camadas convolucionais, pooling e totalmente conectadas. Durante o treinamento, a CNN ajusta os pesos dos filtros nas camadas convolucionais para minimizar uma função de perda, que quantifica a discrepância entre as previsões da rede e os valores reais (LECUN et al., 1998).

## 2.4.2 Aplicações em Bioacústica

Em bioacústica, as CNNs são empregadas para analisar espectrograma Mel que são representações visuais de sinais acústicos em termos de tempo, frequência e amplitude. As CNNs são eficazes na identificação de padrões acústicos complexos, permitindo a classificação de espécies, a detecção de vocalizações e a análise de comportamentos animais (STOWELL, 2022; KVSN et al., 2020).

### 2.4.3 Vantagens das CNNs

- **Captura de Padrões Locais:** As camadas convolucionais capturam padrões locais eficientemente, preservando a estrutura espacial dos dados ([NARANJO-TORRES et al., 2020](#)).
- **Redução de Parâmetros:** A operação de pooling reduz a dimensionalidade dos dados, diminuindo o número de parâmetros e, consequentemente, o risco de overfitting ([DUMOULIN; VISIN, 2016](#); [JORDAN; MITCHELL, 2015](#); [GU et al., 2018b](#)).
- **Transferência de Aprendizado:** As CNNs podem ser treinadas em grandes bancos de dados genéricos e, posteriormente, ajustadas para tarefas específicas, aproveitando características aprendidas anteriormente ([NARANJO-TORRES et al., 2020](#)).

Redes neurais convolucionais revolucionam o campo da bioacústica computacional, proporcionando métodos eficientes e precisos para a análise de sinais acústicos. Sua capacidade de extrair e aprender características complexas a partir de dados torna-as ferramentas poderosas em diversas aplicações, desde a conservação de espécies até a monitorização ambiental ([STOWELL, 2022](#); [KVSN et al., 2020](#)).

## 2.5 Transferência de Aprendizagem

### 2.5.1 Transferência de Aprendizagem

A transferência de aprendizagem é uma metodologia proposta para superar as limitações enfrentadas pelos métodos tradicionais de aprendizagem de máquina, que dependem de grandes volumes de dados rotulados para treinar modelos precisos. Colegar esses dados pode ser custoso, demorado ou, em alguns casos, impossível ([WEISS; KHOSHGOFTAAR; WANG, 2016](#)). A transferência de aprendizagem se concentra na transferência de conhecimento entre diferentes domínios para melhorar o desempenho do modelo em um novo domínio, utilizando informações previamente aprendidas em outro contexto ([ZHUANG et al., 2020](#)).

A ideia de transferência de aprendizagem origina-se da psicologia educacional quando C. H. Judd propôs que a experiência pode ser generalizada, permitindo a transferência de conhecimento de uma situação para outra. No contexto da aprendizagem de máquina, a transferência de aprendizagem envolve o uso de um domínio de origem (*source domain*) e uma tarefa de origem (*source task*) para auxiliar na aprendizagem em um domínio de destino (*target domain*) e tarefa de destino (*target task*) (ZHUANG et al., 2020). Um domínio  $D$  é composto por um espaço de características  $X$  e uma distribuição probabilística  $P(X)$ , enquanto uma tarefa  $T$  envolve um espaço de rótulos  $Y$  e uma função preditiva  $f$  (AGARWAL et al., 2021). Dado um domínio de origem  $D_s$  e uma tarefa de origem  $T_s$ , junto com um domínio de destino  $D_t$  e uma tarefa de destino  $T_t$ , a transferência de aprendizagem objetiva melhorar a função preditiva  $f_t$  em  $D_t$  usando a informação de  $D_s$  e  $T_s$ , onde  $D_s \neq D_t$  ou  $T_s \neq T_t$  (WEISS; KHOSHGOFTAAR; WANG, 2016; ZHUANG et al., 2020; AGARWAL et al., 2021).

#### 2.5.1.1 Categorias de Transferência de Aprendizagem

PAN; YANG categorizaram a transferência de aprendizagem em três tipos principais: transferência transdutiva, transferência indutiva e transferência não supervisionada.

- **Transferência Transdutiva:** As tarefas de origem e destino são as mesmas ( $T_s = T_t$ ), mas os domínios diferem ( $D_s \neq D_t$ ). Isso pode ocorrer quando os espaços de características diferem ( $X_s \neq X_t$ ) ou quando as distribuições são diferentes ( $P(X_s) \neq P(X_t)$ ). Nessa categoria, as informações do rótulo vêm apenas do domínio de origem.
- **Transferência Indutiva:** As tarefas de origem e destino são diferentes ( $T_s \neq T_t$ ), independentemente dos domínios. Pode ocorrer em duas situações: com muitos rótulos no domínio de origem (aprendizagem multitarefas) ou sem rótulos no domínio de origem (autoaprendizagem). Neste caso, as informações dos rótulos do domínio de destino estão disponíveis.
- **Transferência Não Supervisionada:** Tanto os domínios quanto as tarefas podem diferir, e as informações dos rótulos são desconhecidas para ambos os domínios

de origem e destino ([ZHUANG et al., 2020; PAN; YANG, 2010](#)).

Além dessas, outras categorizações incluem a transferência de aprendizagem heterogênea e homogênea, bem como abordagens baseadas em dados e modelos ([ZHUANG et al., 2020; AGARWAL et al., 2021](#)).

### 2.5.1.2 Transferência de Aprendizagem Profunda

Transferência de aprendizagem profunda é a aplicação de uma função não linear representada por uma rede neural profunda para transferir conhecimento entre domínios ([TAN et al., 2018](#)). Para modelos de aprendizagem profunda, foram propostos diversos métodos de transferência de aprendizado, distribuídos em quatro categorias principais: aprendizagem baseada em instâncias, aprendizagem baseada em mapeamento, aprendizagem baseada em rede e aprendizagem baseada em redes adversárias.

Tabela 1 – Categorização da transferência de aprendizagem profunda por [TAN et al.](#)

Tipo de Abordagem	Descrição
Instância	Utiliza as instâncias no domínio de origem por meio dos pesos apropriados.
Mapeamento	Mapeia as instâncias de dois domínios em um novo espaço de dados com melhor similaridade.
Rede	Reutiliza a parte de uma rede pré-treinada no domínio de origem.
Adversárias	Usa tecnologia das redes adversárias para encontrar recursos transferíveis adequados para dois domínios.

A transferência de aprendizagem é aplicada com sucesso em várias áreas, como visão computacional, processamento de linguagem natural e bioacústica. No entanto, ainda enfrenta desafios, como a necessidade de grandes volumes de dados rotulados, a alta demanda computacional e a complexidade na interpretação dos modelos.

A transferência de aprendizagem profunda, especialmente a abordagem em redes, aonde partes de redes pré-treinadas são reutilizadas, é uma abordagem poderosa e eficiente. Nesse trabalho, foi utilizado o modelo Biophony, pré-treinado em um dos maiores bancos de dados bioacústicos, o Xeno Canto. Essa abordagem tem se mostrado eficaz para aproveitar o conhecimento previamente adquirido e aplicá-lo a novos pro-

blemas, reduzindo a necessidade de grandes volumes de dados rotulados e tempo de treinamento. (TAN et al., 2018).

## 2.6 Modelos Ponta-a-Ponta (E2E)

Os modelos ponta a ponta (E2E) representam uma abordagem em aprendizado de máquina onde uma única rede neural é treinada para realizar uma tarefa complexa do início ao fim, sem a necessidade de etapas intermediárias. Ao contrário dos métodos tradicionais de aprendizado de máquina, que geralmente envolvem vários estágios de pré-processamento, extração de características e modelagem, os modelos E2E inferem diretamente as características a partir das entradas, como imagens ou texto, minimizando o número de componentes e camadas no sistema, tornando-o mais simples e eficiente (THOMAS et al., 2006).

Os sistemas E2E são aplicados em áreas como visão computacional e processamento de linguagem natural. No campo da visão computacional, sendo reconhecidos em reconhecimento facial (DU et al., 2022) e condução autônoma, aprendendo a partir de dados brutos (XU et al., 2017). No processamento de linguagem natural, modelos podem traduzir a sequência de fala de entrada em uma sequência de saída utilizando uma única rede neural (LI et al., 2022; BAHDANAU et al., 2016; PRABHAVALKAR et al., 2017).

A bioacústica usa várias das ferramentas de processamento de linguagem natural, mas a utilização dos modelos E2E permanece menos explorada, mantendo os desafios na detecção, reconhecimento e extração de informações significativas a partir de áudios puros (BERMANT, 2021). Mesmo com os avanços em aprendizado profundo e redes neurais, ainda há pouca orientação prática para usuários finais, limitando o uso desses métodos no meio acadêmico e fora dele (ULLOA et al., 2018). Com suas vantagens, os modelos E2E têm o potencial de ajudar em áreas menos exploradas como a bioacústica, permitindo o trabalho direto com amostras de áudio puras, onde a representação e os ajustes são feitos automaticamente (BAHDANAU et al., 2016).

O pré-processamento de dados geralmente ocupa muito tempo, esforço e espaço,

principalmente quando há dados de áudio envolvidos, que possuem uma decodificação mais pesada e tamanhos geralmente maiores em relação a dados de imagens ou texto. Por exemplo, 2 segundos de áudio sem compressão com taxa de amostragem de 44,1kHz, que corresponde à qualidade de um CD de áudio, geram um vetor de 88200 amostras. Consequentemente, processar esse vetor em uma camada de entrada num modelo de aprendizagem profunda torna-se um desafio (CHOI; JOO; KIM, 2017a).

Os procedimentos básicos de preparação de dados de áudio para um modelo de aprendizado de máquina incluem decodificação, re-amostragem e conversão para uma representação de tempo-frequência. Enquanto a decodificação e re-amostragem precisam de uma preparação imediata para que não se crie um gargalo, a etapa de conversão tem várias formas de implementação, com conflitos de escolhas entre armazenamento e tempo de computação.

Assim, os modelos E2E oferecem algumas vantagens em relação aos métodos tradicionais que precisam de um pré-processamento prévio. Em vez de otimizar componentes individuais separadamente, esses modelos requerem apenas a otimização de uma única função-objetivo durante o processo de aprendizagem, potencialmente alcançando um desempenho global superior. Eles simplificam o pipeline, eliminando a necessidade de conhecimento especializado para cada componente individual. Além disso, utilizam uma única rede com uma representação de recursos otimizada diretamente dos dados brutos de entrada, tornando-os mais compactos e eficientes (BAHAR; BIESCHKE; NEY, 2019).

### 2.6.1 Kapre: Pré-processamento em Tempo Real usando Camadas do Modelo

Kapre é um *framework* pré-processador de áudio em camadas Keras desenvolvido em Python que pode gerar transformadas discretas de Fourier de curto prazo, transformadas discretas de Fourier inversas de curto prazo, espectrogramas Mel e outras transformações em GPU, em tempo real (CHOI; JOO; KIM, 2017a). Kapre possibilita o processamento de sinal baseado em rede neural, oferecendo suporte a *kernels* treináveis

para transformações de tempo e frequência.

## 2.6.2 Biophony: Modelo Convolucional para Bioacústica

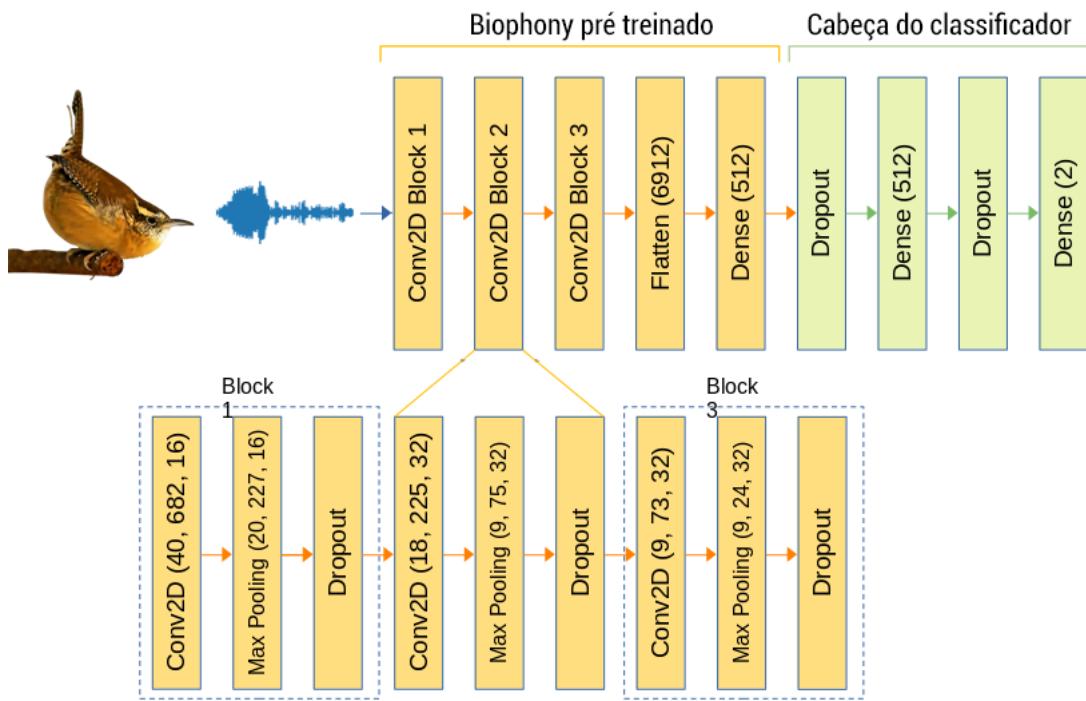


Figura 6 – Arquitetura do Biophony

Biophony é um modelo de rede neural convolucional desenvolvido pela *Conservation Metrics* (CMI) em parceria com a Microsoft para identificar a ave brasileira ameaçada de extinção Soldadinho do Araripe (*Antilophia bokermanni*), uma ave criticamente ameaçada da família dos tangarás (Pipridae) com uma população estimada em apenas 800 indivíduos (FLEISHMAN et al., 2020). Biophony é um modelo relativamente pequeno (16 camadas e 3.995.818 parâmetros treináveis) em comparação a alguns dos modelos mais utilizados na bioacústica (STOWELL, 2022).

O modelo é pré-treinado com amostras de áudio de mais de 300 espécies coletadas pela CMI e utiliza técnicas de transferência de aprendizagem para a classificação do Soldadinho do Araripe. O modelo teve resultados com acurácia de 91% dos cantos do Soldadinho do Araripe em amostras de testes de dois segundos fornecidas ao modelo para avaliação (FLEISHMAN et al., 2020).

## 2.7 Inteligência Artificial Explicável

Inteligência Artificial Explicável (XAI) é um conjunto de técnicas e métodos que permitem o entendimento dos resultados gerados por modelos de inteligência artificial. A XAI visa tornar as decisões e processos dos modelos de inteligência artificiais mais transparentes para os usuários. Utilizando ferramentas de visualização interpretáveis, elucida como métodos de caixa preta realizam a classificação, aprimorando a compreensão do processo de tomada de decisão ([Barredo Arrieta et al., 2020; CANáRIO; RIBEIRO; RIOS, 2022](#)).

