UFMT

UNIVERSIDADE FEDERAL DE MATO GROSSO

INSTITUTO DE COMPUTAÇÃO

COORDENAÇÃO DE ENSINO DE GRADUAÇÃO EM

CIÊNCIA DA COMPUTAÇÃO

# REDES NEURAIS PROFUNDAS INTERPRETÁVEIS PARA AUXÍLIO NO MONITORAMENTO DE AVES

**GABRIEL DE SOUSA GOMES PEDROSO**

**CUIABÁ - MT**

2023

UFMT

UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE COMPUTAÇÃO
COORDENAÇÃO DE ENSINO DE GRADUAÇÃO EM
CIÊNCIA DA COMPUTAÇÃO

# REDES NEURAIS PROFUNDAS INTERPRETÁVEIS PARA AUXÍLIO NO MONITORAMENTO DE AVES

**GABRIEL DE SOUSA GOMES PEDROSO**

**Orientador:** Prof. Dr. Thiago Meirelles Ventura

Monografia apresentada ao Curso de Ciência da Computação, do Instituto de Computação da Universidade Federal de Mato Grosso, como requisito para obtenção do título de Bacharel em Ciência da Computação

**CUIABÁ - MT**

2023

*Este trabalho é dedicado a minha mãe Silene de Sousa Pedroso e meu pai Florisvaldo Gomes Pedroso por estarem sempre presentes em momentos importantes da minha vida e incentivarem meus estudos.*

# AGRADECIMENTOS

A Deus por me ajudar a enxergar o futuro, não desistir de meus objetivos e não perder a fé das coisas boas da vida. À Nossa Senhora Aparecida por interceder a Deus pela minha saúde e vida, para continuar tendo energias e forças para continuar as lutas diárias.

A minha mãe Silene de Sousa Pedroso e meu pai Florisvaldo Gomes Pedroso, por estarem sempre presentes comigo, nos momentos difíceis e nos momentos bons, por me incentivarem nos estudos desde sempre e serem minha base para continuar lutando diariamente pelos meus objetivos.

A minha amiga Maria Gorete, por ter me ajudado numa fase complicada da minha vida durante a pandemia da COVID-19 em 2020 e 2021.

Ao ex-aluno Victor Arinos, por ter me observado e indicado para participar do grupo de pesquisa CO.BRA [1]. Ao ex-aluno Lucas Silva, por ter auxiliado no início e desenvolvimento de minha Iniciação Científica. Aos professores do Instituto de Computação pelos ensinamentos, em especial, ao Allan Oliveira e Thiago Ventura, por terem me orientado durante a graduação na Iniciação Científica e com outras dúvidas que surgiram também. Tornaram especial essa jornada de 4 anos e foram e são muito importantes para mim.

Aos professores Nielsen Simões e Eduardo Bogue, por serem os principais motivadores dos avanços com a Maratona de Programação [2] e acreditar nos alunos para promover melhorias para nós e todo o Instituto de Computação.

[1] https://cobra.ic.ufmt.br/

[2] http://maratona.sbc.org.br/

A todos meus amigos e amigas, e parentes que estiveram comigo durante esses anos e entenderam minha ausência durante a faculdade.

# RESUMO

O monitoramento de aves é de fundamental importância para a sociedade, pois além da proteção às espécies, é possível entender o ambiente. Uma forma de realização deste monitoramento ocorre a partir do estudo da bioacústica, o que envolve o reconhecimento de espécies a partir de suas vocalizações, sejam em um período do dia ou em uma estação. Inúmeros métodos de Aprendizagem de Máquina têm sido utilizados para avanços nos estudos bioacústicos e diversas publicações exemplificam a aplicabilidade e os impactos desses métodos, principalmente com o avanço de Redes Neurais Profundas (RNP). No entanto, tais modelos caracterizam-se por serem caixa-preta, e apesar de já existirem estudos para a interpretação desses modelos, a quantidade de publicações da interpretabilidade das RNP é inferior na bioacústica em comparação com outras áreas. Assim, esse trabalho visa ao descobrimento dos impactos de intepretabilidade em RNP para avaliar de dados de bioacústica para o monitoramento de aves, principalmente de espécies do Pantanal, Mato Grosso, Brasil. Os resultados demonstraram melhoria de desempenho do modelo através do entendimento da aprendizagem do modelo, com aumento de precisão, *recall* e $F1 - score$ de até 0,5%, 3,5% e 2,33%, respectivamente. Essas métricas foram obtidas através da aplicação de métodos de interpretabilidade após o treinamento de um modelo. Assim, a interpretabilidade em Redes Neurais Profundas podem auxiliar no entendimento do modelo e consequentemente na melhoria de seu desempenho. Deste modo, métodos de interpretabilidade em Redes Neurais Profundas devem ser mais explorados na bioacústica.

**Palavras-chaves**: Bioacústica computacional. Detecção de atividade acústica. Reconhecimento sonoro de aves. Interpretabilidade em Redes Neurais.

# ABSTRACT

Bird monitoring is of fundamental importance to the society because, in addition to the protection the species, it is possible to understand the environment. One approach to conducting such monitoring is through the study of bioacoustics, which involves species recognition from their vocalizations, whether in a particular time of day or season. Numerous Machine Learning methods have been used to advance bioacoustic studies, and several publications exemplify the aplicability and the impacts of these methods, mainly with the advancing of Deep Neural Networks (DNNs). However, such models are caracterized as black boxes, and although there are studies on the interpretation of these models, the number of publications on the interpretability of DNNs in bioacoustics is lesser than in other fields. Therefore, this work aims to discover the impacts of interpretability in DNN models for evaluate bioacoustics data for bird monitoring, mainly focusing on Pantanal's species in Mato Grosso, Brazil. The results demonstrated performance improvement of the model through model understanding, with increases in precision, recall, and F1 score of up to 0.5%, 3.5%, and 2.33%, respectively. Thus, interpretability in Deep Neural Networks can aid in model understanding and consequently enhance its performance. Hence, interpretability methods in Deep Neural Networks should be further explored in bioacoustics.

**Key-words**: Computational bioacoustics. Acoustic activity detection. Bird sound recognition. Interpretability in Deep Neural Networks.

# SUMÁRIO

# LISTA DE ILUSTRAÇÕES

# LISTA DE TABELAS

# LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| BG | Background |
| CA | Canto |
| CNN | *Convolutional Neural Network* |
| CO.BRA | *Computational Bioacoustics Research Unit* |
| DNN | *Deep Neural Network* |
| FN | *False Negative* |
| FP | *False Positive* |
| IC | Instituto de Computação |
| LIME | *Local Interpretable Model-agnostic Explanations* |
| PANNs | *Pretrained audio neural networks* |
| ReLU | Rectified Linear Unit |
| RNP | Redes Neurais Profundas |
| S. | Synallaxis |
| SYL1 | Syllable 1 |
| SYL2 | Syllable 2 |

| | |
|---|---|
| TCC | Trabalho de Conclusão de Curso |
| TN | *True Negative* |
| TP | *True Positive* |
| UFMT | Universidade Federal de Mato Grosso |
| TCAV | *Testing with Concept Activation Vectors* |

# CAPÍTULO 1

# INTRODUÇÃO

Este capítulo trata do funcionamento da bioacústica e sua relação com aves, sua importância e esforços atuais. Além disso, é apresentada a contextualização de Redes Neurais Profundas interpretáveis, suas aplicações, importância e impactos. Salientado esses pontos, é demonstrada a importância na área da bioacústica.

## 1.1 Contexto do trabalho

O monitoramento de espécies pode ser realizado *in loco*, em que há a presença do especialista no local (SCHUCHMANN *et al.*, 2018), ou ainda através da coleta dos sons das espécies, o que é importante não só para a conservação das espécies, mas também para conservação do ambiente de estudo. A coleta de sons tem a vantagem de não necessitar o especialista no local, o que permite a maior quantidade de dados a serem coletados. Espécies de aves são escolhidas para a coleta de sons devido a existência de mais 10.000 espécies no planeta Terra (UNWIN, 2011), o que torna possível o estudo dos sons de pássaros como bioindicadores em diversos segmentos (TABUR; AYVAZ, 2010). Dessa forma, é possível realizar o monitoramento de aves por até mesmo 24h, o que cria uma volumosa quantidade de dados, mas que não é possível analisar por um especialista em tempo viável para conservação do ambiente de estudo e das espécies. Esse monitoramento através dos sons ocorre por meio da bioacústica, pela qual ocorre a análise das vocalizações

após a coleta por equipamentos os quais são colocados no ambiente, como mostrado na Figura 1, o que é menos invasivo às espécies.

Figura 1 – Gravador acústico usado para monitorar o comportamento vocal da espécie Greater Rhea

Fonte: Pérez-Granados e SCHUCHMANN (2020)

Nesse sentido, para automatizar o processo de audição e reconhecimento de espécies, inúmeros métodos de Aprendizagem de Máquina têm sido utilizados para avanços nos estudos bioacústicos. Diversas publicações exemplificam a aplicabilidade e os impactos desses métodos, a exemplo dos trabalhos extraindo características distintas de um áudio feitos por Rai *et al.* (2016), Jadhav, Patil e Parasar (2020) e Ji, Jiang e Xie (2021). Além disso, outros estudos trouxeram abordagens com o avanço de Redes Neurais Profundas (RNP), como exemplificam Cakir *et al.* (2017) e Kahl *et al.* (2017).

Apesar dos avanços em reconhecimentos, tais modelos são denominados caixa-preta devido à sua estrutura de aprendizado com a descida de gradiente e respectiva mudança de pesos, e apesar de já existirem estudos para a interpretação desses modelos,

a quantidade de publicações da interpretabilidade das RNP é inferior na bioacústica em comparação com outras áreas. A importância da interpretabilidade desses modelos na biologia e consequentemente na bioacústica perpassa pelo nível de confiança a ser dado para um especialista, além da possibilidade de avanços e descobertas na área, construída através das decisões tomadas nos dados apresentados por esses algoritmos (FORTELNY; BOCK, 2020).

## 1.2 Objetivo

Este trabalho tem como objetivo descobrir os impactos de RNP treinadas para avaliação de dados de bioacústica e associar a aprendizagem dos modelos utilizados. Este trabalho realizará o treinamento de um modelo de RNP, interpretará os resultados e a partir dos resultados demonstrados, modificações serão realizadas para um novo treinamento do classificador, a fim de verificar o impacto e potencial do modelo. Assim, ao final, serão validadas as interpretações obtidas, o que fornece modelos interpretáveis para auxiliar especialistas na confiabilidade dos resultados e na realização de tomadas de decisões, bem como melhorar o modelo com processamentos futuros.

## 1.3 Estrutura do Trabalho

O restante deste trabalho segue da seguinte forma: No Capítulo 2 são apresentadas as formas de monitoramento de animais, com e sem uso de bioacústica automatizada, além dos conceitos de explicabilidade e interpretabilidade os quais serão seguidos, bem como o que são redes neurais interpretáveis e como serão utilizadas. O Capítulo 3 apresenta a espécie de ave a ser estudada e o conjunto de dados a ser utilizado, junto com a escolha do modelo e como serão avaliados os experimentos. No Capítulo 4 serão apresentados os resultados dos experimentos definidos e no Capítulo 5, as conclusões realizadas.

# CAPÍTULO 2

# FUNDAMENTAÇÃO TEÓRICA

Este capítulo se dedica a apresentação de trabalhos desenvolvidos para estabelecimento deste TCC, o que inclui o monitoramento de espécies de animais com o uso de outras metodologias além da bioacústica, e a apresentação de Redes Neurais, passando pela estrutura básica, profundidade e camadas convolucionais. Além disso, serão apresentados trabalhos desenvolvidos para a interpretabilidade de RNP, seus conceitos, suas vantagens e limitações. Também será apresentada qual a relação de alguns trabalhos para adaptação no estudo da bioacústica. Por último, serão apresentados métodos de interpretabilidade os quais serão utilizados neste trabalho, com suas definições e motivos, inicialmente sumarizados por Linardatos, Papastefanopoulos e Kotsiantis (2021).

## 2.1 Monitoramento de animais e bioacústica automatizada

Diversos trabalhos são realizados sobre o monitoramento de animais para a manutenção da biodiversidade. As abordagens perpassam por diferentes espécies de animais, seja por sua vocalização ou gravações visuais, até suas famílias e respectivas estações, além de ecossistemas diferentes. O ecossistema é composto por diversos organismos, cada um com sua tarefa, objetivo e consequências. Nesse sentido, há diferentes animais os quais permitem a continuidade desse ecossistema. O monitoramento é importante desde animais

no solo (veja Potapov *et al.* (2022)), até animais aquáticos (MATLEY *et al.*, 2022) e aves, como apresentado já neste trabalho.

Diferentes ecossistemas ao redor do mundo são estudados, cujos trabalhos realizados utilizam diferentes abordagens. Por exemplo, Harcourt *et al.* (2019) usam telemetria animal em áreas marinhas cobertas de gelo medindo o uso do habitat, fenologia de padrões migratórios, fatores bióticos e abióticos que impactam distribuições animais; e variáveis físicas ambientais. Em uma área do Cerrado, em Chapada dos Guimarães, Mato Grosso, Brasil, Pinheiro Saravy, Schuchmann e Marques (2021) estudam visitantes de flores (por exemplo, borboletas e abelhas), especificamente a pimenta-de-macaco escolhida, em estações secas e molhadas de modo quantitativo e qualitativo com a coleta de dados através de câmera fotográfica. No Pantanal, Senič *et al.* (2023) realizam o estudo da espécie de ave *Bare-faced Curassow* a qual tem grande necessidade de conservação e habita no Pantanal, com o uso de armadilhas fotográficas para entendimento da taxa de sexo da espécie, sua organização social e seus padrões de atividades.

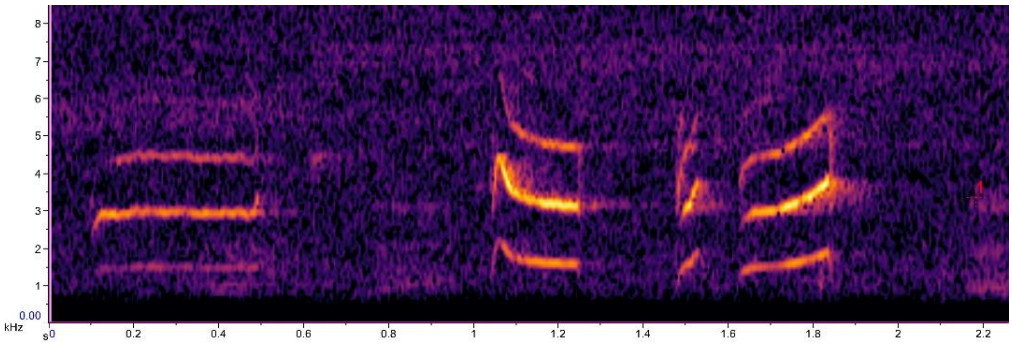
Sob outra metodologia, Pérez-Granados e Schuchmann (2022) utilizam a bioacústica, através do monitoramento acústico passivo com software de reconhecimento de sinal para estudo de um ciclo anual completo da espécie de ave *Band-tailed Nighthawk* no Pantanal, a qual possuía somente somente informações de anedotas e informações gerais. Com o estudo pôde ser confirmado a espécie ser residente do bioma, no entanto, mesmo com o uso do software para reconhecimento da espécie, foi necessária a validação manual para confirmação do que foi reconhecido. O software detectou mais de 75% das vocalizações anotadas por um especialistas. Essa validação ocorre não somente para espécies de aves, mas também com outros animais, como pode ser visto em Pérez-Granados e Schuchmann (2023) para a detecção de criações ilegais de gado no Pantanal.

Com o uso do software, os áudios não foram diretamente utilizados, pois podem ser representados como séries temporais em computadores, o que corresponde a uma composição de sequências de observações de sistemas medidas ao longo do tempo (AMINIKHANGHAHI; COOK, 2017). Assim, o número de sequências torna a tarefa de avaliar áudios custosa e complexa, o que é preferível utilizar o espectrograma ao invés do sinal que define o áudio. Os espectrogramas são compostos por faixas de frequência no eixo das ordenadas, enquanto o tempo compõe o eixo das coordenadas, nos quais para cada combinação possível destes eixos há um valor da intensidade do áudio (BADSHAH *et al.*, 2017). Assim, representam visualmente áudios, como pode ser visto na Figura 2, a qual mostra o espectrograma do chamado e canto da espécie típica do Pantanal Synallaxis Albilora.

Até aqui foram demonstradas diferentes formas de monitoramento, com diferentes modos para a conservação da biodiversidade. Mesmo com o uso de técnicas

automatizadas, como ocorre na bioacústica, há a necessidade de validação humana devido a não haver confiança completa das metodologias automáticas e o não entendimento das metodologias do que é esperado pelos especialistas.

Figura 2 – Espectrograma do chamado e canto da espécie Synallaxis Albilora.




Fonte: O Autor.

## 2.2 *Deep Learning*

A proximidade deste trabalho com a biologia encontra-se além do domínio escolhido, isto pois, inicialmente apresentado por McCulloch e Pitts (1943), as Redes Neurais Artificiais são inspiradas na rede neural biológica, cujas unidades básicas também são os neurônios. As conexões por dendritos aqui são definidas através dos pesos e o estímulo para um dado neurônio ocorre à partir de um valor maior que um limiar após os cálculos realizados por uma função de ativação.

A Rede Neural Artificial mais básica foi apresentada por Rosenblatt (1958): o *Perceptron*. Esta rede consiste de uma camada de entrada de dados, uma camada de processamento representada por um único neurônio e uma camada de saída, como pode ser visto na Figura 3. A camada de processamento (também chamada de *hidden layer*) realiza o produto escalar (representado pelo símbolo de somatório) das entradas e cada um dos pesos, os quais são a aprendizagem da rede, após isso a função de ativação $f(x)$ ativa o neurônio se e somente se (para o *Perceptron*) a saída for maior ou igual a 0, o que é representado pelo valor 1, e caso contrário, é propagado o valor 0 como saída $y$.