Reconhecendo a necessidade de confiabilidade e transparência, houve uma busca maior por interpretabilidade e explicabilidade, com aumento de pesquisas na área. E dada a complexidade do assunto, não há acordo sobre uma única definição ou taxonomia. Apesar do uso concomitante, os termos "interpretável" e "explicável" não compartilham o mesmo significado. **Interpretabilidade** refere-se à capacidade de um modelo ser entendido diretamente por humanos, permitindo que desenvolvedores e usuários com conhecimento técnico compreendam o processo de tomada de decisão do modelo. **Explicabilidade**, por outro lado, envolve a adição de elementos que esclarecem como e por que um modelo chegou a uma determinada decisão, proporcionando uma visão clara e confiável, especialmente em casos onde a interpretação direta não é possível ([ORTIGOSSA; GONçALVES; NONATO, 2024; ALI et al., 2023](#)).

Existem muitas abordagens para interpretabilidade, e vários deles independentes de modelo *model-agnostic* que podem (teoricamente) ser aplicados a qualquer modelo de aprendizado, independentemente da arquitetura ou algoritmo subjacente, pois não são projetados para considerar características específicas de um modelo específico. Uma dessas abordagens é o modelo SHAP (SHapley Additive exPlanations) ([ORTIGOSSA; GONçALVES; NONATO, 2024; BAPTISTA; GOEBEL; HENRIQUES, 2022](#)).

### 2.7.1 SHAP

SHAP (SHapley Additive exPlanations) é um explicador *model-agnostic* (independe do modelo aplicado) que fornece interpretação local utilizando dos valores de Shapley da

teoria dos jogos ([MOSCA et al., 2022](#)). Utilizando valores de Shapley, o SHAP avalia a importância relativa das variáveis de entrada, mostrando a contribuição individual de cada recurso na previsão de uma amostra específica. Ao calcular valores Shapley para cada variável de entrada, como píxeis de uma imagem, o SHAP fornece explicações com fortes propriedades matemáticas, garantindo sua interpretabilidade para amostras individuais ([KAUR et al., 2020](#)).

Desenvolvido por Lloyd Shapley em 1953, valores de Shapley são apresentados como uma solução justa para a distribuição de pagamentos em jogos cooperativos. Em um jogo cooperativo, vários jogadores (ou características) colaboram para obter uma recompensa (ou predição), e os valores de Shapley distribui de forma justa de acordo com esses valores, considerando todas as possíveis coalizões de jogadores.

**Definição Matemática dos Valores de Shapley:** Os valores de Shapley para uma característica  $i$  são a média ponderada das contribuições marginais da característica  $i$  em todas as combinações de características que podem ser encontradas. A diferença na saída do modelo quando a característica  $i$  é adicionada a uma combinação de características é chamada de contribuição marginal. Os valores de Shapley  $\phi_i$  são calculados matematicamente por:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [v(S \cup \{i\}) - v(S)] \quad (2.8)$$

onde:

- $N$  é o conjunto de todas as características.
- $S$  é um subconjunto de  $N$  que não contém a característica  $i$ .
- $v(S)$  é a função de valor que dá a predição do modelo com o subconjunto de características  $S$ .
- $|S|$  é o número de elementos no subconjunto  $S$ .
- $|N|$  é o número de características totais. ([LUNDBERG; LEE, 2017](#))

Como um método aditivo de atribuição de características, o SHAP apresenta solução única com três propriedades essenciais: **Acurácia local** (*Local accuracy*), **Ausê-**

**Cia** (*Missingness*) e **Consistência** (*Consistency*). A propriedade **Acurácia local** refere-se à capacidade de uma ferramenta de interpretabilidade de fornecer explicações precisas para previsões individuais, refletindo corretamente o comportamento do modelo em cada instância específica; A propriedade **Ausência** garante que apenas as características relevantes sejam destacadas nas explicações, pois se uma característica não contribui para nenhuma coalizão, seu valor de Shapley é zero. E a **Consistência** garante que o valor de Shapley de uma característica não diminuirá se sua contribuição marginal aumentar ou permanecer a mesma em todas as coalizões possíveis, refletindo adequadamente as mudanças na importância da característica. ([LUNDBERG; LEE, 2017](#); [BAPTISTA; GOEBEL; HENRIQUES, 2022](#)) Estas três propriedades são satisfeitas pela única solução da equação acima usada para calcular os valores SHAP.

### 2.7.2 DeepSHAP

*Deep Learning Important Features* (DeepLIFT) é um mecanismo de *backpropagation* que opera camada por camada da saída até a entrada, estimando a contribuição de cada neurônio para a probabilidade gerada pela função logit (a inversa da função sigmoide) na camada final. Com imagens, o DeepLIFT avalia como as ativações da camada de saída respondem a uma mudança na imagem de entrada comparando-as com as ativações esperadas para a classe de imagem. Ele quantifica as discrepâncias entre os resultados esperados e as previsões do modelo, atribuindo pontuações de importância a cada píxel. Essas pontuações destacam os píxeis que mais influenciam a saída do modelo, fornecendo assim uma visão sobre a importância dos píxeis. ([CANÁRIO; RIBEIRO; RIOS, 2022](#)).

Deep SHAP (DeepLIFT + Shapley values) é uma técnica que combina DeepLIFT e valores de Shapley para fornecer explicações interpretáveis para os modelos de aprendizado profundo. O Deep SHAP calcula os valores de Shapley usando as atribuições de DeepLIFT como aproximação, mantendo a interpretação rigorosa dos valores de Shapley enquanto se beneficia da rapidez e escalabilidade do DeepLIFT ([LUNDBERG; LEE, 2017](#)).

## 2.8 Métricas

As métricas de avaliação são medidas quantitativas utilizadas para avaliar a qualidade de um modelo de Machine Learning em uma determinada tarefa, oferecendo uma medida objetiva de quão bem um modelo está performando em tarefas específicas. Elas fornecem uma maneira objetiva de comparar o desempenho do modelo com dados reais ou de validação e ajudam a entender como o modelo se comporta em diferentes situações.

### 2.8.1 Métricas Comuns

Existem várias métricas comuns usadas para avaliar modelos de Machine Learning, dependendo do tipo de problema e da natureza dos dados. Algumas das métricas mais amplamente utilizadas incluem:

- **Acurácia:** A proporção de previsões corretas em relação ao total de previsões feitas pelo modelo. É uma métrica simples e intuitiva, adequada para problemas de classificação binária ou multiclasse balanceados.

$$\text{Acurácia} = \frac{\text{Verdadeiros Positivos} + \text{Verdadeiros Negativos}}{\text{Total de Amostras}} \quad (2.9)$$

- **Precisão e Revocação:** Métricas frequentemente usadas em problemas de classificação desbalanceados, onde a precisão mede a proporção de instâncias positivas corretamente previstas entre todas as instâncias previstas como positivas, enquanto a revocação mede a proporção de instâncias positivas corretamente previstas entre todas as instâncias positivas reais.

$$\text{Precisão} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Positivos}} \quad (2.10)$$

$$\text{Revocação} = \frac{\text{Verdadeiros Positivos}}{\text{Verdadeiros Positivos} + \text{Falsos Negativos}} \quad (2.11)$$

- **F1-Score:** A média harmônica da precisão e revocação, fornecendo uma única métrica que equilibra ambos os aspectos. É útil quando há um desequilíbrio significativo entre as classes.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \quad (2.12)$$

- **Matriz de Confusão:** Uma tabela que descreve o desempenho do modelo em um problema de classificação, mostrando o número de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

		Valores verdadeiros		
		Positivo	Negativo	
Predição	Positivo	Verdadeiro positivos (VP)	Falso-positivos (FP)	total de Predições positivas
	Negativo	Falso-negativos (FN)	Verdadeiro negativos (VN)	total de Predições negativas
		Total de positivos	Total de negativos	

Figura 7 – Matriz de confusão de um problema de classificação binário

## 2.9 Considerações Finais

Os avanços na bioacústica são facilitados pelo crescente poder de processamento dos computadores e pela disponibilidade de dispositivos digitais de gravação de alta qualidade. A integração de métodos de aprendizagem profunda, adaptados ao reconhecimento de imagem, fala e áudio, está permitindo resolver problemas anteriormente considerados intratáveis no monitoramento ambiental. Abordagens ponta-a-ponta (E2E) otimizam processos de processamento de imagem e som, aumentando a eficiência e precisão dos resultados.

Além disso, a incorporação de técnicas de IA explicável, como SHAP e DepSHAP, contribui significativamente para a transparência e interpretabilidade dos modelos de aprendizado profundo aplicados à bioacústica. Isso permite maior confiança por parte dos especialistas, auxiliando-os em suas pesquisas e tornando os resultados mais compreensíveis e acionáveis.

Este estudo destaca a relevância dos métodos automatizados para a análise bioacústica e sua aplicação prática na compreensão dos impactos ambientais sobre espécies como a curruíra (*Troglodytes aedon*). A adoção dessas tecnologias não apenas aprimora a eficiência e precisão da análise bioacústica, mas também oferece novas oportunidades

para a pesquisa e conservação ambiental. Ressalta-se, portanto, a importância de uma abordagem interdisciplinar que une avanços tecnológicos e conhecimentos ecológicos, promovendo um monitoramento ambiental mais eficaz e uma conservação mais informada da biodiversidade.

# 3

## TRABALHOS RELACIONADOS

Nesse capítulo, serão apresentados trabalhos que exploram diferentes aspectos do uso de aprendizado profundo em bioacústica, os efeitos do ruído antropogênico nas curruíras (*Troglodytes aedon*), e a importância da explicabilidade nos modelos de aprendizado profundo aplicados à bioacústica. A análise desses tópicos é fundamental para entender o estado da arte e as lacunas existentes na pesquisa atual.

A seção 3.1 aborda o aprendizado profundo em bioacústica, em particular as CNNs, destacando os avanços e desafios dessa abordagem na análise de dados acústicos de vida selvagem. Serão discutidos as vantagens dessa abordagem, assim como os resultados obtidos em diferentes estudos.

A seção 3.2 foca na explicabilidade na bioacústica, apresentando pesquisas que utilizam ferramentas como SHAP para interpretar e entender as decisões tomadas por modelos de aprendizado profundo. A importância de proporcionar transparência e interpretabilidade nos modelos é discutida, bem como a aplicação prática dessas técnicas na área de bioacústica.

Por fim, na seção 3.2 é explorado o impacto do ruído antropogênico nas curruíras (*Troglodytes aedon*). Esta seção examina como o ruído causado pela atividade humana afeta o comportamento vocal dessas aves e as adaptações que elas desenvolvem em resposta a esses distúrbios ambientais.

### 3.1 Aprendizagem profunda em Bioacústica

A aprendizagem profunda revolucionou vários domínios, incluindo reconhecimento de fala, processamento de linguagem e tarefas multimodais (MCLOUGHLIN; STEWART; MCELLIGOTT, 2019). Este sucesso transformador também se estendeu ao campo da bioacústica computacional, onde as técnicas de aprendizagem profunda têm se mostrado muito promissoras no avanço da análise e compreensão das vocalizações animais(MUTANU et al., 2022b; STOWELL, 2022; KVSN et al., 2020).

Na bioacústica computacional, os pesquisadores estão explorando a aplicação de algoritmos de aprendizagem profunda, como redes neurais profundas (DNNs), redes neurais recorrentes (RNNs) e redes neurais convolucionais (CNNs), para enfrentar vários desafios (AGGARWAL; HASIJA, 2022; THAKUR et al., 2019; MUTANU et al., 2022a). A popularidade dessas técnicas deve-se ao seu desempenho superior em comparação com outras abordagens. Para todas as configurações de tarefas (classificação, detecção e agrupamento), os métodos de aprendizagem profunda baseados em CNN apresentam melhor desempenho, superando outras técnicas de aprendizado de máquina (MARCHAL; FABIANEK; AUBRY, 2022; KNIGHT et al., 2017; PRINCE et al., 2019).

Uma das principais vantagens da aprendizagem profunda neste domínio é a sua capacidade de aprender representações robustas de características de grandes conjuntos de dados, o que pode levar a um melhor desempenho em tarefas como classificação de espécies animais, detecção de chamadas e localização de fontes sonoras (KVSN et al., 2020). A facilidade de adaptação também deve ser mencionada, com grande uso de redes prontas e pré-treinadas, como ResNet, VGG, DenseNet e MobileNet, para tarefas de classificação bioacústicas (MUTANU et al., 2022a). Algumas redes prontas são facilmente adaptadas para a tarefa (ResNet e DenseNet), e algumas arquiteturas são projetadas visando eficiência (MobileNet, EfficientNet, Xception), reduzindo a quantidade de cálculos sem perder precisão (CANZIANI; PASZKE; CULURIELLO, 2016).

As abordagens de aprendizagem profunda têm o potencial de superar as limitações do processamento tradicional de sinais e dos métodos estatísticos, que exigem

muitas vezes extensa engenharia de recursos e conhecimento específico de domínio (Hannun et al., 2014; Deng et al., 2013). Outra das vantagens dos modelos de aprendizagem profunda é a capacidade de aprender diretamente dos dados brutos, permitindo aos modelos selecionar as características do som adequadas para a tarefa, evitando a necessidade de extração manual de características e vieses (SANCHEZ et al., 2021). Apesar das vantagens, esses modelos exigem abundância de dados bioacústicos brutos rotulados (BERMANT et al., 2019) e a falta de dados de treinamento geralmente leva ao *overfitting* nas CNNs (THAKUR et al., 2019). A falta de grandes conjuntos de dados rotulados também é um problema frequente em bioacústica. Espécies raras ou difíceis de encontrar, ou capturar, impedem obter quantidade significativa de exemplos de áudios; eventos sonoros que exijam especialistas para rotulagem dificultam a criação desses rótulos (STOWELL, 2022).

O monitoramento contínuo de larga escala almeja coletar grandes volumes de dados para viabilizar o treinamento de modelos mais profundos, porém o volume gerado excede a quantidade de especialistas disponíveis para rotular. Por exemplo, a captura de 24 horas consecutivas de áudio ambiental implica que um especialista deveria inspecionar essas gravações. Se ele tiver que ouvir as gravações para identificar os animais presentes, ele precisaria de outras 24 horas para rotular, tornando a tarefa inviável. Assim, soluções que produzem resultados eficazes, mesmo com a disponibilidade de apenas poucos exemplos de treinamento, devem ser exploradas (STOWELL, 2022; MOLNÁR et al., 2008; MUTANU et al., 2022a).

Várias estratégias foram propostas para mitigar a escassez de dados usáveis ou falta de rótulos, como *Data Augmentation* e *Embeddings*. Essas técnicas são úteis para permitir o treinamento de conjuntos de dados com limitações, com resultados bem estabelecidos (LASSECK, 2018; LI et al., 2021; PADOVESE et al., 2021; ZHONG et al., 2020; KAHL et al., 2021). Além dessas técnicas, o *transfer learning* permite a transferência de conhecimento de uma rede pré-treinada para o domínio e a tarefa de interesse. Em cenários de escassez de dados, a transferência de aprendizagem de redes existentes é mais fácil e mais eficaz do que treinar a rede a partir do zero (LASSECK, 2018; PROVOST; YANG; CARSTENS, 2022).

Apesar da disponibilidade de dados gerados pelo monitoramento passivo para bioacústica, o estudo sobre pequenos conjuntos de dados permanece como desafio. Tarefas envolvendo discriminação de alta resolução (como identificação de animais da mesma espécie ou com sonoridade muito parecida) e tarefas onde a transferência de aprendizado não é suficiente (devido ao viés, por exemplo) continuarão surgindo, abrindo espaço para novas técnicas e integração dessas técnicas, já que pré-treinamento, aprendizado multitarefa e aumento de dados oferecem baixo risco para uma generalização aprimorada ([MORFI; LACHLAN; STOWELL, 2021](#); [STOWELL, 2022](#)).

### 3.2 Explicabilidade na bioacústica

A IA explicável (XAI) utiliza ferramentas de visualização interpretáveis para elucidar como os métodos de caixa preta realizam a classificação, melhorando a compreensão do processo de tomada de decisão ([CANÁRIO; RIBEIRO; RIOS, 2022](#)). Devido à complexidade e ao grande número de parâmetros inerentes aos modelos DL, existe uma necessidade crescente de ferramentas de interpretabilidade que auxiliem os usuários a entender como esses modelos operam ([DAS et al., 2024a](#); [CHACÓN, 2018](#)).

Isso permite melhora da confiança, a responsabilidade e a tomada de decisões. Alguns setores são favorecidos pela transparência e credibilidade proporcionadas pela compreensão dos modelos de inteligência artificial, especialmente em cenários sensíveis, como saúde, carros autônomos, sistemas financeiros, segurança pública e combate ao terrorismo ([GOMES et al., 2023](#); [ORTIGOSSA; GONÇALVES; NONATO, 2024](#)).

Na classificação bioacústica, a interpretabilidade e explicabilidade pode nos ajudar a compreender os padrões de sinal e as características específicas de espécies, como mudanças de frequência e mudanças de tempo, que levam um modelo DL a uma decisão específica. Isso destaca quais padrões nos registros de áudio brutos são mais relevantes para o modelo. Para garantir a confiabilidade das ferramentas de interpretabilidade, o desempenho dos modelos DL deve ser validado e comparações com explicações geradas por humanos também podem ajudar a identificar os pontos fortes e fracos das abordagens XAI aplicadas à bioacústica ([DAS et al., 2024a](#); [CHACÓN,](#)

2018).

Apesar das vantagens que a XAI pode trazer à bioacústica computacional, há poucos trabalhos nessa área, com recentes trazendo para a bioacústica a explicabilidade, ao analisar como as abordagens de XAI aumentam a interpretabilidade dos dados, facilitando para os pesquisadores e biólogos a compreensão dos motivos pelos quais determinadas espécies foram selecionadas por classificadores.

Um dos primeiros trabalhos na linha, feito por [FONSECA et al.](#) utilizou uma abordagem E2E (modelo CNN e Kapre) com uma alta acurácia (80%) para identificar mosquitos de várias espécies que transmitem doenças tropicais (banco de dados Wings-beat) usando o som da batida das asas, e utilizou o SHAP para mostrar as diferenças de padrão de frequência entre os sons emitidos. Este trabalho é de coautoria do autor dessa dissertação, publicado em 2023 e apresenta protótipo de visualizações bioacústicas propostas nessa dissertação utilizando valores SHAP.

O trabalho de [AKBAL et al.](#) utilizou 1D-LBP (Local Binary Pattern) e TQWT (Tuned Q-factor Wavelet Transform) para trazer explicabilidade para o trabalho de classificação de Anuros. Ele extraiu 270 características de cada som de anuro e gerou um vetor de *features* para classificar e explicar a classificação. Esse método proposto de classificação sonora de anuros baseado em 1D-LBP e TQWT alcançou uma precisão de classificação de 99,35%.

O trabalho de [DAS et al.](#) usou um modelo CNN com 92% de acurácia para classificar várias espécies de pássaros do banco de dados Xeno-Canto e avaliou empiricamente duas metodologias, o LIME (Local Interpretable Model-agnostic Explanations) e SHAP (SHapley Additive exPlanations) para determinar a interpretabilidade do modelo proposto. Os resultados mostraram que o SHAP teve um desempenho ligeiramente melhor do que o LIME em termos de identidade, estabilidade e separabilidade. Apesar disso, o uso do SHAP se restringiu a uso de MFCCs, sendo utilizado nessa pesquisa um conjunto de características do som, e não o próprio som.

### 3.3 Efeito de ruído antropogênico em *Troglodytes Aedon*

As aves estão entre as espécies mais documentadas em bioacústica devido ao canto que reflete o seu ambiente e os impactos que enfrentam (EBERLY ABRAM FLEISHMAN, 2020; CHOI; JOO; KIM, 2017b; MUTANU et al., 2022b). Eles podem alterar ou modular seus chamados para se comunicar durante as interações de acasalamento, defender-se contra intrusos territoriais ou responder a mudanças ambientais (SEMENTILI-CARDOSO; DONATELLI, 2021).

Considerando o ruído ambiental em áreas urbanas, algumas espécies de aves modificam seus sinais acústicos como resposta adaptativa ao tipo e nível de ruído. Alterações na frequência e duração de suas músicas e chamadas podem ajudá-los a evitar ruídos, mas essas mudanças também podem limitar outras características acústicas que transmitem informações importantes sobre o remetente (MUTANU et al., 2022b).

Estudos apontam que o ruído antropogênico, como o ruído dos motores e outras atividades humanas, contribui significativamente para o declínio das espécies de aves, impedindo-as de comunicar adequadamente. Esse ruído artificial mascara suas vocalizações, impede a comunicação e aumenta os níveis de estresse. A falta de comunicação ou o aumento do stress reduzem as suas taxas de reprodução, acelerando assim o declínio populacional. Espécies com vocalizações de baixa frequência têm maior probabilidade de serem afetadas. (PROPPE; STURDY; CLAIR, 2013; GRABARCYK; GILL, 2019; SEMENTILI-CARDOSO; DONATELLI, 2021).

Dentre as espécies observadas, a corruíra (*Troglodytes Aedon*) se destaca por apresentar características distintas em vocalização e frequência quando encontrado em ambientes altamente impactados (GRABARCYK; GILL, 2019; ??; STOWELL, 2022). Estudos em populações da América Central e do Norte revelam que o ruído antropogênico perturba a transmissão do canto das corruíras machos (??). Além disso, os machos das populações do norte ajustam as suas vocalizações em resposta à urbanização e ao ruído antropogênico (??SANDOVAL; REDONDO; BARRANTES, 2013; CADIEUX et al., 2020).

Nas populações sul-americanas, as corruíras machos cantam em amplitudes

maiores quando o ruído do trânsito é maior ([SEMENTILI-CARDOSO; DONATELLI, 2021; PERIS; COLINO-RABANAL, 2016](#)). Juarez et al. ([JUAREZ et al., 2021](#)) indica que em ambientes mais barulhentos, os machos desta espécie tendem a incorporar mais elementos de alta frequência em seus cantos.

Além disso, [SANDOVAL; REDONDO; BARRANTES](#) demonstram que essas mudanças impactam a estrutura fundamental da vocalização, bem como elementos de frequência e amplitude. Nas áreas rurais, as corruíras produzem trinados com uma frequência mínima mais baixa e uma largura de banda de frequência mais ampla, que diminui com o aumento dos níveis de ruído. Em contraste, em ambientes urbanos, corruíras cantam trinados mais curtos. À medida que os níveis de ruído aumentam, a largura de banda do trinado diminui, enquanto a taxa aumenta para manter o desempenho.

Grupo de corruíras de diferentes locais apresentam diferentes estratégias para sobrepor o ruído. Por exemplo, as corruíras no Caribe cantam introduções mais curtas, trinados mais rápidos e aumentam as frequências baixas da introdução e da música inteira para compensar a urbanização ([CYR KIMBERLEY WETTEN; KOPER, 2021](#)).

Esses estudos fornecem evidências convincentes dos impactos do ruído antropogênico nas vocalizações e no comportamento territorial das corruíras domésticas. Eles abrangem aspectos comportamentais mais amplos e detalhes mais sutis das características da música, observados por meio de análises acústicas e estatísticas. No entanto, a maioria desses estudos se baseou em abordagens manuais e algoritmos tradicionais de aprendizado de máquina. Isto destaca a necessidade de abordagens mais modernas e automáticas para discriminar melhor as mudanças de vocalização entre os indivíduos, continuando a ser um desafio em aberto ([EBERLY ABRAM FLEISHMAN, 2020; MUTANU et al., 2022b](#))

### 3.4 Considerações Finais

Os trabalhos sobre o impacto do ruído antropogênico nas vocalizações das corruíras (*Troglodytes aedon*) revela diversas mudanças notáveis e indicam que, em ambientes

urbanos com altos níveis de ruído, as corruíras ajustam suas vocalizações de várias maneiras para melhorar a comunicação e evitar o mascaramento do som. Entre as principais mudanças destacadas estão o aumento da amplitude e a incorporação de frequências mais altas. Além disso, em áreas com maior poluição sonora, os trinados tendem a se tornar mais curtos, com uma largura de banda de frequência reduzida. Esses ajustes são feitos para manter a eficiência da comunicação, mas podem comprometer outras características acústicas importantes. Estudos também mostram que em regiões rurais, onde os níveis de ruído são menores, as corruíras produzem vocalizações com uma frequência mínima mais baixa e uma largura de banda de frequência mais ampla.

Apesar das evidências estatísticas e matemáticas dessas variações, ainda não há trabalhos utilizando aprendizado profundo ou outras técnicas de aprendizado de máquina que tragam esses resultados. Embora o aprendizado profundo seja amplamente utilizado para a classificação e análise geral de vocalizações animais, os trabalhos focados na explicação detalhada das variações nas vocalizações de cada indivíduo são escassos, indicando uma lacuna em análises detalhadas das mudanças específicas do canto de cada indivíduo.

Os estudos existentes abordam principalmente variações gerais entre populações ou espécies, e há necessidade de mais pesquisas que se concentrem na discriminação e interpretação das mudanças individuais nas vocalizações, utilizando técnicas modernas de aprendizado profundo e explicabilidade. Já há abordagens que usem o E2E para facilitar o preprocessamento, e essa é a metodologia utilizada neste trabalho de dissertação. O SHAP, ferramenta popular de forte base matemática e bons resultados também foi usado. O uso de SHAP também está presente entre os trabalhos recentes, apresentando desempenho ligeiramente melhor que o LIME.

# 4

---

## METODOLOGIA

Esta seção detalha a metodologia utilizada neste trabalho. A seção 4.1 descreve o conjunto de dados sobre a curruíra (*Troglodytes aedon*), abordando o local de coleta e as características específicas dos registros de áudio. A seção 4.2 discute o processamento dos áudios, elucidando os métodos aplicados para preparar os dados brutos para o modelo de aprendizado profundo. Na seção 4.3 mostramos o uso do modelo ponta-a-ponta (E2E), detalhando a arquitetura da rede neural escolhida (Biophony) e ajustes finos feitos para alcançar o melhor desempenho. Por fim, a seção 4.4 apresenta a abordagem de explicabilidade com SHAP (Shapley Additive Explanations), detalhando o framework usado e como foi aplicado para interpretar as decisões do modelo e entender a influência das diferentes características nas previsões.

### 4.1 Conjunto de Dados Corruíra

Nossos experimentos utilizaram o conjunto de dados "Curruíra" (*Troglodytes aedon musculus*), compilado por Sementili Cardoso e Donatelli ([SEMENTILI-CARDOSO; DONATELLI, 2021](#)). Este conjunto de dados inclui gravações de 50 machos de curruíra em uma área periurbana de Bauru, São Paulo, Brasil, entre julho e dezembro de 2017. Os biólogos selecionaram machos com territórios distintos, garantindo uma distância mínima de 100 metros entre eles.

As curruíras foram divididas em dois grupos: um grupo de 25 indivíduos localizados a até 500 metros de uma rodovia, e outro grupo de 25 indivíduos localizados

a pelo menos 1.500 metros da rodovia. No total, foram coletadas 889 amostras de vocalizações (média:  $16 \pm 5$  amostras por indivíduo, intervalo: 7–35 amostras).

Este conjunto de dados oferece uma coleção abrangente de vocalizações completas, apesar de seu tamanho reduzido.

## 4.2 Processamento dos áudios

Cada gravação de áudio foi dividida em segmentos de aproximadamente dois segundos, dependendo da duração da vocalização. Se a vocalização fosse mais curta que dois segundos, uma abordagem de preenchimento (*padding*) circular foi adotada para completar a duração de dois segundos. O preenchimento circular repete a parte inicial da gravação no final para evitar preenchimento com zeros. Das 889 amostras, 844 foram completadas com preenchimento circular. Foi utilizado um filtro passa-banda, restringindo as gravações entre 1,5 kHz e 10 kHz para focar nas vocalizações das curruíras, ajudando a remover ruídos ambientais fora das principais bandas de frequência. As gravações foram reamostradas de 44.100 Hz para 22.050 Hz.

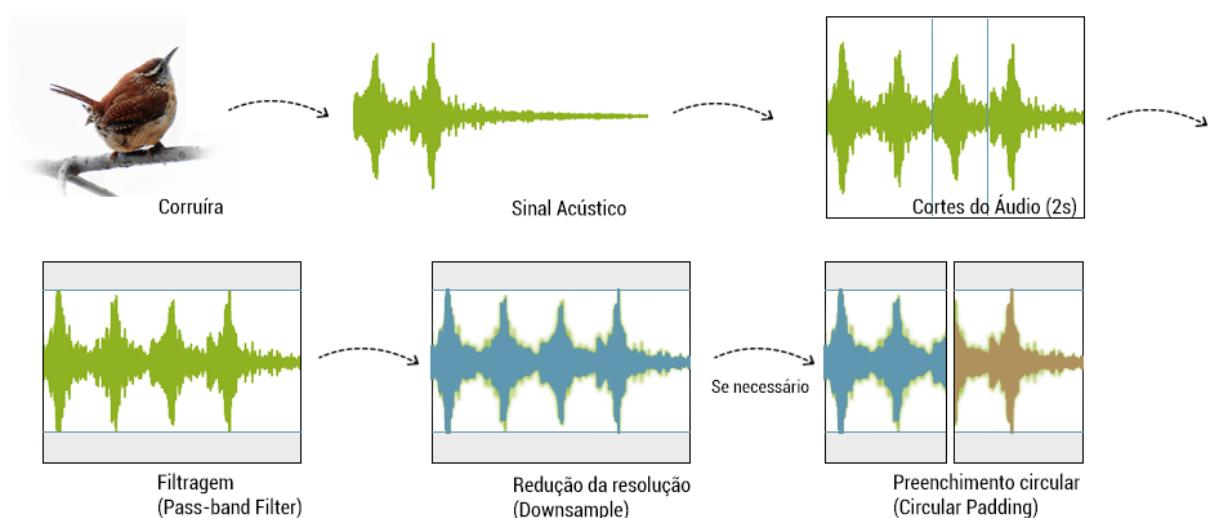


Figura 8 – Processamento das amostras de áudio do banco de dados "curruíra".