As Redes Neurais Profundas ou *Deep Learning* correspondem a redes neurais com mais de 1 camada de processamento (*hidden layer*), como pode ser visto na Figura 4. Ademais, múltiplos neurônios são utilizados para extração de características, com cada saída sendo utilizada como entrada para outro neurônio. Enquanto a aprendizagem do *Perceptron* pode ser definida através das equações de atualização do peso 2.1 e de variação do peso 2.2, com múltiplos neurônios vale salientar as diferenças na abordagem. Nas equações, $\eta$ é a taxa de aprendizagem, responsável por controlar a velocidade de

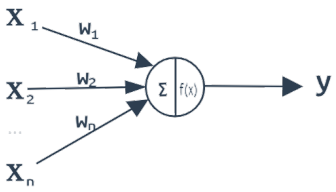
aprendizagem do modelo limitando quão "largos"devem ser os passos com cada iteração e $d$ é a saída esperada.

$$w_i = w_i + \Delta w_i \tag{2.1}$$

$$\Delta w_i = \eta x_i (d - y) \tag{2.2}$$

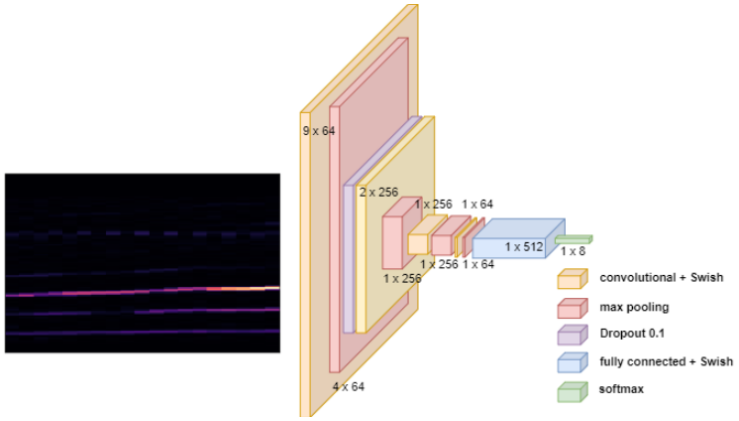
Com o uso de múltiplos neurônios e múltiplas camadas, a aprendizagem da rede ocorre através de *backpropagation*. Neste método, ps pesos da rede os quais são reajustados ao longo do treinamento do modelo, sofrem alterações à partir da computação do gradiente da função de erro com relação aos pesos a fim de corrigir e atualizar os valores para encontrar o erro mínimo global (ROJAS, 1996). O gradiente corresponde ao vetor de direção o qual determina a maior mudança de uma função (STEWART, 2007), nesse sentido, a direção de cada peso é mudada de modo a tentar permitir alcançar o erro mínimo global. Tentar pois, é possível que o gradiente levo a um erro mínimo local, de modo que a rede não consegue alcançar resultados melhores.

Figura 3 – Estrutura do *Perceptron*




Fonte: O autor

Figura 4 – Estrutura de uma Rede Neural Convolucional para classificação de sons de aves




Fonte: O autor

Existem diferentes trabalhos sendo desenvolvidos com *Deep Learning*, as quais abordam diferentes funções de ativação para aprendizagem do modelo, funções de perda, estruturação de camadas, arquiteturas, entre outros parâmetros possíveis. Para reconhecimento de áudios de pássaros, por exemplo, arquiteturas de Redes Neurais Recorrentes têm sido utilizadas (NOUMIDA; RAJAN, 2022), além de novas arquiteturas para tipos existentes de redes neurais (KIM, 2017). Em específico para este trabalho, serão tratadas as Redes Neurais Convolucionais (CNNs). As CNNs são arquiteturas de Redes Neurais as quais simulam o funcionamento de aprendizado do córtex cerebral para reconhecimento de imagens, nas quais ocorre o processo de convolução para extração de características ao longo de camadas ocultas que são cada vez mais específicas, devido à combinação de diferentes filtros para o extrator de características (MURPHY, 2016; KIM, 2017). Ademais, as Redes Neurais Convolucionais têm demonstrado serem estado-da-arte em classificação de imagens (KHAN; LAGHARI; AWAN, 2021), o que mantém-se para a área de estudo desse trabalho, com grande impacto na classificação de aves, como pode ser visto em (KAHL *et al.*, 2017; HIDAYAT; CENGGORO; PARDAMEAN, 2021).

Mesmo com abordagens as quais provaram ter melhorado o desempenho do modelo, os métodos tradicionais não conseguem determinar sobre um determinado conjunto de dados quais foram os aprendizados adquiridos para a melhoria, ou ainda uma estratégia para melhorar entender os resultados e reduzir o escopo de tentativas a serem tomadas para melhoria do modelo. Nesse sentido, métodos de interpretabilidade podem auxiliar com a determinação de uma heurística para tomar decisões com o conjunto de dados sob análise.

## 2.3 Explicabilidade e Interpretabilidade

Para a avaliação das formas de interpretabilidade e seus impactos em RNP, é necessário antes situar os conceitos de interpretabilidade e explicabilidade, pois ambos os termos são utilizados como sinônimos em diversas situações. Doshi-Velez e Kim (2017) revisou as definições em diferentes áreas, não limitadas somente à computação, o que permitiu pontuar a interpretabilidade como a capacidade de tornar algo entendível ou explicável a nível humano. Tal entendimento ou a explicabilidade salientam a necessidade do conhecimento do que é necessário no domínio o qual está utilizando Aprendizagem de Máquina, para eliminar as barreiras para avaliação ou otimização do domínio no qual está sendo aplicado, *trade-offs* possíveis ou ainda questões éticas e de segurança.

Miller (2019) realizou o estudo de explicabilidade e interpretabilidade sob óptica das Ciências Sociais e pontua a explicabilidade como a resposta para um dado acontecimento em função de dar motivação à esta resposta e demonstrar o porquê não é outra. Já a interpretabilidade define o quanto é possível entender a causa de um acontecimento e

assim, reforça a diferença entre os dois termos, em que, o entendimento da causa pode ser uma explicação.

As diferentes definições ocorrem principalmente pela falta de formalismo matemático para a definição de cada um, o que possibilita a definição de conjuntos de termos semelhantes a fim de tornar claro o entendimento da aprendizagem de modelos de Aprendizagem de Máquina (ADADI; BERRADA, 2018). Ainda assim, este trabalho distingue interpretabilidade e explicabilidade da mesma que apresentou Miller (2019), o que é necessário para demonstração dos materiais encontrados e a metodologia proposta para elaboração do trabalho o qual será desenvolvido.

Apresentadas as definições, existem escopos para abordar a interpretabilidade no domínio de estudo. Esses escopos demonstram os tipos de interpretabilidade possíveis de realizar. Já foi comentado por Doshi-Velez e Kim (2017) anteriormente motivações para a interpretabilidade, mas Linardatos, Papastefanopoulos e Kotsiantis (2021) agrupa em 4 principais tipos a interpretabilidade.

(i) O objetivo é explicar os modelos através das relações entre saídas de uma ou mais partes de um modelo, em RNP, as camadas, ou seja, entender a saída das camadas a partir de uma entrada.

(ii) Corresponde à criação modelos os quais possam explicar suas decisões, o que independe de entradas, a exemplo de *AutoEncoders*.

(iii) Perpassa por restringir questões discriminatórias e tornar justo o uso de modelos de Aprendizagem de Máquina, principalmente, através de pré-processamento para remoção de *bias* do conjunto de dados.

iv) Verifica a sensibilidade das predições de modelos, de modo que é possível perceber o quanto um dado pode sujar ou impactar o conjunto inteiro para o treinamento dos modelos.

## 2.4 Redes Neurais Interpretáveis

Apresentados os conceitos e esforços para os estudos acerca de interpretabilidade, é possível diminuir o escopo para investigar avanços com RNP. Os trabalhos realizados nos últimos anos para o avanço da interpretabilidade em RNP ocorreram em diversas áreas da biologia, como na genômica (FORTELNY; BOCK, 2020; KOO; PLOENZKE, 2020), biologia computacional (YUAN *et al.*, 2021), entre outras. Estes trabalhos perpassam pelo entendimento da interpretabilidade na área de domínio, a aplicação, entendimento, limitações e motivações para a ocorrência. Trabalhos anteriores já investigaram

o uso de interpretabilidade em RNP na acústica. Para antes entender as metodologias destes trabalhos, alguns conceitos devem ser apresentados: *Prototypical Networks* e *Auto Encoders*.

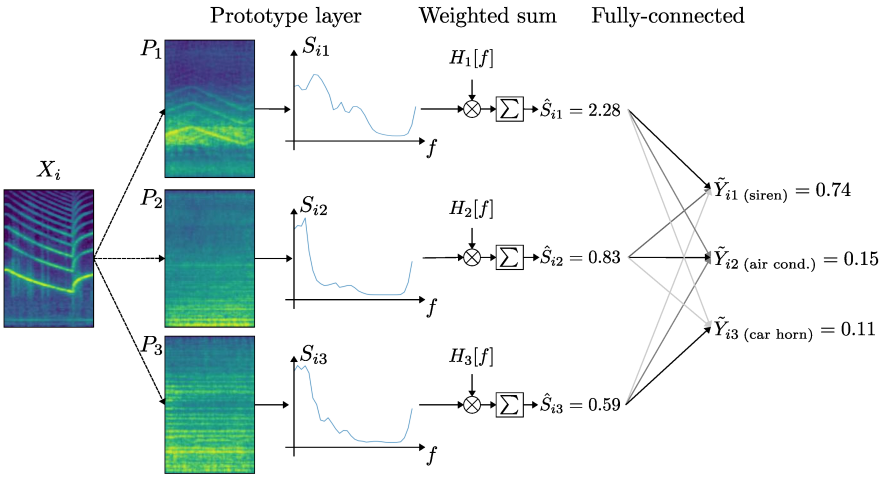
*Prototypical Networks* ou Redes Neurais de Protótipos, inicialmente apresentadas por Snell, Swersky e Zemel (2017), computam à partir de uma função de distância escolhida (por exemplo, euclidiana ou similaridade de cossenos) um espaço de representação (ou protótipo) de cada classe, em que cada protótipo é o vetor médio dos pontos pertencentes a esta classe. Inicialmente elaboradas para trabalhar com conjunto com poucas quantidades de dados, com o mapeamento de quais pontos são mais representativos dada uma consulta de uma entrada, tal método serviu como base para outros trabalhos além dos promovidos na acústica que serão aqui apresentados, por exemplo, para imagens (LI *et al.*, 2018).