### 4.3 Modelo E2E (Biophony e Kapre)

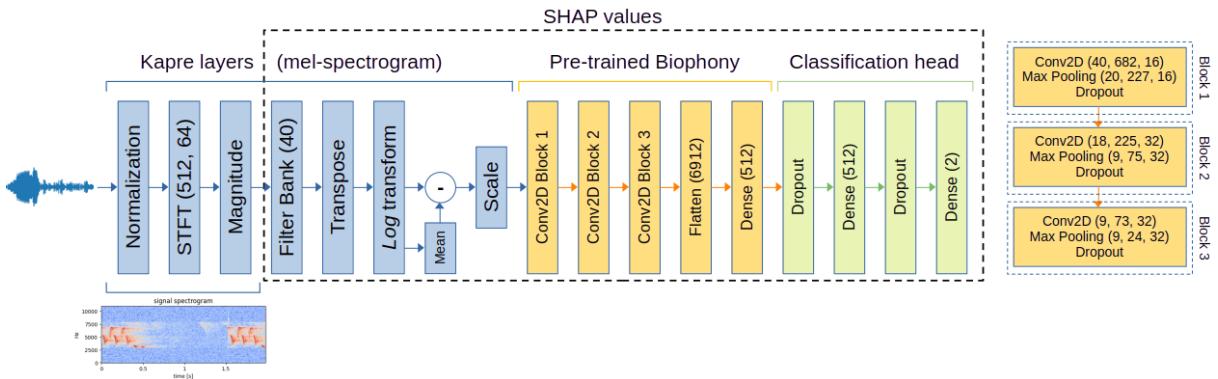


Figura 9 – Modelo E2E composto por camadas Kapre e o modelo Biophony. Os valores SHAP são obtidos a partir do espectrograma, antes de convertê-lo para a escala Mel.

Após processar todos os áudios, aplicou-se a abordagem E2E utilizando Biophony e Kapre para a classificação, como retratado na imagem. Para obter explicabilidade, foi utilizado o Deep Explainer para obter os valores SHAP a partir da classificação dos espectrogramas gerados para o biophony, como mostra a imagem 9.

**Kapre:** A abordagem proposta pelo Kapre segue os passos descritos: i) Decodificar (e possivelmente gerando amostras) arquivos de áudio e salvá-los como formatos binários; ii) implementar um gerador para carregar os dados e iii) adicionar uma camada Kapre ao lado de entrada do modelo Keras. Para processar áudio nas camadas Keras, a camada `get_melspectrogram_layer`, estendida baseada em `Spectrogram` com multiplicação por matriz de conversão em escala Mel, é utilizada.

Tabela 2 – Parâmetros da função `get_melspectrogram_layer`

Parâmetros	Descrição
<code>n_mels</code>	Número de recipientes de Mel no FilterBank Mel
<code>n_fft</code>	Número de pontos FFT em STFT
<code>win_length</code>	Comprimento da janela de STFT
<code>hop_length</code>	Comprimento do salto de STFT
<code>mel_f_max</code>	Frequência mais alta do FilterBank Mel
<code>return_decibel</code>	Se deve aplicar a escala de decibéis no final

A camada `FilterBank` providencia uma camada de filtro que pode ser inicializada com escalas de frequência Mel/log/linear ou aleatório.

```
get_melspectrogram_layer(n_mels=41, n_fft=1024, win_length=1150,
hop_length=63, mel_f_max=11025, return_decibel='True')
```

Figura 10 – Um trecho de código que retorna a camada de espectrograma Mel

Através das camadas Kapre, as gravações foram convertidas em um espectrograma Mel para serem usadas como imagem de entrada para a primeira camada do modelo Biophony. A cabeça de pré-processamento usando Kapre inclui uma janela Hann. A camada de espectrograma foi configurada para usar 60 ms (512 pontos de tempo) com um tamanho de salto de 64 ms (256 pontos de tempo) e 128 bins de frequência Mel, resultando em imagens de espectrograma de tamanho 40 x 682 píxeis.

```
duration = 1.0 # audio clip duration
num_channel = 1 # number of channels
input_shape = (int(input_shape[1]), num_channel) # channels_last convention
preemph = 0.5

model = models.Sequential()

model.add(get_melspectrogram_layer(n_mels=40, n_fft=1024, win_length=512, hop_length=6, mel_f_max=sample_rate/2,
                                   input_data_format='channels_last', output_data_format='channels_last',
                                   return_decibel=False, # decibel false funciona melhor
                                   input_shape=input_shape))

a) KAPRE

model.add(layers.Permute((2,1,3)))

#Modelo Biophony
res_path = "/content/drive/MyDrive/QualiLas/resources/"
model_b = model_from_config(json.load(open(res_path+'cmi_mbam01.json', 'r')))
model_b.load_weights(res_path+'cmi_mbam01.h5')
model_b = Model(inputs=model_b.input, outputs=model_b.layers[-4].output)
model_b.summary()

b) Biophony (Transferência de Aprendizagem)
```

Figura 11 – Um trecho de código com a) KAPRE e b) Biophony

**Biophony:** No modelo utilizamos um total de 5 camadas pre-treinadas (3 de convolução, 1 *flatten* e 1 de *Dropout*). Na cabeça da classificação possuem 4 camadas, com duas densas, uma possui 512 neurônios e a outra duas, enquanto as outras são de *Dropout*. A camada final, ou de saída, consiste em 2 rótulos, correspondendo a classes específicas. Todas as camadas internas utilizam a ativação ReLU, enquanto a camada de saída utiliza a ativação Softmax. Como parte do treino, temos os pesos já obtidos do treinamento com o Xeno-Canto do biophony original.

**Fine-tuning:** O processo de *fine-tuning* foi realizado utilizando validação cruzada

*K-Fold* visando maximizar a robustez e a generalização do modelo E2E desenvolvido para a classificação do corruíra. Para tanto, o conjunto de dados foi dividido em cinco subconjuntos, sendo que em cada iteração quatro subconjuntos foram utilizados para treinamento e um para teste, garantindo assim uma avaliação abrangente do desempenho do modelo.

Inicialmente, os dados foram preparados e divididos em cinco partes (folds) com o uso da técnica de validação cruzada *K-Fold*. Esta abordagem permite que cada amostra do conjunto de dados seja utilizada tanto para treinamento quanto para validação, proporcionando uma avaliação mais confiável do modelo. Em cada uma das cinco iterações, um fold diferente foi reservado para validação enquanto os outros quatro foram utilizados para treinar o modelo.

Para aprimorar a eficácia do treinamento, utilizamos dois callbacks importantes: *EarlyStopping* e *ModelCheckpoint*. O *EarlyStopping* monitorou a métrica de precisão na validação (`val_accuracy`) e interrompeu o treinamento se não houvesse melhoria após cinco épocas consecutivas, prevenindo assim o *overfitting* e garantindo o melhor desempenho do modelo. Simultaneamente, o *ModelCheckpoint* salvou os pesos do modelo que apresentaram a melhor precisão na validação, garantindo que o melhor estado do modelo fosse preservado.

Foi usado `models.build_model(kapre_model, biophony_model)`, que integra módulos pré-treinados (Kapre e Biophony) com novas camadas específicas para a tarefa de classificação de vocalizações. Durante o treinamento, o modelo foi ajustado com um conjunto de dados de treinamento e validado utilizando 20% dos dados de treinamento reservados para validação. Este processo foi repetido para cada um dos cinco folds, permitindo uma avaliação completa e robusta da eficácia do modelo.

Após o treinamento de cada iteração, o modelo realizou previsões sobre os dados de teste, e os resultados foram armazenados para posterior análise. A combinação de validação cruzada *K-Fold*, callbacks de *early stopping* e *checkpointing*, juntamente com o fine-tuning dos modelos pré-treinados, proporcionou um método eficiente e robusto para a classificação de vocalizações da espécie estudada, maximizando a precisão e a capacidade de generalização do modelo desenvolvido.

Após o processo de validação cruzada K-Fold, realizamos um treinamento final do modelo utilizando todos os dados de treinamento disponíveis, com uma divisão interna para validação (*validation split*). Este treinamento final visou consolidar o aprendizado do modelo, garantindo que ele estivesse otimizado para a tarefa de classificação.

#### 4.3.1 Verificando se há viés em relação ao ruído de fundo

Para garantir que a abordagem usada dependa exclusivamente das vocalizações do Corruíra e seja imparcial em relação a ruídos de fundo (por exemplo, carros passando nas gravações), analisamos trinta amostras com somente ruído de fundo do banco de dados, abrangendo segmentos de áreas impactadas e não impactadas.

Vetores de probabilidade foram gerados para essas amostras usando o modelo proposto. A comparação desses vetores com aqueles das vocalizações de House Wren foi realizada utilizando o teste pareado de Wilcoxon, uma ferramenta estatística não paramétrica utilizada para comparar duas amostras relacionadas. O teste de Wilcoxon é particularmente adequado para dados que não seguem uma distribuição normal, oferecendo uma alternativa robusta ao teste *t* pareado em tais circunstâncias.

O teste pareado de Wilcoxon revelou diferenças significativas entre os vetores de probabilidade gerados para os ruídos de fundo e as vocalizações de House Wren, com valores de  $p = 1,21 \times 10^{-5}$  e  $p = 1,86 \times 10^{-9}$  para impactados e não impactados, respectivamente.

Esses resultados afirmam que o modelo usado neste trabalho distingue eficazmente as vocalizações da House Wren e não é tendencioso devido às particularidades do ruído de fundo. A utilização do teste de Wilcoxon reforça a robustez dos nossos achados, evidenciando que as diferenças observadas não são fruto do acaso, mas sim indicativas de uma capacidade clara do modelo em identificar e discriminar particularidades específicas às vocalizações, independentemente da presença de ruídos ambientais no fundo das gravações.

## 4.4 SHAP

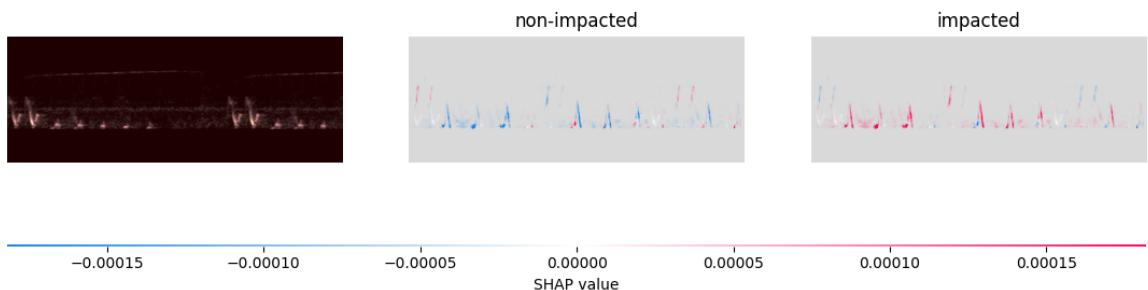


Figura 12 – Amostra de vocalização de corruíra e os valores SHAP para a amostra. Esta representação mostra a distribuição dos valores SHAP em dois quadros ao lado do espectrograma; a intensidade da cor é definida pelos valores de contribuição, quanto mais intenso o vermelho, maior a contribuição positiva, e quanto mais intenso o azul, maior a contribuição negativa; na parte de baixo da figura, uma régua revelando a escala de valores de contribuição para essa amostra

Após treinar o modelo, foi utilizado o SHAP, uma biblioteca Python de código aberto para IA explicável (XAI) para SHAP proposto por [LUNDBERG; LEE](#). Utilizamos o `shap.DeepExplainer` para criar um explicador SHAP, fornecendo como entrada o modelo treinado e as amostras. O método **DeepExplainer** é uma versão aprimorada do algoritmo DeepLIFT implementado utilizando os valores SHAP. Apesar de ser um explicador local, o SHAP também pode fornecer uma percepção global sobre todo o banco de dados ([MOLNAR, 2022](#); [LUNDBERG; LEE, 2017](#)), e para isso, foi calculado os valores SHAP para todos os espectrogramas de uma só vez para obter uma visão geral da importância das características em todo o conjunto de dados. Na figura 12, mostra como o modelo classificou uma amostra da vocalização do corruíra. além do espectrograma, há também a visualização da distribuição dos valores SHAP ao lado do espectrograma original, mostrando as contribuições positivas ou negativas conforme a intensidade do azul(contribuições negativas) ou vermelho(contribuições positivas).

Para criar as novas visualizações, foram utilizados os valores SHAP gerados pelo explainer SHAP, e combinado com os espectrogramas, foram criadas melhores formas de visualização para as características observadas, sendo uma das contribuições-chave deste trabalho.

# 5

---

## RESULTADOS

Nesta seção, serão discutidos os resultados do modelo Biophony e Kapre para a classificação das vocalizações da corruíra em ambiente impactado ou não impactado.

### 5.1 O Modelo

O modelo Biophony foi treinado utilizando o callback EarlyStopping, configurado com uma paciência de cinco épocas e um valor  $\Delta$  (delta) de 0,00001. Esta configuração interrompe o processo de treinamento quando a perda de validação deixa de melhorar, resultando em um total de 5 épocas de treinamento. A estabilidade observada nas curvas durante o treinamento, conforme mostrado na Figura 13, sugere que o modelo está efetivamente aprendendo com os dados, sem apresentar *overfitting* (memorização do conjunto de treinamento).

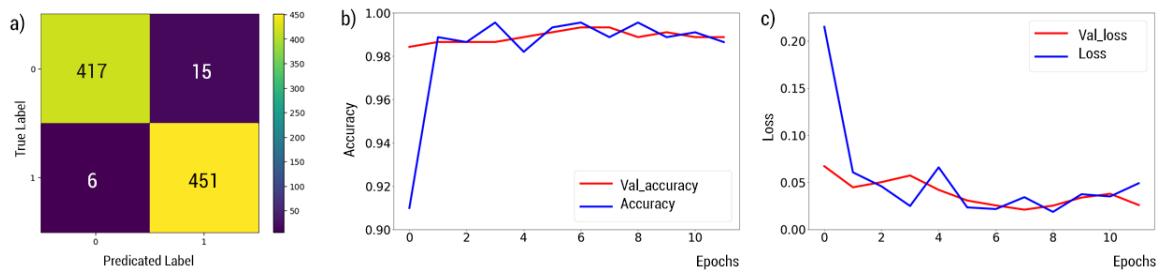


Figura 13 – a) Matriz de confusão e b) Curvas de aprendizagem: Acurácia; c) Loss

Tabela 3 – Métricas: Precisão, Revocação e F1-Score

Class	Precision	Recall	F1-score	Support
<b>Non-Impacted</b>	0.99	0.98	0.98	432
<b>Impacted</b>	0.98	0.99	0.98	457
<b>Accuracy</b>			0.98	889
<b>Macro avg</b>	0.98	0.98	0.98	889
<b>Weighted avg</b>	0.98	0.98	0.98	889

Além disso, a Figura 13 oferece uma visão sobre o desempenho do modelo, revelando a matriz de confusão gerada pelo modelo final. A matriz de confusão destaca a precisão do modelo em classificar corretamente as vocalizações do Corruíra. Complementando, o relatório de classificação do conjunto de testes apresenta métricas detalhadas, como precisão, revocação e F1-score, evidenciando a capacidade do modelo de generalizar eficazmente e fazer previsões precisas sobre dados não vistos pelo modelo anteriormente. Esses resultados sublinham a eficácia do modelo Biophony na identificação e classificação das vocalizações da espécie estudada.