*AutoEncoders* são redes neurais as quais aprendem a extrair informações mínimas necessárias para que seja possível identificar uma determinada entrada. Isso é possível à partir de dois componentes principais dessa rede: *encoder* e *decoder*. Enquanto há diferença na origem e composição do termo *AutoEncoders* ao longo do tempo (RUMELHART; HINTON; WILLIAMS, 1985; BALDI, 2012), é possível determinar a evolução destas redes tanto para aprendizagem não supervisionada com a redução de dimensionalidade e extração de características necessárias para representar a entrada, quanto para representações as quais retiram ruídos. O *encoder* é utilizado para extração de características as quais são representadas no espaço latente, enquanto o *decoder* recompõe as características extraídas deste espaço de modo que a saída tem a mesma dimensão da entrada, no entanto, a rede aprendeu à partir do erro computado entre a reconstrução e o dado original, quais características foram necessárias e são mais importantes para esta reconstrução.

Apresentados os conceitos necessários, é possível prosseguir com a revisão de trabalhos os quais seguem esta abordagem para o domínio acústico. Zinemanas *et al.* (2021) apresenta uma rede a qual reconstrói mel-espectrogramas em escala logarítmica através de *AutoEncoders* para a determinação de protótipos. Para a determinação de protótipos, a similaridade foi computada atribuindo um peso para cada faixa de frequência no espaço latente, de modo a ser verificada a importância de cada faixa de acordo com a atividade no áudio. Em vista da reconstrução de entradas (mel-espectrogramas), é possível inspecionar a saída do *decoder*, a interpretabilidade ocorre através do mapeamento do protótipo para o domínio tempo x frequência, ou seja, cada protótipo associado a entrada pertencente a uma classe é o mel-espectrograma com maior similaridade ao áudio passado para ser realizada a classificação. A arquitetura proposta foi destinada ao domínio sonoro, avaliando em diferentes conjuntos de dados o modelo proposto, e devido a natureza do modelo, caracteriza-se por ser do tipo (ii). A Figura 5 mostra como o modelo realiza

as predições, cujo exemplo considera as classes de sons de sirenes, ar condicionado e buzina de carro. A entrada é comparada a um protótipo de cada classe a fim de verificar a similaridade de dependência de frequência que é posteriormente integrada usando a camada de soma ponderada definida que projeta a similaridade através de uma camada não convolucional para classificar a entrada.

Figura 5 – Funcionamento da arquitetura da Rede Neural de Protótipos para o domínio acústica




Ainda nesse tipo de interpretabilidade, Ren, Nguyen e Nejdl (2022) apresenta também uma RNP de protótipos, mas no domínio de classificação de sons respiratórios e utiliza similaridade de cossenos para comparação dos exemplos e protótipos a fim de determinar os protótipos com maior semelhança, e não *AutoEncoders*. Deste modo, após as extrações de características pelo *encoder* (utilizadas CNNs no trabalho), são calculadas as similaridades entre pares de faixas de tempo x frequência para cada protótipo e entrada para serem utilizadas nas próximas camadas do modelo para a etapa de classificação.

Em virtude dos esforços apresentados no domínio acústico com o tipo (ii), esse trabalho objetiva realizar o tipo (i) de interpretabilidade, em específico à bioacústica.

## 2.5 Interpretabilidade *Post-hoc*

Métodos de interpretabilidade do tipo (i) também são conhecidos como *post-hoc* (ZINEMANAS *et al.*, 2021). Existem inúmeras abordagens existentes na literatura e para a limitar as possibilidades, neste trabalho, foram considerados dois algoritmos para cada um dos métodos baseados em gradientes locais ou referências. Dentre os métodos de referências o primeiro pode ser classificado como gradiente *global* e o segundo como **conceito**. Os métodos baseados em gradiente a serem explicados nos próximos parágrafos

assimilam a propagação da direção positiva ou negativa para determinação de características mais significativas para cada entrada.

Os métodos escolhidos nomeados gradientes locais podem assim ser definidos pois não utilizam conjunto de referências a ser considerado para avaliar a contribuição de camadas, sinais dos dados ou funções sobre as entradas a serem recebidas. Deste modo, considera-se que há um único conjunto para o qual um algoritmo é usado. Nesse sentido, Zeiler e Fergus (2014) propõe inverter o mapeamento das imagens para mapas de características da imagem de múltiplas formas, guardando as localizações dos pixels mais intensos para que seja possível reconstruir a imagem, visto que a operação de *pooling* é inversível. É comum utilizar funções ReLU (*Rectified Linear Unit*) após a operação de *pooling* para garantir a saída positiva, o que também é realizado após a reconstrução da entrada. Por último, para inverter filtros já processados, foi realizada a transposição dos filtros para serem utilizados para os mapas os quais sofreram a operação ReLU.

Springenberg *et al.* (2014) aponta como um dos problemas possíveis para esta metodologia a propagação de gradientes negativos de funções ReLU, o que os autores resolveram através do que foi chamado como *guided backpropagation*, que corresponde a operação de *backpropagation* das saídas, no entanto, com a sobreposição de valores negativos de ReLU por 0. Com ambos os métodos é possível visualizar diferentes camadas convolucionais, de níveis mais altos a mais baixos.

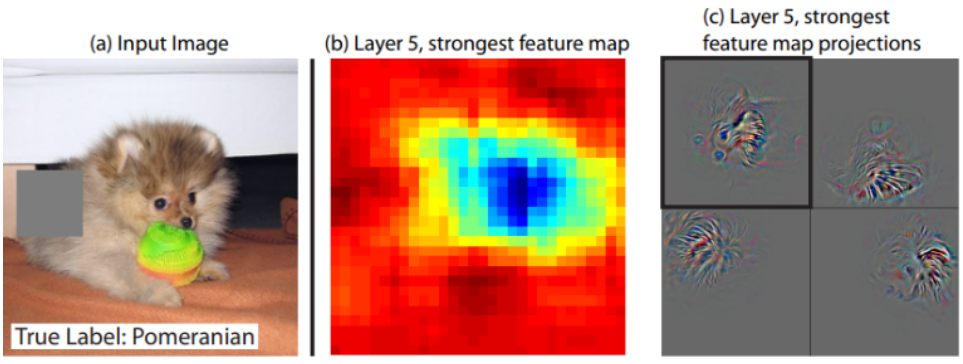
Com relação a métodos de referência, foram escolhidos os trabalhos realizados por Shrikumar, Greenside e Kundaje (2017) e Kim *et al.* (2018). Shrikumar, Greenside e Kundaje (2017) definiram o *DeepLift*, o qual utiliza uma entrada de referência à qual deve ser estabelecida através do domínio de conhecimento abordado, para determinar regiões de importância da entrada original separando pontos positivos e negativos de contribuição às entradas. As contribuições são calculadas considerando multiplicadores, os quais definem o quanto um neurônio de entrada contribui em relação ao neurônio de saída, e regras para camadas lineares (regra linear) e camadas não lineares (regra de reescala), possibilitando estarem em qualquer camada da rede, apesar de não ser recomendado utilizar na última camada para modelos com saídas de ativação como *sigmoid* ou *softmax* como mostrado no trabalho. Além disso, é proposta também a regra *RevealCancel* a qual separa contribuições positivas de contribuições negativas em operações de funções não lineares, que possibilita resolver problemas em que a ordem de cálculo de atribuição das funções muda a contribuição.

Kim *et al.* (2018) perpassa além da etapa de atribuição de características (o que é nomeado por **conceito** pelos autores), com o estabelecimento de um método de teste quantitativo para avaliação dos interpretações obtidas (*Testing with Concept Activation Vectors* - TCAV). Com a definição de um conceito alvo e um conjunto de exemplos

*s* passados através para uma camada, é encontrado um vetor cujo conceito representa *s* (Vetor de Ativação de Conceito), considerando as ativações da camada, tal vetor é normal ao hiperplano e separa exemplos conceitualmente definidos e não conceitualmente definidos (aleatórios). A sensitividade e melhoria a esses conceitos são melhorados através de derivadas direcionais, o que permite também estar em qualquer camada de CNNs. Realizados estes passos, testes de significância estatística foram realizados, os autores elaboraram testes *t* de *student* bilateral cuja hipótese nula é de que o conceito escolhido não está relacionado à predição da classe escolhida.

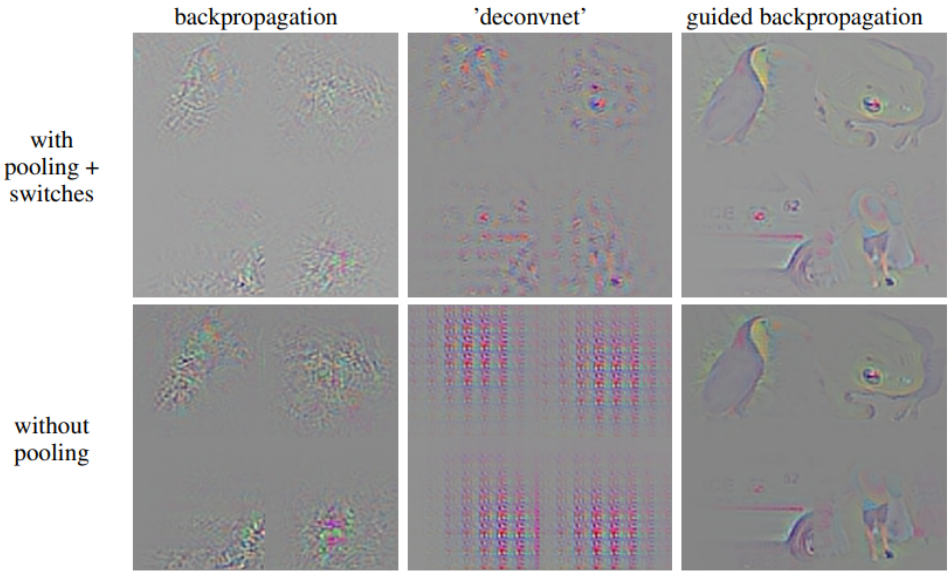
As Figuras 6, 7, 8, e demonstram o funcionamento de cada um dos métodos apresentados. Para o *Deconvolution*, é demonstrado quais características foram ressaltadas quando a imagem foi propositalmente coberta com um quadrado cinza, cujo destaque dado pelo modelo mostrou ser referente a imagem do cachorro. A Figura 7 compara os dois métodos apresentados baseados em gradientes locais, em que mais detalhes são apresentados com a reconstrução da imagem usando o *Guided Backpropagation*. A Figura 8 demonstra pontos ressaltados para uma imagem de barco quando utilizada uma imagem borrada do barco e uma imagem toda em preto com o método *DeepLift* cuja explicação possível para o maior destaque para o fundo é que características como o oceano ou paisagem foram relevantes para classificar a imagem, enquanto o formato foi destacado quando a imagem ficou embaçada. Por último, a Figura 9 mostra duas imagens *adversárias* a da classe zebra para diferentes camadas do modelo, os autores forçaram as duas imagens a serem classificadas como zebra, e apesar disso, mostra a diferença em quais conceitos tiveram mais impacto para classificar cada imagem, em que a combinação de listras e *zig-zags* foi única para a imagem correta de zebra, o que pode ser utilizado para verificar entradas as quais fogem do padrão de contribuição para determinados conceitos utilizados de acordo com a base de dados.

Figura 6 – Visualização de ativações para a 5º camada da rede usando o *Deconvolution* da rede proposta pelos autores



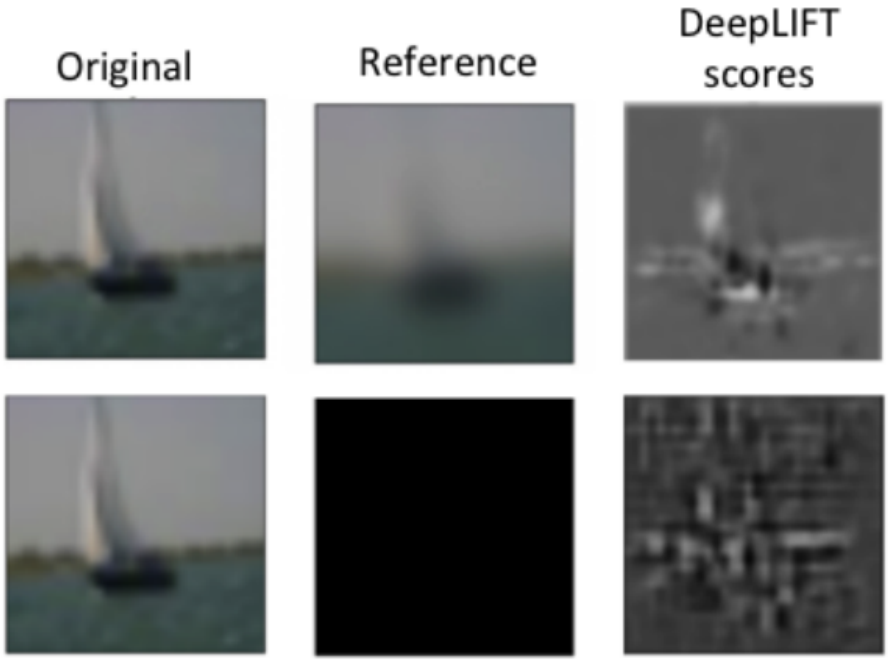

Fonte: Adaptada de Zeiler e Fergus (2014)

Figura 7 – Comparação da *backpropagation* original e dos métodos *Deconvolution* e *Guided Backpropagation* para a última camada convolucional




Fonte: Springenberg *et al.* (2014)

Figura 8 – Pontos ressaltados para uma imagem de barco quando utilizada a mesma imagem borrada e quando é totalmente obscurecida com o método *DeepLift*




Fonte: Shrikumar, Greenside e Kundaje (2017)

Figura 9 – Importância de diferentes conceitos para a classificação correta e incorreta de zebra com o *TCAV*


Fonte: Kim *et al.* (2018)

# CAPÍTULO 3

# METODOLOGIA

Este capítulo aborda as atividades as quais foram desenvolvidas neste trabalho, o que inclui como os conceitos abordados neste trabalho foram utilizados. Será também abordada *Transfer Learning*, pois tem demonstrado aceleração no processo de treinamento para obtenção de um classificador viável e potencial de melhoria na classificação de sons (DUFOURQ *et al.*, 2022). Será listada a ave de observação para estudo da interpretação dos modelos, bem como o conjunto de dados necessário. Em seguida, serão explicados os métodos de pré-processamento dos dados, em complemento aos espectrogramas já explicados anteriormente. Logo após, será apresentado o modelo utilizado para *Transfer Learning*, motivo e expectativa. Por último, serão apresentadas as formas de avaliação dos resultados, como a interpretabilidade será utilizada a fim de comparar o desempenho do modelo à partir das análises realizadas e as ferramentas utilizadas para desenvolvimento do trabalho.

## 3.1 Espécie de estudo e conjunto de dados de bioacústica

Antes de realizar a seleção do conjunto de dados de bioacústica, foi definida a espécie *Synallaxis Albilora* (S. Albilora) para estudo do seu comportamento e identificação do ambiente escolhido em meio a outras espécies de aves, devido à dificuldade em modelar o som dessa ave e por ser um pássaro comum da região do Pantanal e haver poucas

informações sobre a ecologia dela, como pode ser visto em Rubio e Pinho (2008). Os sons do S. Albilora foram separados em três classes pelas quais há destaque em sua vocalização, o canto, a primeira sílaba do canto e a segunda sílaba do canto. Esses áudios totalizam 10 minutos e 10 segundos, com duração média de 0,36 segundos ao longo de 1683 trechos.

As vocalizações da espécie ocorrerão com a presença de sons de fundo, em diferentes ambientes e estações, com maior ou menor intensidade. Assim, foram coletados também somente os sons desses ambientes para que haja a distinção da vocalização com diferentes intensidades de som externo e de sons do ambiente que a ave pertence. Os sons ambientes foram representados a partir de 4 classes as quais correspondem aos biomas savana e floresta do Pantanal, em ambos períodos de estação, chuvoso e seco, referenciados como BG-savanna-Wet, BG-savanna-Dry, BG-forest-Wet e BG-forest-Dry, onde *BG* é abreviatura para *Background*. Cada classe contém 4 gravações com duração média de 6 minutos, o que totaliza 1.6 horas para os áudios de ambiente.

Além disso, é preciso que o modelo saiba distinguir as vocalizações de S. Albilora de outras espécies de pássaros. Nesse sentido, foi selecionado o conjunto de outras espécies no mesmo ambiente, proposto inicialmente por Ventura *et al.* (2015), aqui mencionado como others, o qual possui 1532 gravações, em que cada uma possui pelo menos uma espécie de ave vocalizando. Esse subconjunto tem duração média de aproximadamente 30 segundos e totaliza 12.8 horas. A base de dados foi dividia em 10 subconjuntos (*folders*) contendo aproximadamente a mesma quantidade de dados, no qual 9 destes foram utilizados para treino e 1 para validação dos resultados do treinamento. A Tabela 1 apresenta as informações do conjunto e sua organização.

Tabela 1 – Organização do conjunto de dados. CA = Canto, SYL1= Primeira sílaba do canto, SYL2= Segunda sílaba do canto. Uso para treinamento e validação do modelo

| Classe | Duração Média (s) | Quantidade | Duração (s) |
|---|---|---|---|
| albilora-CA | 0,44 | 205 | 89,6 |
| albilora-SYL1 | 0,34 | 759 | 255,37 |
| albilora-SYL2 | 0,37 | 719 | 264,68 |
| BG-forest-Dry | 293,74 | 4 | 1.174,97 |
| BG-forest-Wet | 300 | 4 | 1.200 |
| BG-savanna-Dry | 582,11 | 4 | 2.328,46 |
| BG-savanna-Wet | 267,88 | 4 | 1.071,51 |
| others | 30,1 | 1.532 | 46.109,28 |

Os áudios do conjunto de dados possuem diferentes frequências de amostragem, variando de 24 kHz a 72 kHz. Nesse sentido, os áudios foram reamostrados a 24 kHz, visto que atividades do Synallaxis Albilora ainda ocorrem até 12 kHz com maior intensidade e outras atividades com frequência acima foram desconsideradas. Seja $d$ a duração padrão de

quanto deve ser carregado de cada áudio, cada um dos áudios de ambientes foram divididos em 125 trechos de $d$ segundos. Para cada áudio de aves, foi selecionado 1 trecho aleatório de $d$ segundos. Inicialmente, foi considerado $0.125s$ somente para os ambientes, por ser o tempo de duração mínimo da base de dados. Para os áudios cuja duração seja menor que $d$, foi adicionado *padding* de zeros para o áudio, em que o trecho foi posteriormente deslocado aleatoriamente para uma posição $r$ aleatória de início de atividade acústica.