## 5.2 Explicabilidade do Modelo

Para trazer explicabilidade para os resultados do modelo utilizamos o *DeepExplainer*. Gráficos SHAP foram produzidos para uma instância de dados classificados, nos quais observamos as escolhas do modelo. A ideia é determinar o nível de contribuição de cada píxel na imagem prevista.

Para explicação, os valores SHAP e o gráfico de resultados para as duas classes (corruíra de ambiente impactado e corruíra de ambiente não impactado) são mostrados na Fig 14. A primeira imagem é uma imagem original do conjunto de testes, a segunda imagem e a terceira imagem mostram a interpretação do SHAP e fornecem informações sobre o resultado da previsão.

As imagens destacam a explicação do SHAP em diferentes tons de vermelho e azul (no fundo, pode-se ver a imagem original). As etiquetas nas imagens exibem a classe prevista, com píxeis vermelhos significando valores SHAP com contribuição positiva para a classificação da classe rotulada, enquanto píxeis azuis representam

contribuições que impactam negativamente a previsão. Quanto mais intensa a tonalidade de vermelho ou azul, maior a contribuição para a classificação de uma classe. Assim, áreas mais escuras sugerem uma contribuição mais significativa do que as mais claras. Isso é mostrado pela barra de cores, que mapeia os valores SHAP associados à tonalidade da cor.

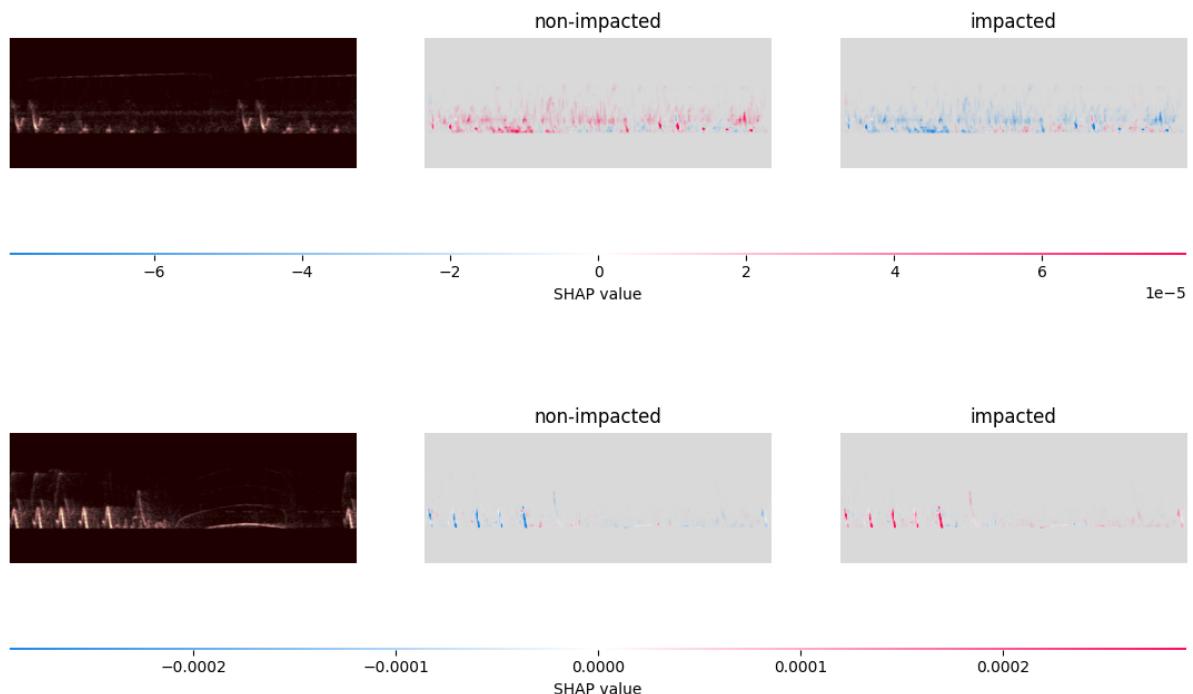


Figura 14 – Interpretação SHAP para a) Canto do corruíra em ambiente não-impactado.  
b) Canto do corruíra em ambiente impactado.

Para explicar a interpretação do SHAP em mais detalhes, consideremos o primeiro resultado na Figura 14, onde é exibido um espectrograma do som de uma corruíra em um ambiente não impactado. A interpretação destacou a imagem em tons de vermelho e azul segundo os valores SHAP. As partes mais escuras foram destacadas em vermelho nas áreas de frequência, indicando por que o modelo classifica o som como "não impactado". Além disso, há mais áreas em vermelho, mas com uma tonalidade mais clara. Em contraste, algumas frequências mais baixas se destacam em azul, sugerindo que esses píxeis contribuíram negativamente para o resultado.

Na segunda imagem, rotulada como "impactado", os píxeis destacados têm exatamente as cores opostas nos mesmos locais. Essa observação indica que, se a imagem

tivesse sido classificada como "impactada", então as partes da imagem consideradas realmente "relevantes" nesse contexto contribuiriam negativamente para a previsão. No entanto, a imagem ainda é classificada como "não impactada", pois as contribuições das frequências mais altas são predominantes. Esta dualidade é típica de um classificador binário, se fosse multi classe haveria diferenças entre as atribuições de importâncias no pixels e não somente a inversa destes para cada rótulo.

Na imagem da classe "impactada" na Figura 14, por exemplo, observa-se a predominância do azul, deixando claro que o modelo de ML a classificaria como impactada. Aqui, as frequências foram destacadas em azul, o que significa que contribuíram negativamente. Ao observar imagens de amostras interpretadas com SHAP, tem-se uma ideia de quais características o modelo de ML utiliza para fazer sua previsão de classificação.

### 5.2.1 Espectrograma SHAP

A visualização do espectrograma SHAP permite uma comparação direta entre o espectrograma original e os valores SHAP atribuídos a ele, oferecendo insights sobre a distribuição dos valores SHAP em todo o domínio tempo-frequência-energia do sinal. Ao sobrepor o espectrograma SHAP com o espectrograma original, pode-se avaliar visualmente como as regiões destacadas pelos valores SHAP se correlacionam com a intensidade e dispersão da energia espectral em cada registro amostral dos espécimes. Valores elevados de SHAP indicam contribuições substanciais de padrões específicos de tempo-frequência para a previsão de saída do modelo, enquanto valores baixos de SHAP em regiões específicas sugerem influência mínima na previsão do modelo.

Ao falar sobre a distribuição de energia nos cantos e trinados da corruíra, eles tendem a frequências mais baixas, aderindo a um padrão de distribuição log-normal. Além disso, os trinados geralmente manifestam-se com amplitude aumentada e uma largura de banda de frequência mais estreita em comparação com outros segmentos do canto. Em particular, nosso modelo aprendeu sobre os trinados, observando todos os exemplos e a alta contribuição (positiva e negativa) nos trinados.

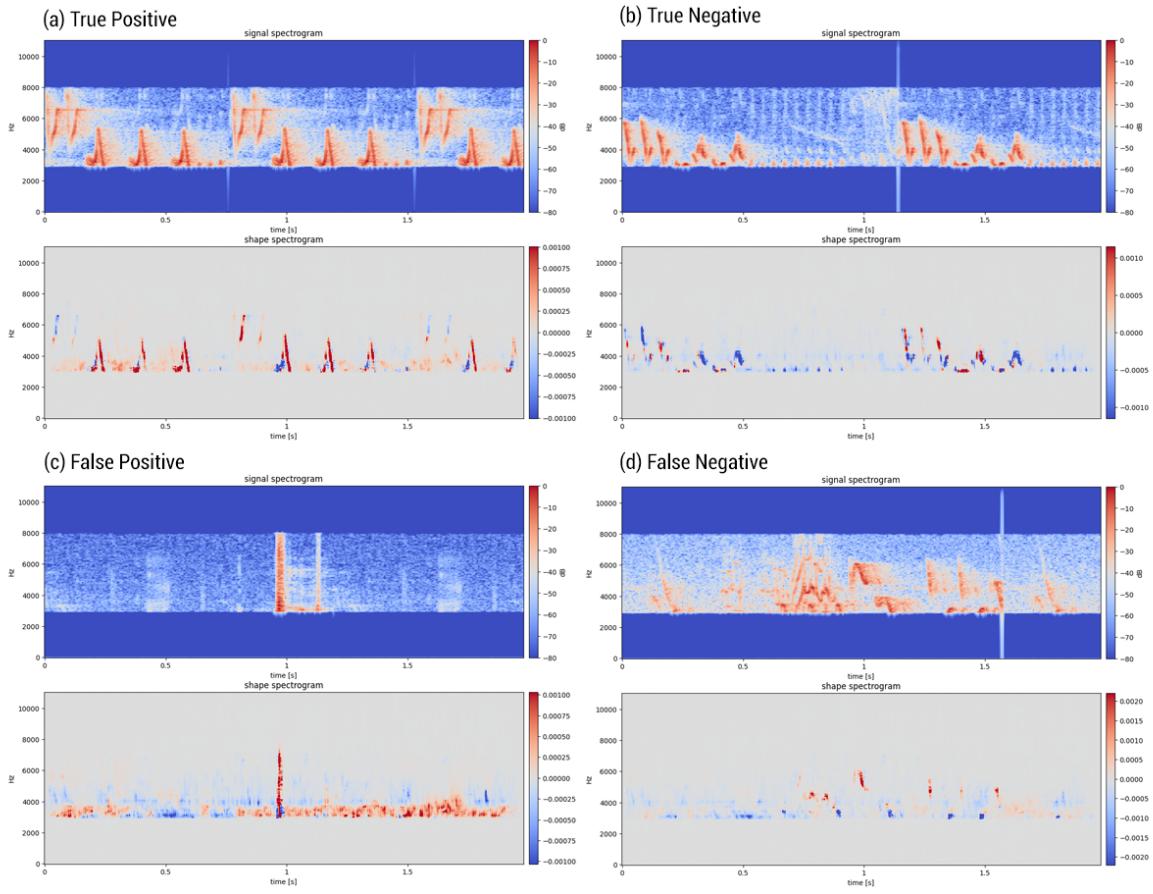


Figura 15 – SHAP spectrograma a) de uma amostra de vocalização de uma região impactada (TP); b) de uma amostra de vocalização de uma região não impactada (TN); c) de uma amostra de vocalização de uma região não-impactada, mas classificada como impactada (FP); d) Uma amostra de vocalização de uma região não impactada, mas classificada como impactada (FN)

Na figura 15, o espectrograma SHAP (a) para o caso de vocalização de corruíra na região impactada (TP) revela um aumento significativo nas contribuições em uma região específica do espectrograma, indicado pelo tom mais escuro de vermelho, mostrando que esta região de trinados corresponde a uma característica relevante para a classe positiva, levando o modelo a classificá-la como "impactado". Além disso, a amplitude do som da corruíra nos trinados é observada pelo modelo, considerando a maior contribuição em áreas onde o espectrograma mostra maior brilho (representando a amplitude do som).

No caso de verdadeiro negativo, o espectrograma SHAP (b) na figura 15, este mostra baixas contribuições em uma região irrelevante para a classe positiva, com um

tom mais escuro de azul aparecendo em alguns pontos de potência onde anteriormente o modelo observou no primeiro caso. Mas, agora, os trinados fizeram contribuições negativas, com áreas marcadas com o azul mais escuro, mostrando que o modelo continuou observando o trinado.

No caso de falso positivo (c) na figura 15, o spectrograma SHAP mostra um aumento significativo nas contribuições de frequência em uma região irrelevante para a classe positiva, e uma forte contribuição em vermelho da amplitude do som em todas as áreas do spectrograma, o que fez o modelo classificar erroneamente como "impactado". O áudio parece ser um som com alta frequência, o que "enganou" o modelo. Com isso, a importância das frequências altas no som da corruíra em um ambiente impactado aparenta ser relevante, corroborando trabalhos recentes de biólogos.

Com contribuições negativas sutis em locais diferentes dos vistos e poucas contribuições positivas em frequências mais próximas do pico do canto da corruíra, na seção (d) da figura 15, o modelo não interpreta as frequências mais baixas como uma característica discriminante, impedindo-o de classificar corretamente, sendo assim considerado um falso negativo.

O spectrograma SHAP pode influenciar a interpretação dos resultados do modelo, identificando áreas no qual o modelo está correto ou incorreto em sua predição com base nas características espectrais relevantes.

### 5.2.2 Visualizações SHAP

Ao visualizar a interpretação SHAP, é notável que as decisões tomadas pelo Modelo de ML são baseadas em píxeis que representam quais frequências oferecem a contribuição mais significativa para a classificação. Para garantir uma visão mais detalhada e clara dessas contribuições, outras duas visualizações foram propostas: o SHAP Spectrum e o SHAP Time.

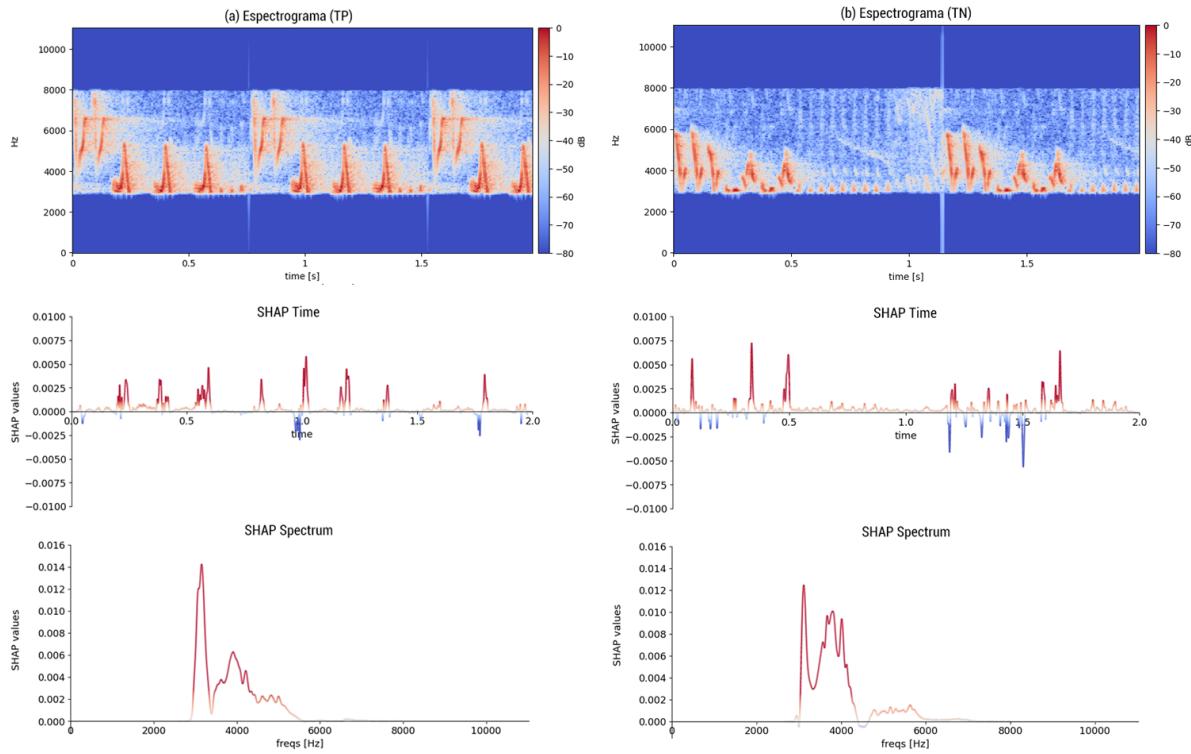


Figura 16 – SHAP Spectrum e SHAP time de (a) uma amostra de vocalização de uma região impactada (TP); (b) uma amostra de vocalização de uma região não impactada (TN)

### 5.2.2.1 SHAP Spectrum

A função `shap_spectrum` calcula os valores SHAP interpolados e como eles são distribuídos entre as frequências. Dados os valores SHAP ( $S$ ) correspondentes a um rótulo específico e uma frequência de amostragem ( $f_s$ ), o SHAP spectrum é calculado somando ( $S$ ) ao longo do eixo das frequências. Em seguida, a interpolação por spline cúbica é aplicada aos índices originais e ao SHAP spectrum, resultando em uma representação contínua.

Como descrito antes, as partes da imagem destacadas em vermelho representam os píxeis que contribuíram positivamente, e quanto mais escura a tonalidade de vermelho, maior é a contribuição para uma previsão correta. Agora, com o espectro SHAP, temos a representação dos valores SHAP no eixo  $y$  e, no eixo  $x$ , a frequência ( $Hz$ ), mantendo a tonalidade escura de vermelho próxima a valores maiores no eixo  $y$ .

Como observado na figura 16, para uma amostra de uma área impactada (TP),

a frequência que mais contribui está entre 2kHz e 6kHz, com um pico em 3,151kHz. Portanto, conforme o SHAP, o modelo percebeu que essas frequências distinguem este indivíduo na região impactada, dando maior importância (ou seja, maior peso) às bandas de frequências que originalmente eram representadas por píxeis.

Na figura 16 parte (b), como uma amostra de uma região não impactada, o modelo também parece prestar atenção a frequências entre 2kHz e 4kHz, conforme indicado pelos valores SHAP, em particular o pico desse valor. Comparado à amostra da região impactada, a frequência que mais contribuiu é um pouco mais baixa (3,115kHz). Aqui, também há alguma contribuição negativa, com azul aparecendo em 3kHz. O modelo mostra que suas escolhas estão alinhadas com a diferença entre as frequências, no qual as corruíras na região não impactada vocalizam em uma frequência mais baixa do que os espécimes na região impactada.

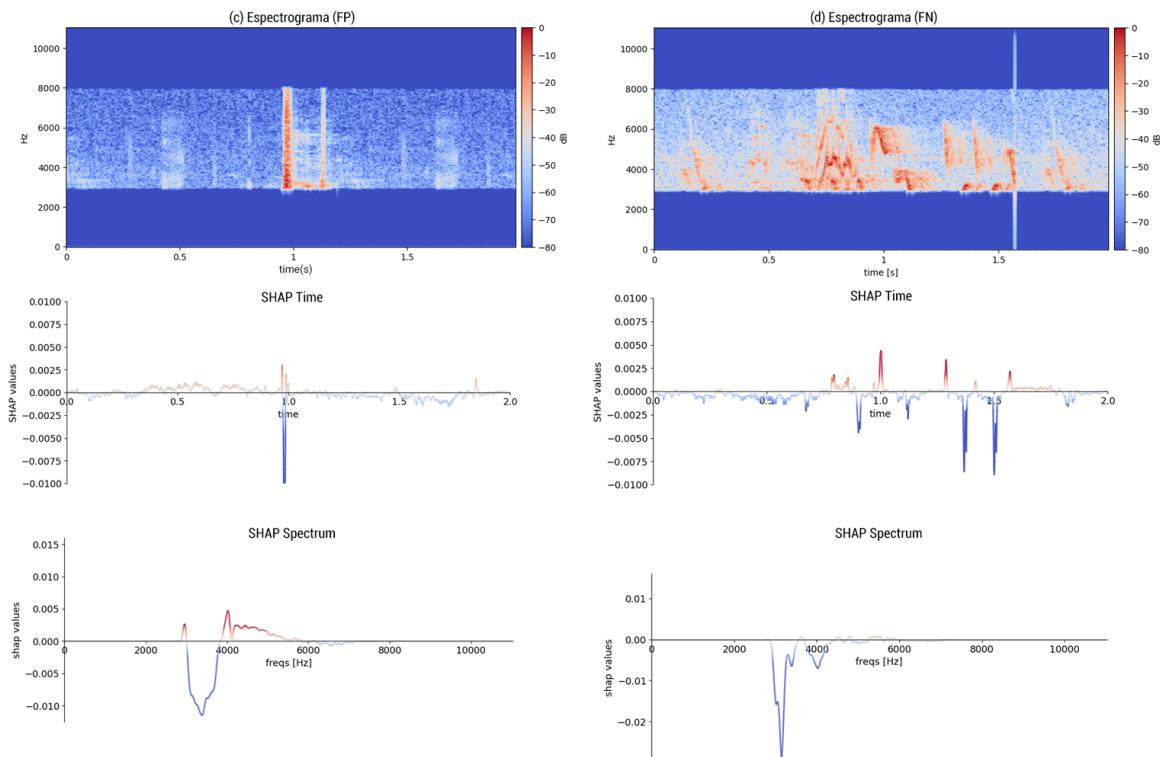


Figura 17 – SHAP Spectrum e SHAP time de (a) uma amostra de vocalização de uma região não-impactada classificada como impactada (FP); (b) uma amostra de vocalização de uma região impactada classificada como não-impactada (FN)

Na figura 17 (c), um exemplo de falso positivo, parece um som de alta frequência, com as partes destacadas em azul-escuro aparecendo em 3,39kHz. Isso facilita entender por que o modelo classificou este exemplo como “impactado”; na verdade, foram as frequências entre 2kHz e 4kHz observadas pelo modelo que mais contribuíram negativamente para essa previsão. Da mesma forma, parte (d) da figura 17, um exemplo de falso negativo, onde a frequência de 3,14kHz mostra o maior impacto negativo, conforme os valores SHAP. Essa frequência apresenta a maior contribuição negativa, sendo muito próxima ao valor de 3,11kHz apresentado no verdadeiro negativo (não impactado), mostrando que o modelo não apenas percebeu a faixa de frequência no qual as corruíras cantam, mas também a diferença sutil na frequência entre um indivíduo de ambiente impactado e indivíduo de ambiente não-impactado.

### 5.2.2.2 SHAP Time

A função *shap\_time* calcula os valores SHAP para o rótulo específico e somados ao longo do eixo das frequências para obter a representação temporal.

Dada uma amostra específica, os valores SHAP ( $S$ ) associados a um rótulo específico são somados ao longo do eixo das frequências (*freqs*) para obter uma representação acumulada dos valores SHAP. Esses valores acumulados são interpolados usando uma spline cúbica para criar uma representação contínua (*shape\_time*).

O gráfico de tempo SHAP foi desenvolvido para fornecer uma perspectiva temporal sobre os valores SHAP. Essa função visualiza como os valores SHAP variam ao longo do tempo, mostrando sobre quais intervalos de tempo contribuem de forma mais significativa para as previsões do modelo.

O gráfico de SHAP time está alinhado com o espectrograma do sinal de áudio, onde as colunas do espectrograma SHAP foram somadas para produzir o gráfico temporal. Por exemplo, conforme mostrado na Figura 16 (a), o gráfico SHAP time para uma amostra impactada destaca contribuições importantes em intervalos de tempo específicos, ilustrando como o modelo interpreta padrões temporais para fazer previsões.

### 5.2.3 SHAP Time-Amplitude

Ainda observando mais a fundo as contribuições temporais dos valores SHAP na classificação dos sinais de áudio, foi proposta a "SHAP Time-Amplitude". Esta visualização combina a amplitude da onda sonora com os valores SHAP interpolados ao longo do tempo, permitindo identificar os intervalos de tempo que mais influenciam as previsões do modelo.

Em ambientes ruidosos, as corruíras ajustaram suas vocalizações para produzir cantos com amplitudes mais altas, o que pode ser refletido no gráfico por mudanças nos valores SHAP em direção a frequências mais altas ou amplitudes aumentadas em regiões associadas à exposição ao ruído antropogênico, como ficou claro pelas escolhas do modelo. Por exemplo, na figura 18 (c), o modelo classificou erroneamente como "impactado" devido ao elemento de frequência mais alta, e o SHAP nos mostra, com um tom escuro de azul, a contribuição negativa desse elemento.

#### 5.2.3.1 Envoltório-Amplitude SHAP

Essa visualização é projetada para analisar a envoltória de amplitude do sinal de áudio em conjunto com os valores SHAP interpolados. Com o uso dessa visualização, padrões distintos de modulação de amplitude nos chamados da Corruíra, caracterizados por flutuações na intensidade do som ao longo do tempo, podem ser visualizados. Os deslocamentos nos gráficos de envoltória de amplitude correspondem a períodos de atividade vocal intensa e relativa quietude, refletindo a mudança na natureza dinâmica da comunicação e as contribuições desses padrões para a classificação. O volume ou intensidade (a amplitude de um som) e o timbre (a qualidade ou caráter de um som que o distingue) são características relevantes que distinguem a Corruíra de outros pássaros. Nessas representações, as mudanças de volume tornam-se mais claras ao observar a modulação de amplitude, e associadas aos valores SHAP, contribuem para as escolhas do modelo.

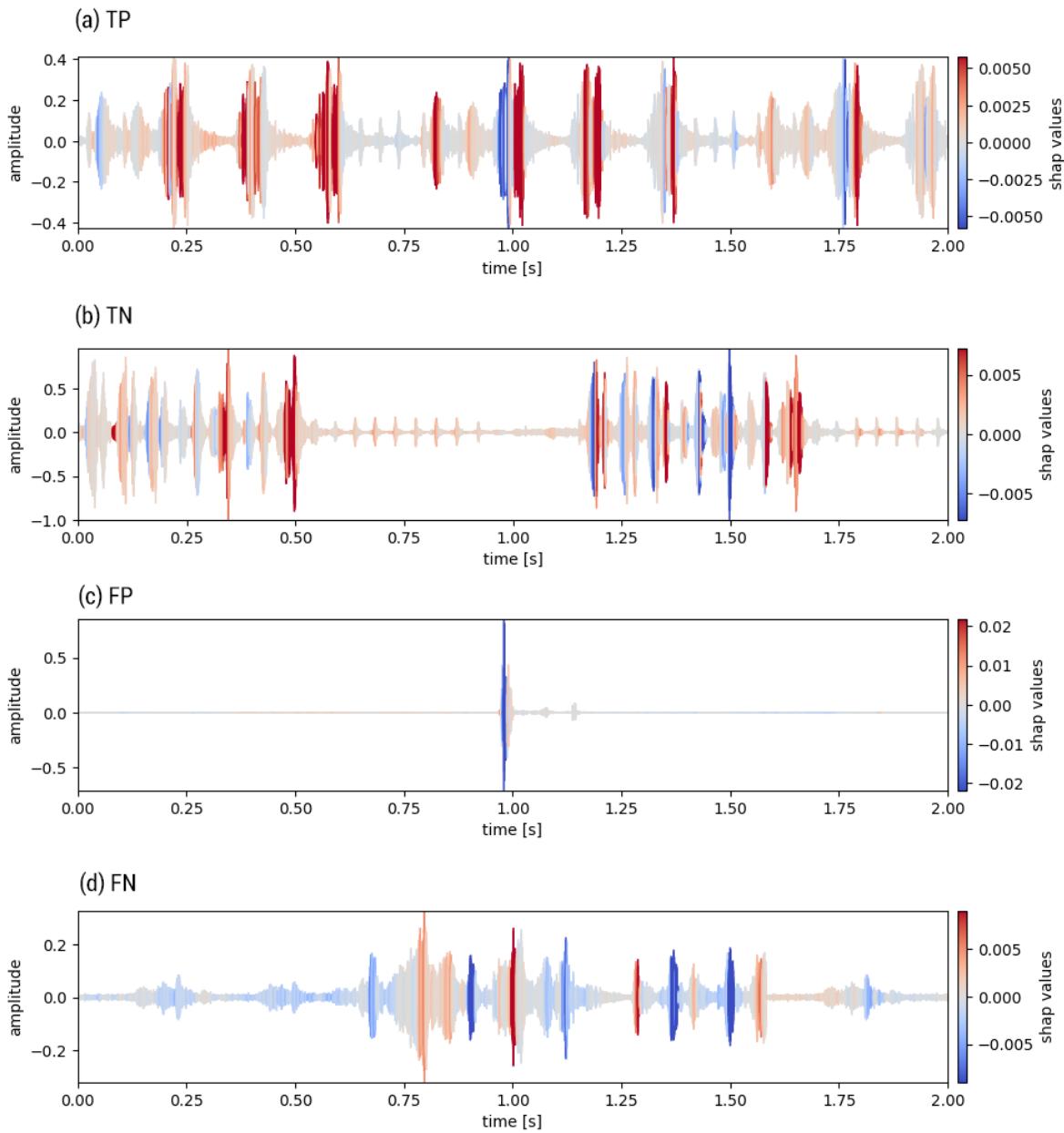


Figura 18 – SHAP Time-Amplitude de a) de uma amostra de vocalização de uma região impactada (TP); b) de uma amostra de vocalização de uma região não impactada (TN); c) de uma amostra de vocalização de uma região não-impactada, mas classificada como impactada (FP); d) Uma amostra de vocalização de uma região não impactada, mas classificada como impactada (FN)