## 3.2 *Transfer Learning*

Modelos já treinados em um conjunto de dados têm grande implicância para classificação de outros conjuntos de dados, pois os pesos já foram calculados para cada camada devido ao processamento anterior em que, muitas vezes, é feito com uma grande quantidade de dados, o que evita tempo para encontrar uma arquitetura e parâmetros que satisfaçam ao conjunto alvo de dados (DIMENT; VIRTANEN, 2017). Kong *et al.* (2020) propuseram modelos para diferentes atividades possíveis no domínio de reconhecimento de áudios, como detecção de evento de som, classificação de cena acústica, *audio tagging*, classificação de músicas, classificação de emoções de fala, dentre outras. Todas as atividades realizadas foram estabelecidas sobre redes neurais pré-treinadas na base de dados do *AudioSet* (*Pretrained audio neural networks* ou *PANNs*).

Dentre os diferentes modelos propostos para cada atividade, foi selecionada a arquitetura *CNN14* por obter o estado-da-arte em diferentes atividades em comparação com outros modelos com a transferência do aprendizado. A arquitetura consiste de 6 blocos convolucionais, cada um composto por 2 camadas convolucionais de *kernel* 3x3. Além disso, *Batch Normalization* é aplicada entre cada camada dentro do bloco com uso seguinte da função de ativação *ReLU*. Também é utilizada *average pooling* 2x2 após o final de cada bloco, com incremento de *global pooling* após o último bloco convolucional. Por último, é aplicada uma operação máxima e média sobre os dados antes de serem passados para a camada preditiva ou de *embedding* para transferir aprendizado.

É esperada a maior capacidade de generalização das classes, de modo que não haja necessidade de descongelar outras camadas e seja treinada somente a camada com as classes apresentadas na Tabela 1.

## 3.3 Avaliação e Interpretação

A fim de avaliar os modelos treinados, foram utilizadas a acurácia (equação 3.1), precisão (equação 3.2), recall (equação 3.3) e $f1 - score$ (equação 3.4), onde *TP (True Positive)* é o número de acertos da classe alvo, *TN (True Negative)* é o número de acertos das outras classes, *FP (False Positive)* é o número de falsos positivos, *FN (False*

*Negative)* é o número de falsos negativos e $N$ é o número de instâncias para a base sendo classificada.

$$Acurácia = \frac{TP+TN}{N}100[\%] \tag{3.1}$$

$$Precisão = \frac{TP}{TP+FP}100[\%] \tag{3.2}$$

$$Recall = \frac{TP}{TP+FN}100[\%] \tag{3.3}$$

$$F1 = 2 \times \frac{Precisão \times Recall}{Precisão + Recall}100[\%] \tag{3.4}$$

Essas equações listadas foram utilizadas para respectivamente: ter uma visão geral do quanto o modelo acertou considerando todas as classes, quão precisas foram as detecções, qual a capacidade geral de detecção das classes, e quão balanceada está a capacidade de detecção e a qualidade de detecção do modelo.

Após os resultados para avaliação da capacidade do modelo, foram utilizados algoritmos de interpretabilidade para entendimento dos aprendizados adquiridos, para que fossem alterados os métodos para processamento dos dados e modificada a base de dados, com diferentes expectativas para cada um.

Ambos os métodos baseados em gradiente locais foram utilizados para analisar a importância de cada período do áudio para o modelo, além da importância dos áudios e suas respectivas classes com maior ou menor atribuição para o modelo, com visualização nos blocos 1, 2 e 5 a fim do entendimento da extração de características gerais para mais específicas.

O método *DeepLift* foi utilizado neste trabalho pois além de possibilitar o entendimento do que foi aprendido pelo modelo, é possível também entender quais foram as influências das entradas, suas classes e as características de cada entrada. Com o método de Vetores de Ativação de Conceito, foram realizados testes de hipótese cujo objetivo foi identificar e comprovar as identificações de quais **conceitos** detêm maior ou menor impacto em comparação com um ou mais conceitos definidos.

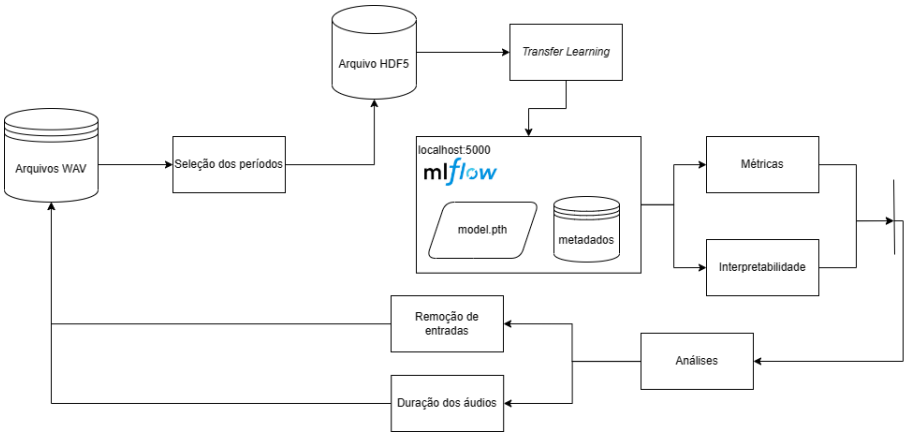
## 3.4 Planejamento de operações

A sequência de operações que define como os resultados foram obtidos e respectivamente como foram analisados para melhoria do desempenho do modelo pode ser conferida na Figura 10. Com os áudios coletados da base de dados em arquivos *.wav* e

selecionados como já foi explicado anteriormente, é criado um arquivo *hdf5* para posterior treinamento do modelo. Antes da inicialização do treinamento, é iniciado um servidor local de rastreamento com o *mlflow* para salvar o modelo treinado com a última camada adaptada à classificação dos áudios, utilizando o *Pytorch* (PASZKE *et al.*, 2019), juntamente com os parâmetros para construção dos espectrogramas e o desempenho para cada iteração do modelo.

Ao fim do treinamento, são calculadas as métricas para a última iteração do modelo. Logo após a interpretabilidade é realizada a partir da biblioteca *captum* da linguagem de programação *python* (KOKHLIKYAN *et al.*, 2020), em que até o momento da escrita deste trabalho, não há suporte para a regra *RevealCancel* do método *DeepLift*. Com os resultados obtidos, foram realizadas novas modificações, em que as mais significativas foram a remoção de entradas e a seleção de diferentes trechos do áudio, como será descrito no próximo capítulo.

Figura 10 – Fluxo de atividades do trabalho. Processos desde a coleta dos dados até o desenvolvimento de novos experimentos com os resultados obtidos



# CAPÍTULO 4

# RESULTADOS E DISCUSSÕES

Este capítulo aborda os resultados obtidos através das etapas definidas nos capítulos anteriores. Serão apresentados os resultados à partir de um experimento inicial, e os melhores resultados obtidos à partir das análises realizadas com os métodos de interpretabilidade previamente definidos neste trabalho. Os resultados obtidos à partir da interpretabilidade serão comparados com o experimento inicial, apresentando possíveis motivos para a melhoria dos resultados. Todos os códigos utilizados para o desenvolvimento deste trabalho e obtenção dos resultados são públicos e encontram-se hospedados no repositório do *Github* `https://github.com/SousaPedroso/PIDL`.

Inicialmente, foi realizado um experimento com os mesmos parâmetros definidos por Kong *et al.* (2020). Assim, foram consideradas somente atividades no áudio entre 50Hz e 14kHz, com espectrogramas em escala mel com 64 faixas, com 1024 amostras da *Short Time Fourier Transform* cuja janela utilizada foi a *hann* de 1024 amostras correspondente a 32$ms$ e sobreposição de 320 amostras as quais permitem o deslocamento da janela em passos de 22$ms$. Além do mais, devido à reamostragem na seleção do áudio à 24 kHz, a utilização da taxa de amostragem de 32kHz das *PANNs* permite o "zoom"do áudio com redução da faixa temporal e maior qualidade de frequência do áudio, de modo que os áudios para o modelo contêm 0.75$s$.

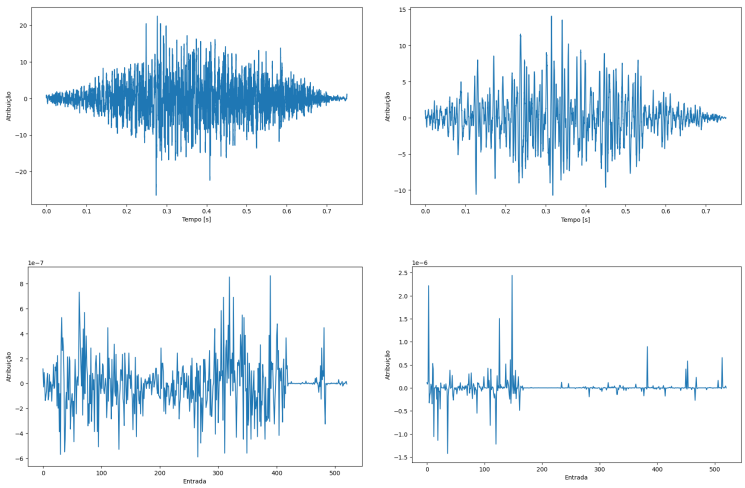
Foram realizadas 100 iterações nesse experimento, em que pôde-se observar convergência do modelo após 70 iterações com detecção apenas da classe *others*. A acurácia, precisão, *recall* e *F1* foram respectivamente 29,36%, 3,67%, 12,5% e 5,67%. Para verificar as possíveis motivações do ajuste do modelo a somente uma classe, foram definidos diferentes parâmetros a serem analisados para os métodos de interpretabilidade já definidos.

Com os métodos baseados em gradientes locais, os impactos em blocos foram medidos através da construção de uma distribuição *bootstrap* de 10000 amostras à partir de 360 neurônios de cada um dos blocos obtendo como estatística de amostra a média de cada neurônio de um bloco dado o conjunto de validação, devido ao tempo de processamento para todos os neurônios de um bloco. Em relação aos métodos de referência, além de ter sido verificada a importância dada a cada *frame* do áudio considerando o *DeepLift* com a regra *Rescale*, foi considerada também a importância atribuída aos áudios dadas diferentes características a serem atribuídas considerando as faixas de frequência: 2kHz-4kHz, 4kHz-6kHz, 6kHz-8kHz e 8kHz-10kHz; e como ocorre a importância para características aleatórias, aqui escolhidos: ruído gaussiano e impulsivo, onde o primeiro ressalta ruídos constantes e o segundo ruídos periódicos.

A Figura 11 demonstra a importância do modelo para cada período do áudio considerando os algoritmos *Deconvolution* e *Guided BackPropagation*. O eixo *Y* dos gráficos corresponde ao gradiente calculado à partir de cada um dos métodos, à partir daqui mencionado **Atribuição**. Já o termo **Entrada** no eixo *X* corresponde as entradas usadas no conjunto de validação.

Figura 11 – Atribuição média para cada momento do áudio em segundos

(a) *Deconvolution* (b) *Guided BackPropagation*

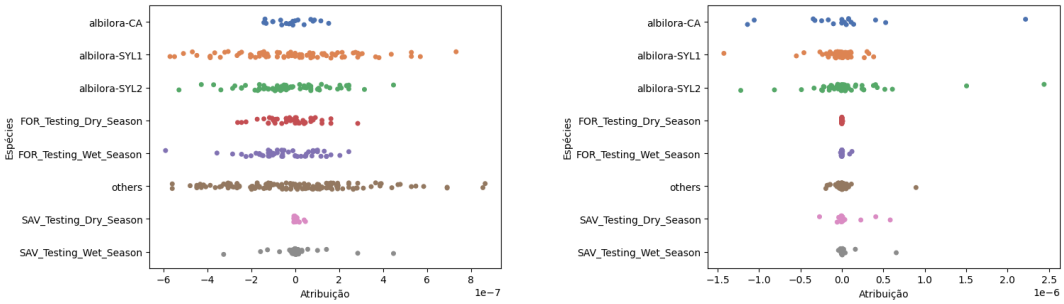



Fonte: O autor.

É possível observar diferentes picos positivos e negativos de atribuição para diferentes momentos do áudio, mas há maior amplitude entre 0.1 segundos e 0.55 segundos (aproximadamente). Além disso, os métodos divergiram para atribuir quais entradas do conjunto de validação tiveram gradiente positivo ou negativo através de *backpropagation*, em que, dentro da escala de cada método, a semelhança entre os dois métodos ocorreram principalmente após 400 entradas com valores extremamente baixos, salvo algumas entradas. Mais especificamente, ambos os métodos expressaram a maior parte das atribuições menor ou igual a 0 para as classes, como mostra a Figura 12.

Figura 12 – Atribuição média para cada classe

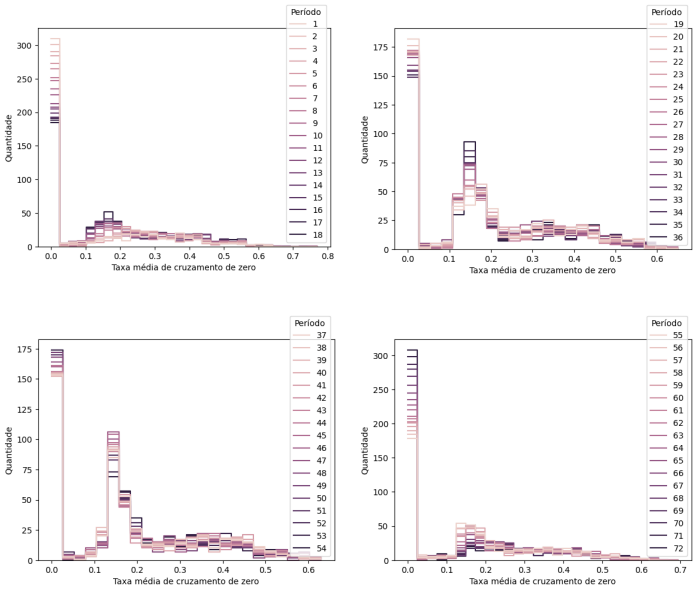
(a) *Deconvolution* (b) *Guided BackPropagation*




Fonte: O autor.

Em confirmação ao período de atribuição, foi calculada a taxa de cruzamento de zero para entendimento da atividade em cada áudio, em que maiores valores podem indicar maior mudança entre a presença ou não de atividade sonora. Para esse cálculo, os áudios foram separados em períodos de 0.18 segundos, cujo histograma na Figura 13 demonstra maior variação para a faixa de período citado, em que cada período foi considerado em $10^{-1}s$ para melhor visibilidade.

Figura 13 – Taxa média de cruzamento de zero em períodos de 0.18$s$

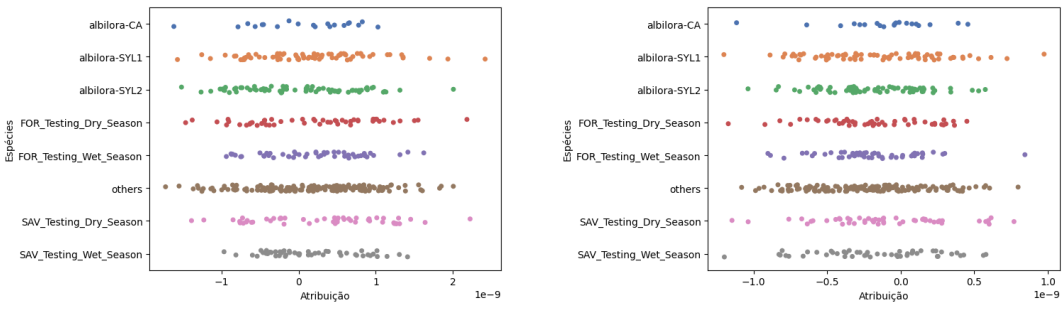



Fonte: O autor.

À partir da distribuição obtida para cada um dos 3 blocos escolhidos, foi realizada a análise da atribuição a fim de comparar com os resultados de todo o modelo. As Figuras 14, 15 e 16 mostram a atribuição ao longo do 1º, 2º e 5º bloco, em que o modelo demonstrou aumentar o número de atribuições cujo valor fosse maior ou igual a 0. Uma possível explicação é que o modelo não conseguiu distinguir todas as características juntas previamente à camada de classificação, o que pode explicar o resultado para a atribuição geral do modelo. Ademais, mesmo com a distribuição *bootstrap*, a mesma não corrige vieses da amostra obtida à partir dos 360 neurônios de cada bloco. Dadas essas duas observações, foram estabelecidos os métodos baseados em referência.

Figura 14 – Atribuição média para o 1º bloco de cada método

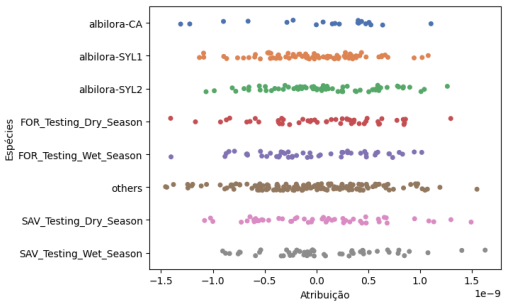
(a) *Deconvolution* (b) *Guided BackPropagation*



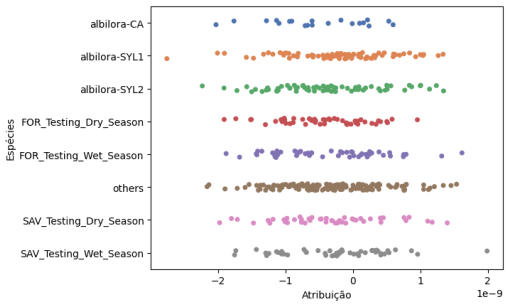

Fonte: O autor.

Figura 15 – Atribuição média para o 2º bloco de cada método

(a) *Deconvolution*



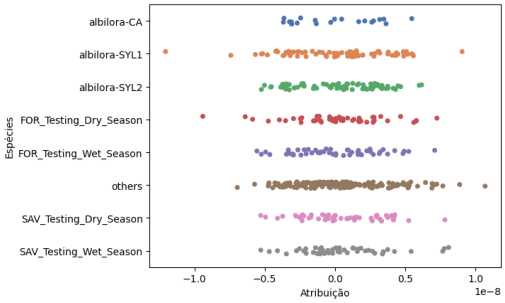

(b) *Guided BackPropagation*



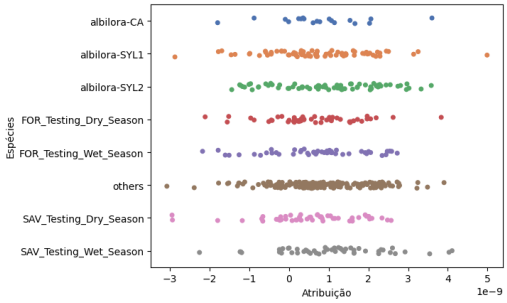

Fonte: O autor.