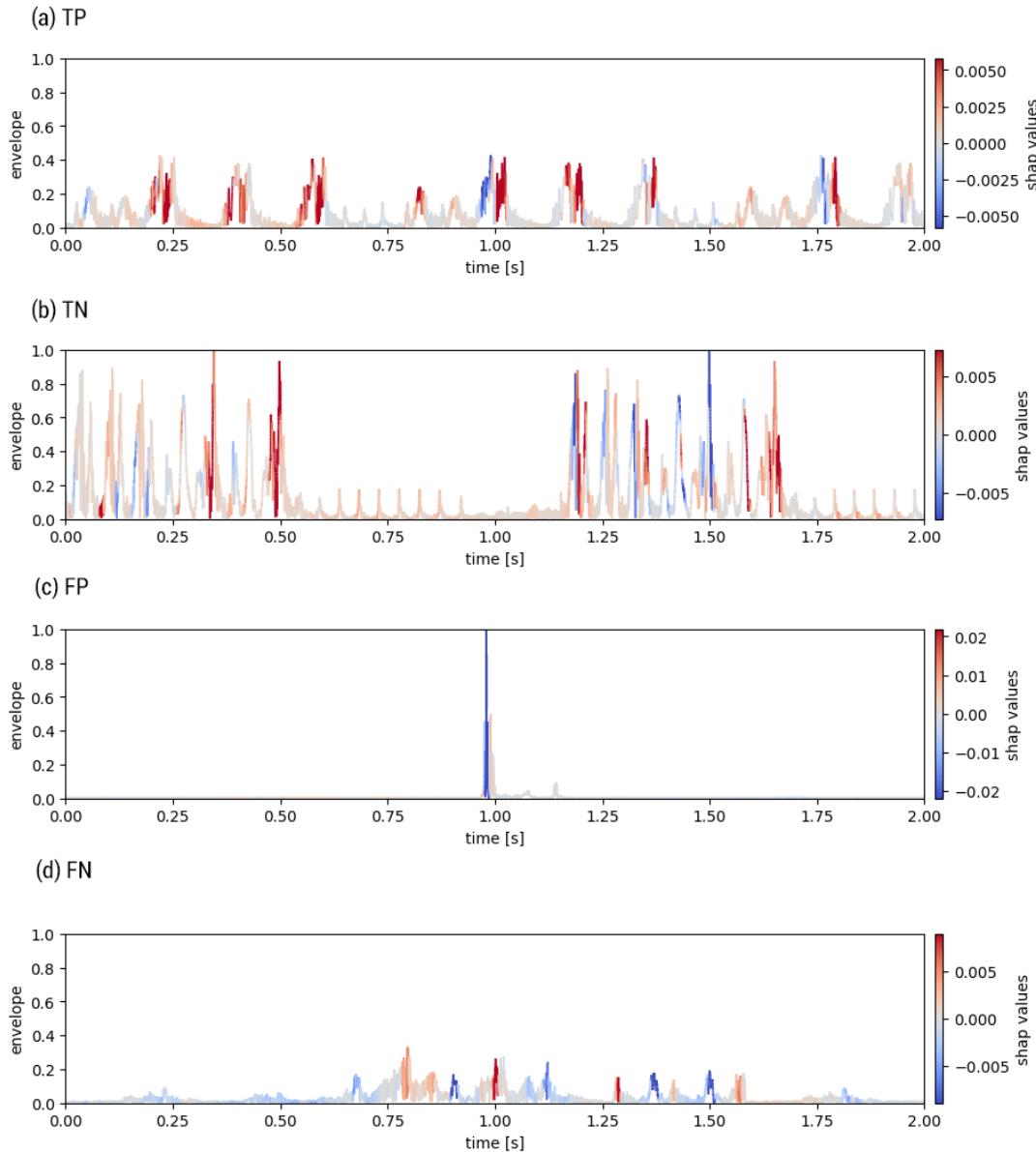


Figura 19 – SHAP Envoltório de: a) de uma amostra de vocalização de uma região impactada (TP); b) de uma amostra de vocalização de uma região não impactada (TN); c) de uma amostra de vocalização de uma região não-impactada, mas classificada como impactada (FP); d) Uma amostra de vocalização de uma região não impactada, mas classificada como impactada (FN)

#### 5.2.4 Contribuições de frequência global (SHAP Espectrograma global)

O espectrograma SHAP global é gerado agregando valores SHAP em todo o conjunto de dados, somando os valores para cada intervalo de frequência e período para capturar sua contribuição coletiva para as previsões do modelo. Matematicamente, o SHAP

espectrograma global  $S_{global}$  é calculado somando os valores SHAP  $S_i$  de todas as amostras  $i$  na matriz de valores SHAP onde  $N$  é o número total de amostras,  $t$  representa o tempo e  $f$  representa a frequência, conforme descrito na Equação 5.1.

$$S_{global}(t, f) = \sum_{i=1}^N S_i(t, f) \quad (5.1)$$

Este spectrograma oferece uma visão abrangente de como os componentes de frequência impactam coletivamente as previsões de alvos em todo o conjunto de dados.

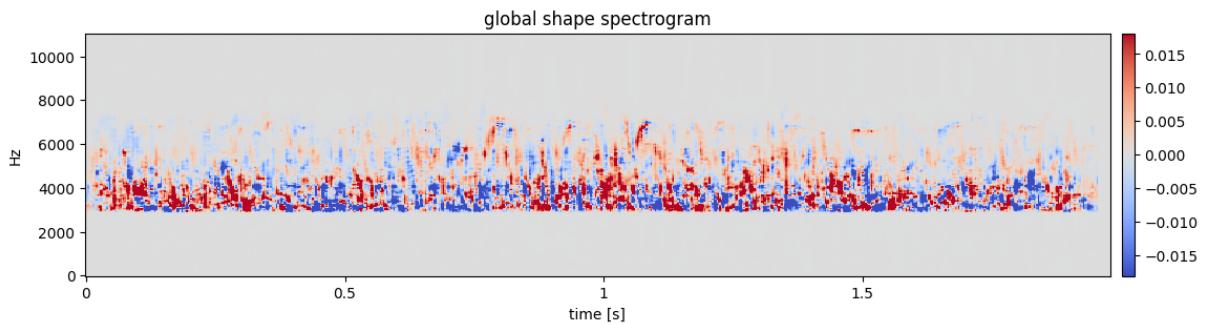


Figura 20 – Espectrograma Global SHAP

#### 5.2.4.1 SHAP Frequênci a global

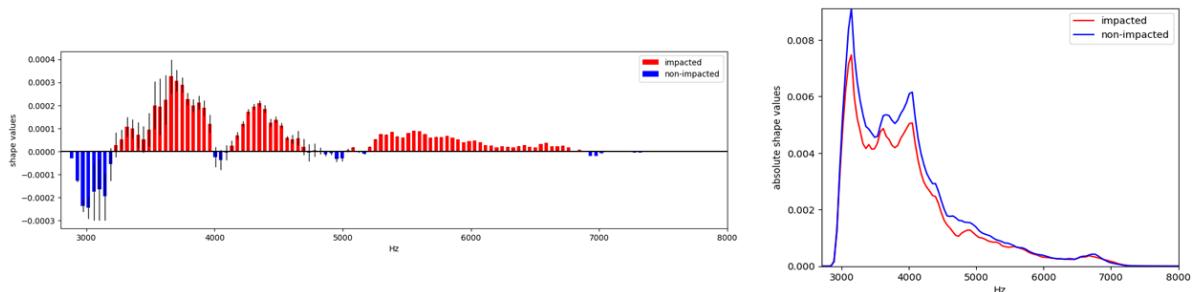


Figura 21 – (a) SHAP Frequênci a Global (b) SHAP frequênci a global em valores absolutos

**SHAP Frequênci a Global:** Dada uma faixa de frequências ( $f$ ) e os valores SHAP ( $S$ ), a função ( $shap\_g\_freq$ ) calcula a influênci a média ( $avg$ ) dos valores SHAP ao longo de todas as amostras.

A função soma os valores SHAP ao longo do eixo temporal, resultando em uma representação acumulada das frequências que tem uma contribuição significativa. Após isso, a média e a variância dos valores SHAP acumulados são calculadas para cada frequência.

Com os valores de média e variância, um gráfico de barras onde as frequências que apresentam contribuições positivas são mostradas em vermelho e as que apresentam contribuição negativa, em azul. Barras de erro são adicionadas para representar a variabilidade dos valores SHAP. Por fim, o gráfico é delimitado pelas frequências mínimas ( $f_{min}$ ) e máximas ( $f_{max}$ ), e uma linha horizontal em zero é desenhada para facilitar a visualização das frequências. Isso proporciona uma análise global das frequências e suas contribuições com os valores SHAP. Aqui, a largura de banda entre 3kHz e 5kHz é o contribuinte mais importante para as previsões negativas ou positivas do modelo em todo o banco de dados.

**SHAP Frequência Global em valores absolutos:** A função calcula e plota a média absoluta dos valores SHAP. Primeiro, os valores SHAP correspondentes os rótulos que representa as amostras impactadas são somados ao longo do eixo temporal, resultando em uma representação acumulada das frequências que contribuíram positiva e negativamente. A média absoluta dessas contribuições e o desvio padrão dos valores SHAP acumulados são então calculados. O mesmo processo é repetido para os valores SHAP das amostras não impactadas, obtendo a média e o desvio padrão. O SHAP, como um método aditivos de atribuição de características, tem valores absolutos de Shapley de todas as instâncias nos dados, então pode ser usado a média para uma visão global.

### 5.3 Considerações Finais

Os experimentos realizados demonstraram a eficácia do modelo E2E ( Biophony e Kapre) na classificação das vocalizações da corruíra em ambientes impactados e não impactados, apresentando um desempenho robusto com alta precisão, revocação e F1-score para ambas as classes. Estes resultados indicam que o modelo tem uma capacidade de generalização sólida, permitindo previsões precisas e consistentes sobre dados

fornecidos.

A interpretação dos valores SHAP proporcionou uma visão detalhada das contribuições dos diferentes píxeis no espectrograma para a decisão do modelo. Com **espectrograma SHAP** mostrou as frequências mais impactantes para a classificação, com destaque para a faixa entre 2kHz e 6kHz, que desempenhou um papel crucial na distinção entre duas amostras de corruíras de ambientes impactados e não-impactados.

Além disso, os casos de falsos positivos e negativos foram observados, mostrando como o modelo pode ser enganado por características específicas aprendidas com as informações que o modelo aprendeu, como a associação de frequências mais altas as mudanças na vocalização do corruíra como resposta ao ruído antropogênico.

As análises adicionais com **SHAP Spectrum** e **SHAP Time** destacaram a variação das contribuições das frequências e dos valores SHAP ao longo do tempo. Essas visualizações ajudaram uma compreensão mais profunda sobre quais frequências e intervalos de tempo são mais significativos (seja em contribuição positiva ou negativa) para as previsões do modelo.

Finalmente, a análise global das frequências com **SHAP Frequência Global** explicitou quais bandas de frequência são mais relevantes para as previsões tanto positivas como negativas, com uma faixa específica de 3kHz a 5kHz mostrando contribuições significativas. Esta visão global, complementada pela análise de média e variância, forneceu uma compreensão abrangente das frequências que mais influenciam a classificação do modelo.

As visualizações **SHAP Time-Amplitude** e **Envoltório-Amplitude SHAP** trouxe novas perspectivas sobre a influência da amplitude e da modulação do som na classificação. Essas abordagens mostram indícios de relação entre a intensidade do som e as contribuições dos valores SHAP, ilustrando como ajustes na amplitude podem impactar a decisão do modelo.

Esses resultados fornecem uma base sólida para entender a importância das características acústicas na classificação das vocalizações da corruíra e conduz para novas pesquisas e aplicações em bioacústica.

# 6

---

## CONCLUSÕES

A inteligência artificial (IA), especialmente no campo do aprendizado de máquina (ML), muitas vezes é percebida como uma caixa preta complexa, limitando sua aplicação eficaz na gestão ambiental sustentável. No entanto, com o advento da explicabilidade em IA (XAI), os modelos podem ser tornar mais transparentes e compreensíveis. A XAI visa estabelecer confiança entre desenvolvedores e usuários, melhorando a interpretabilidade e a aceitação dos resultados.

Neste estudo, utilizamos uma abordagem E2E, associando Biophony e Kapre para obter bons resultados na classificação e técnicas avançadas de XAI, como o SHAP (SHapley Additive exPlanations), para aprimorar a interpretabilidade e o desempenho dessa abordagem, que foi desenvolvido para identificar variações nas vocalizações da corruíra (*Troglodytes aedon musculus*) em resposta ao ruído antropogênico. Além de aumentar a precisão das previsões, essas técnicas proporcionaram explicações claras e visualizações compreensíveis para ornitólogos e ecologistas, facilitando uma melhor compreensão dos padrões de vocalização das aves em ambientes urbanos.

Nossas descobertas introduzem novas abordagens para visualizar as contribuições do modelo, tanto em escala local quanto global, utilizando o espectro SHAP, o espectrograma SHAP e o SHAP aplicado aos padrões temporais dos sinais. Essas visualizações revelaram como mudanças no tempo e na frequência são críticas para as decisões do modelo, proporcionando noções sobre a adaptação das aves ao ruído urbano e suas consequências para a conservação da biodiversidade.

## 6.1 Extensões futuras

O trabalho realizado fornece uma base sólida para a classificação das vocalizações da corruíra em ambientes impactados e não impactados, mas há várias áreas que podem ser exploradas para aprimorar e expandir a análise.

- Aplicar a metodologia desenvolvida a outras espécies de animais e estudar vocalizações de diferentes espécies, especialmente aquelas em ambientes impactados, permitirá validar e ajustar a abordagem proposta para cada contexto ecológico específico, ampliando a aplicabilidade do trabalho para outras áreas da biologia e conservação;
- Melhorar nas técnicas de interpretação dos resultados pode ser explorada com o uso de outros explicadores ou mesmo desenvolver novos explicadores que permitam o avanço da explicabilidade na bioacústica;
- Desenvolver ferramentas interativas baseadas nas visualizações SHAP que permitam explorar a análise e interpretação dos resultados por biólogos e pesquisadores. Interfaces de usuário intuitivas permitiriam uma exploração mais dinâmica dos dados, promovendo uma compreensão mais detalhada das contribuições das características acústicas e melhorando a usabilidade do modelo.

---

## REFERÊNCIAS

- AGARWAL, N. et al. Transfer learning: Survey and classification. *Smart innovations in communication and computational sciences*, Springer, p. 145–155, 2021. [35](#), [36](#)
- AGGARWAL, D.; HASIJA, Y. A Review of Deep Learning Techniques for Protein Function Prediction. 2022. [47](#)
- AKBAL, E. et al. Explainable automated anuran sound classification using improved one-dimensional local binary pattern and tunable q wavelet transform techniques. *Expert Systems with Applications*, v. 225, p. 120089, 2023. ISSN 0957-4174. [50](#)
- ALI, S. et al. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, v. 99, p. 101805, 2023. ISSN 1566-2535. [40](#)
- BAHAR, P.; BIESCHKE, T.; NEY, H. A comparative study on end-to-end speech to text translation. 11 2019. [38](#)
- BAHDANAU, D. et al. End-to-end attention-based large vocabulary speech recognition. In: IEEE. 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). [S.I.], 2016. p. 4945–4949. [37](#)
- BAPTISTA, M. L.; GOEBEL, K.; HENRIQUES, E. M. Relation between prognostics predictor evaluation metrics and local interpretability shap values. *Artif. Intell.*, Elsevier Science Publishers Ltd., GBR, v. 306, n. C, may 2022. ISSN 0004-3702. [40](#), [42](#)
- Barredo Arrieta, A. et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, v. 58, p. 82–115, 2020. ISSN 1566-2535. [40](#)
- BERMANT, P. C. Biocppnet: automatic bioacoustic source separation with deep neural networks. *Scientific Reports*, Nature Publishing Group, v. 11, n. 1, p. 1–13, 2021. [37](#)
- BERMANT, P. C. et al. Deep machine learning techniques for the detection and classification of sperm whale bioacoustics. *Scientific reports*, Nature Publishing Group, v. 9, n. 1, p. 1–10, 2019. [48](#)
- BROWN, C.; RIEDE, T. Comparative bioacoustics: An overview. Bentham Science Publishers, 2017. [18](#), [29](#)

- CADIEUX, P. et al. Projected effects of climate change on boreal bird community accentuated by anthropogenic disturbances in western boreal forest, canada. *Diversity and Distributions*, 2020. [51](#)
- CANZIANI, A.; PASZKE, A.; CULURCIELLO, E. *An Analysis of Deep Neural Network Models for Practical Applications*. [S.l.]: arXiv, 2016. [47](#)
- CANÁRIO, J. P.; RIBEIRO, O.; RIOS, R. Explaining noise effects in cnn: a practical case study on volcano signals. p. 49–54, 2022. [40](#), [42](#), [49](#)
- CHACHADA, S.; KUO, C.-C. J. Environmental sound recognition: A survey. *APSIPA Transactions on Signal and Information Processing*, Cambridge University Press, v. 3, 2014. [28](#)
- CHACÓN, J. U. *Assessment of animal acoustic diversity in neotropical forest*. Tese (Theses) — Université Paris Saclay (COmUE), 2018. [49](#), [50](#)
- CHOI, K.; JOO, D.; KIM, J. Kapre: On-gpu audio preprocessing layers for a quick implementation of deep neural network models with keras. 06 2017. [38](#)
- CHOI, K.; JOO, D.; KIM, J. *Kapre: On-GPU Audio Preprocessing Layers for a Quick Implementation of Deep Neural Network Models with Keras*. 2017. [51](#)
- COOLEY, J. W.; TUKEY, J. W. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, JSTOR, v. 19, n. 90, p. 297–301, 1965. [29](#)
- CORP., N. I. *Understanding FFTs and Windowing*. 2021. Url<https://www.ni.com/pt-br/innovations/white-papers/06/understanding-ddft-and-windowing.html>. [28](#)
- CULLINAN, V. I.; MATZNER, S.; DUBERSTEIN, C. A. Classification of birds and bats using flight tracks. *Ecological Informatics*, v. 27, p. 55–63, 2015. ISSN 1574-9541. [16](#), [26](#)
- CYR KIMBERLEY WETTEN, M. H. W. M.-E.; KOPER, N. Variation in song structure of house wrens living in urban and rural areas in a caribbean small island developing state. *Bioacoustics*, 2021. [18](#), [52](#)
- DAS, N. et al. Exploring explainable ai methods for bird sound-based species recognition systems. *Multimedia Tools and Applications*, Springer Science+Business Media, 01 2024. [49](#), [50](#)
- DAS, N. et al. Exploring explainable AI methods for bird sound-based species recognition systems. *Multimedia Tools and Applications*, v. 83, n. 24, p. 64223–64253, jul. 2024. ISSN 1573-7721. [50](#)
- DINIZ, P.; DUCA, C. Anthropogenic noise, song, and territorial aggression in southern house wrens. *Journal of Avian Biology*, 2021. [18](#)
- DU, H. et al. The elements of end-to-end deep face recognition: A survey of recent advances. *ACM Computing Surveys (CSUR)*, ACM New York, NY, v. 54, n. 10s, p. 1–42, 2022. [37](#)
- DUMOULIN, V.; VISIN, F. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285*, 2016. [32](#), [33](#), [34](#)

- E. Cramer. Physically Challenging Song Traits, Male Quality, and Reproductive Success in House Wrens. *PLoS ONE*, v. 8, mar. 2013. [24](#)
- E. Cramer. Vocal deviation and trill consistency do not affect male response to playback in house wrens. *Behavioral Ecology*, v. 24, p. 412–420, mar. 2013. [24](#)
- EBERLY ABRAM FLEISHMAN, D. S. C. *Tutorial: Accurate Bioacoustic Species Detection from Small Numbers of Training Clips Using the Biophony Model*. 2020. <Https://github.com/microsoft/acoustic-bird-detection>. [51](#), [52](#)
- EENS, M. et al. Great and blue tits as indicators of heavy metal contamination in terrestrial ecosystems. *Ecotoxicology and environmental safety*, v. 44, p. 81–5, 10 1999. [16](#), [26](#)
- ERBE, C. What is animal bioacoustics. *Journal of the Acoustical Society of America*, v. 139, p. 2004–2004, 2016. [26](#)
- FERNANDES, M.; CORDEIRO, W.; RECAMONDE-MENDOZA, M. Detecting aedes aegypti mosquitoes through audio classification with convolutional neural networks. 08 2020. [16](#), [26](#)
- FLEISHMAN, A. et al. *Tutorial: Accurate Bioacoustic Species Detection from Small Numbers of Training Clips Using the Biophony Model*. 2020. [39](#)
- FONSECA, V. et al. Classification of tropical disease-carrying mosquitoes using deep learning and shap. In: *Anais do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*. Porto Alegre, RS, Brasil: SBC, 2023. p. 25–34. ISSN 2763-8952. [50](#)
- GOMES, C. et al. *A Survey of Explainable AI and Proposal for a Discipline of Explanation Engineering*. 2023. [49](#)
- GRABARCZYK, E.; GILL, S. Anthropogenic noise affects male house wren response to but not detection of territorial intruders. *PLoS ONE*, 2019. [51](#)
- GU, J. et al. Recent advances in convolutional neural networks. *Pattern Recognition*, v. 77, p. 354–377, 2018. ISSN 0031-3203. [30](#)
- GU, J. et al. Recent advances in convolutional neural networks. *Pattern recognition*, Elsevier, v. 77, p. 354–377, 2018. [32](#), [33](#), [34](#)
- HEATON, J. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618. *Genetic Programming and Evolvable Machines*, Springer, v. 19, n. 1-2, p. 305–307, 2018. [32](#)
- JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015. [32](#), [33](#), [34](#)
- JUAREZ, R. et al. House wrens troglodytes aedon reduce repertoire size and change song element frequencies in response to anthropogenic noise. *Ibis*, 2021. [52](#)
- KAHL, S. et al. Birdnet: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, Elsevier, v. 61, p. 101236, 2021. [48](#)

- KAUR, H. et al. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. [S.l.: s.n.], 2020. [41](#)
- KNIGHT, E. et al. Recommendations for acoustic recognizer performance assessment with application to five common automated signal recognition programs. *Avian Conservation and Ecology*, The Resilience Alliance, v. 12, n. 2, 2017. [47](#)
- KVSN, R. R. et al. Bioacoustics data analysis—a taxonomy, survey and open challenges. *IEEE Access*, IEEE, v. 8, p. 57684–57708, 2020. [26](#), [30](#), [33](#), [34](#), [47](#)
- L. Kermott; L. Scott Johnson. The Functions of Song in Male House Wrens (*Troglodytes Aedon*). *Behaviour*, v. 116, p. 190–209, 1991. [25](#)
- LASSECK, M. Audio-based bird species identification with deep convolutional neural networks. *CLEF (working notes)*, v. 2125, 2018. [48](#)
- LECUN, Y. et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 86, n. 11, p. 2278–2324, 1998. [8](#), [31](#), [33](#)
- LEE, C.-Y.; GALLAGHER, P. W.; TU, Z. Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree. In: PMLR. *Artificial intelligence and statistics*. [S.I.], 2016. p. 464–472. [33](#)
- LI, J. et al. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, Now Publishers, Inc., v. 11, n. 1, 2022. [37](#)
- LI, L. et al. Automated classification of *tursiops aduncus* whistles based on a depth-wise separable convolutional neural network and data augmentation. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 150, n. 5, p. 3861–3873, 2021. [48](#)
- LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. [S.I.]: Curran Associates Inc., 2017. p. 4768–4777. ISBN 9781510860964. [41](#), [42](#), [60](#)
- MADHUSUDHANA, S. et al. Choosing equipment for animal bioacoustic research. In: \_\_\_\_\_. *Exploring Animal Behavior Through Sound: Volume 1: Methods*. Cham: Springer International Publishing, 2022. p. 37–85. ISBN 978-3-030-97540-1. [28](#)
- Mala H. Sawhney; M. C. Baker; Bradley R. Bisbee. DEVELOPMENT OF VOCALISATIONS IN NESTLING AND FLEDGLING HOUSE WRENS IN NATURAL POPULATIONS. *Bioacoustics*, v. 15, p. 271–287, jan. 2006. [25](#)
- MALFANTE, M. et al. Automatic fish sounds classification. *Journal of the Acoustical Society of America*, v. 139, p. 2115–2116, 04 2016. [16](#), [26](#)
- MARCHAL, J.; FABIANEK, F.; AUBRY, Y. Software performance for the automated identification of bird vocalisations: the case of two closely related species. *Bioacoustics*, Taylor & Francis, v. 31, n. 4, p. 397–413, 2022. [47](#)

- MCLOUGHLIN, M.; STEWART, R.; MCELLIGOTT, A. Automated bioacoustics: Methods in ecology and conservation and their potential for animal welfare monitoring. *Journal of The Royal Society Interface*, v. 16, 05 2019. [47](#)
- MOLNAR, C. *Interpretable Machine Learning*: A guide for making black box models explainable. 2. ed. [S.l.: s.n.], 2022. [60](#)
- MOLNÁR, C. et al. Classification of dog barks: a machine learning approach. *Animal Cognition*, Springer, v. 11, n. 3, p. 389–400, 2008. [48](#)
- MORFI, V.; LACHLAN, R. F.; STOWELL, D. Deep perceptual embeddings for unlabelled animal sound events. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 150, n. 1, p. 2–11, 2021. [49](#)
- MOSCA, E. et al. SHAP-based explanation methods: A review for NLP interpretability. In: *Proceedings of the 29th International Conference on Computational Linguistics*. [S.l.]: International Committee on Computational Linguistics, 2022. p. 4593–4603. [41](#)
- MUTANU, L. et al. A review of automated bioacoustics and general acoustics classification research. *Sensors*, v. 22, n. 21, 2022. ISSN 1424-8220. [26](#), [47](#), [48](#)
- MUTANU, L. et al. A review of automated bioacoustics and general acoustics classification research. *Sensors*, 2022. [47](#), [51](#), [52](#)
- NARANJO-TORRES, J. et al. A review of convolutional neural network applied to fruit image processing. *Applied Sciences*, v. 10, n. 10, 2020. ISSN 2076-3417. [31](#), [34](#)
- ORTIGOSSA, E. S.; GONÇALVES, T.; NONATO, L. G. Explainable artificial intelligence (xai)—from theory to methods and applications. *IEEE Access*, v. 12, p. 80799–80846, 2024. [40](#), [49](#)
- P. Tubaro. Song description of the House Wren (*Troglodytes aedon*) in two populations of eastern Argentina, and some indirect evidences of imitative vocal learning. *El Hornero*, set. 1990. [25](#)
- PADOVESE, B. et al. Data augmentation for the classification of north atlantic right whales upcalls a. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 149, n. 4, p. 2520–2530, 2021. [48](#)
- PAN, S. J.; YANG, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, IEEE, v. 22, n. 10, p. 1345–1359, 2010. [35](#), [36](#)
- PERIS, S.; COLINO-RABANAL, V. Does the song of the wren *troglodytes troglodytes* change with different environmental sounds? *Acta Ornithologica*, 2016. [52](#)
- PRABHAVALKAR, R. et al. A comparison of sequence-to-sequence models for speech recognition. In: *Interspeech*. [S.l.: s.n.], 2017. p. 939–943. [37](#)
- PRINCE, P. et al. Deploying acoustic detection algorithms on low-cost, open-source acoustic sensors for environmental monitoring. *Sensors*, MDPI, v. 19, n. 3, p. 553, 2019. [47](#)
- PROPPE, D. S.; STURDY, C. B.; CLAIR, C. C. S. Anthropogenic noise decreases urban songbird diversity and may contribute to homogenization. *Global Change Biology*, 2013. [51](#)

- PROVOST, K. L.; YANG, J.; CARSTENS, B. C. The impacts of fine-tuning, phylogenetic distance, and sample size on big-data bioacoustics. *Plos one*, Public Library of Science San Francisco, CA USA, v. 17, n. 12, 2022. [48](#)
- ROCH, M. A. et al. Organizing metadata from passive acoustic localizations of marine animals. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 141, n. 5, p. 3605–3605, 2017. [18](#)
- SANCHEZ, F. J. B. et al. Bioacoustic classification of avian calls from raw sound waveforms with an open-source deep learning architecture. *Scientific Reports*, Nature Publishing Group, v. 11, n. 1, p. 1–12, 2021. [48](#)
- SANDOVAL, L.; REDONDO, P.; BARRANTES, G. Urban noise influences vocalization structure in the house wren troglodytes aedon. *Ibis*, 2013. [8](#), [24](#), [51](#), [52](#)
- SCHERER, D.; MÜLLER, A.; BEHNKE, S. Evaluation of pooling operations in convolutional architectures for object recognition. In: SPRINGER. *Artificial Neural Networks–ICANN 2010: 20th International Conference, Thessaloniki, Greece, September 15–18, 2010, Proceedings, Part III* 20. [S.l.], 2010. p. 92–101. [33](#)
- SEMENTILI-CARDOSO, G.; DONATELLI, R. Anthropogenic noise and atmospheric absorption of sound induce amplitude shifts in the songs of southern house wren (troglodytes aedon musculus). *Urban Ecosystems*, v. 24, p. 1–9, 10 2021. [16](#), [17](#), [18](#), [51](#), [52](#), [54](#)
- SKUTCH, A. Life history of the southern house wren. *The Condor*, v. 55, p. 121–149, 1953. [24](#)
- SONG, Z. et al. An improved nyquist–shannon irregular sampling theorem from local averages. *IEEE transactions on information theory*, IEEE, v. 58, n. 9, p. 6093–6100, 2012. [27](#)
- SOUZA, L. S.; GATTO, B. B.; FUKUI, K. Classification of bioacoustic signals with tangent singular spectrum analysis. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2019. p. 351–355. [16](#), [26](#)
- STOWELL, D. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, PeerJ Inc., v. 10, p. e13152, 2022. [18](#), [26](#), [27](#), [30](#), [33](#), [34](#), [39](#), [47](#), [48](#), [49](#), [51](#)
- TAN, C. et al. A survey on deep transfer learning. In: SPRINGER. *International conference on artificial neural networks*. [S.l.], 2018. p. 270–279. [11](#), [36](#), [37](#)
- THAKUR, A. et al. Deep metric learning for bioacoustic classification: Overcoming training data scarcity using dynamic triplet loss. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 146, n. 1, p. 534–547, 2019. [47](#), [48](#)
- THOMAS, R. W. et al. Cognitive networks: adaptation and learning to achieve end-to-end performance objectives. *IEEE Communications Magazine*, v. 44, n. 12, p. 51–57, 2006. [37](#)
- ULLOA, J. S. et al. Towards an end-to-end framework for sound classification in passive acoustic recordings. *Assessment of animal acoustic diversity in neotropical forest*, p. 123, 2018. [37](#)

- WEISS, K.; KHOSHGOFTAAR, T. M.; WANG, D. A survey of transfer learning. *Journal of Big data*, SpringerOpen, v. 3, n. 1, p. 1–40, 2016. [34](#), [35](#)
- XU, H. et al. End-to-end learning of driving models from large-scale video datasets. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2017. p. 2174–2182. [37](#)
- YAMASHITA, R. et al. Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, Springer, v. 9, p. 611–629, 2018. [32](#)
- ZHONG, M. et al. Multispecies bioacoustic classification using transfer learning of deep convolutional neural networks with pseudo-labeling. *Applied Acoustics*, Elsevier, v. 166, p. 107375, 2020. [48](#)
- ZHUANG, F. et al. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, IEEE, v. 109, n. 1, p. 43–76, 2020. [34](#), [35](#), [36](#)