Figura 16 – Atribuição média para o 5º bloco de cada método

(a) *Deconvolution*




(b) *Guided BackPropagation*

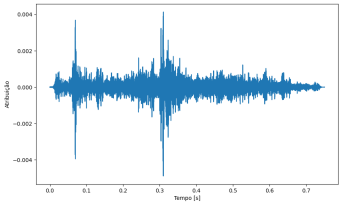



Fonte: O autor.

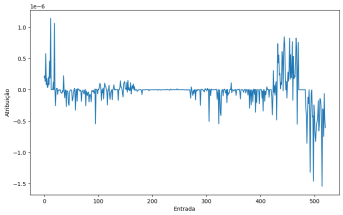
O método *DeepLift* demonstrou menor variação de importância para diferentes períodos dos áudios, com picos entre $0s$-$0.1s$ e $0.3s$-$0.4s$, como demonstrado na imagem à esquerda da Figura 17; a imagem à direita mostra a atribuição para cada entrada, em que as atribuições negativas tiveram mais impacto como mostrado nos métodos anteriores, no entanto, as últimas entradas mostraram maior atribuição negativa ao modelo.

Figura 17 – Atribuição média com o algoritmo *DeepLift*

(a) Atribuição para os períodos dos áudios



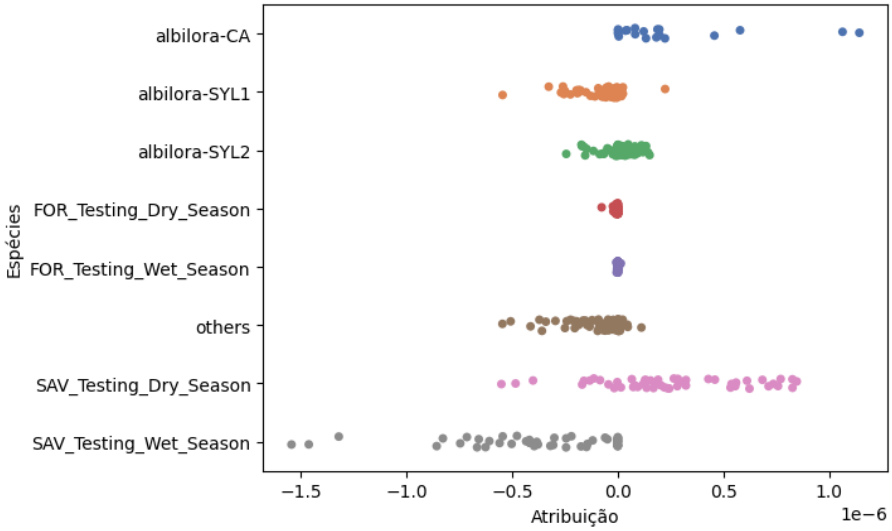
(b) Atribuição para o conjunto de validação



Fonte: O autor.

Da mesma forma como nos outros métodos, foi checada a distribuição da atribuição para cada classe da base de dados. A Figura 18 demonstra permanência de mais entradas com atribuição negativas, no entanto, em maior quantidade em comparação com os outros dois métodos. Além disso, as classes para as sílabas do canto da ave S. albilora e o período chuvoso do bioma savana também demonstraram maior quantidade de entradas com atribuição negativa.

Figura 18 – Atribuição para cada classe do conjunto de validação com o *DeepLift*

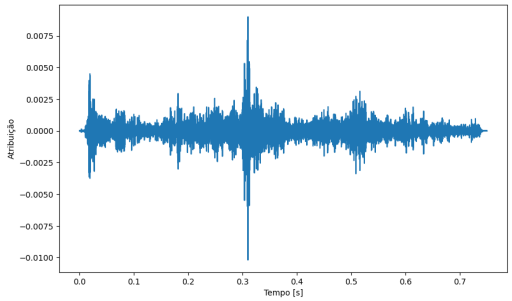



Fonte: O autor.

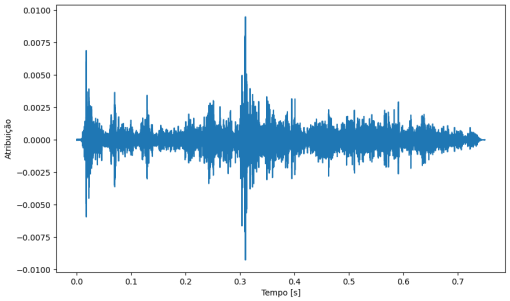
Para visualizar a importância das diferentes faixas de frequência e sons aleatórios, foram analisadas os impactos para diferentes períodos dos áudios e como as entradas se comportaram dadas as características ressaltadas. A Figura 19 mostra que em diferentes faixas de frequência foi retratado os picos de atribuição à partir de 0.3$s$, além do destaque para o início do áudio para as faixas 2kHz-4kHz e 4kHz-6kHz, para 0.1$s$ para a faixa 6kHz-8kHz e entre 0.48$s$ e 0.54$s$ aproximadamente para a faixa 8kHz-10kHz. Por outro lado, a Figura 20 mostra diferentes períodos dos áudios ressaltados, possivelmente porque o modelo esteve considerando também ruído para classificação.

Figura 19 – Atribuição média com o algoritmo *DeepLift* para diferentes períodos dos áudios

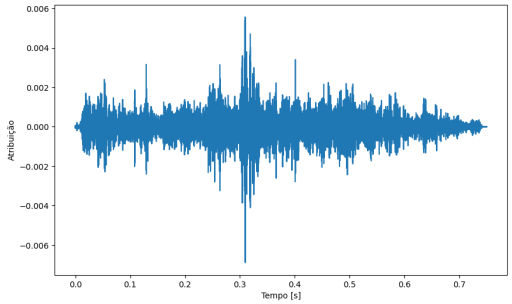
(a) Faixa 2kHz-4kHz



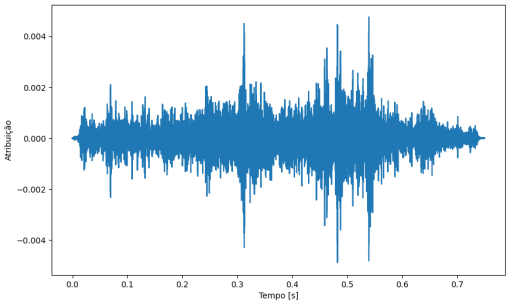

(b) Faixa 4kHz-6kHz




(c) Faixa 6kHz-8kHz



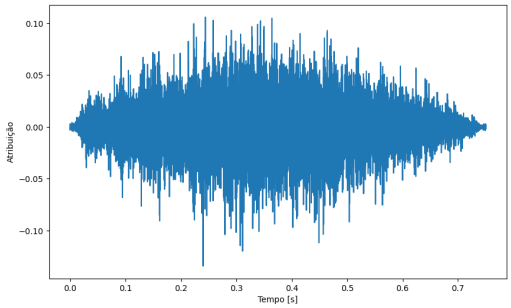

(d) Faixa 8kHz-10kHz



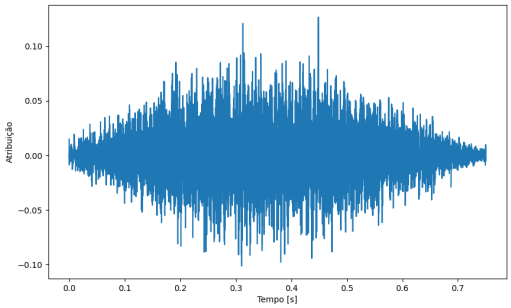

Fonte: O autor.

Figura 20 – Atribuição média com o algoritmo *DeepLift* com ruídos como referência

(a) Ruído gaussiano




(b) Ruído impulsivo



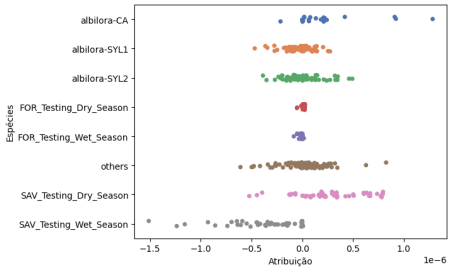

Fonte: O autor.

O principal impacto das referências escolhidas pode ser observado nas Figuras 21 e 22, a classe *others* contribui positivamente à partir da faixa de frequência 6kHz, enquanto a classe *SAV_Testing_Wet_Season* contribuiu negativamente para todas as faixas de frequências. Além disso, as sílabas do S. albilora diminuíram a atribuição à medida que faixas de frequência maiores foram consideradas, e a floresta variou entre contribuições quase nulas ou negativas, com exceção para faixa 8kHz-10kHz, em que mais entradas

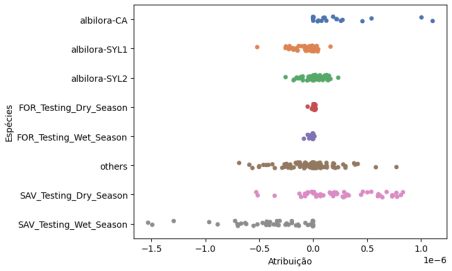
tiveram atribuição positiva. Em vista da magnitude da atribuição calculada, além do que foi observado na Figura 20, o modelo apresenta ter considerado parte das atividades acústicas de cada classe como ruído, em que para algumas entradas a contribuição foi positiva e para outras foram negativas. Vale ressaltar também a maior atribuição para as classes para o ruído impulsivo, o que pode ser possível devido a algumas atividades periódicas de aves ou ruídos mais intensos ao longo de cada áudio. Tais comportamento foram observados com a visualização ao longo dos blocos convolucionais escolhidos, no entanto, com maior quantidade de entradas com atribuição negativa para as classes das sílabas do S. albilora.

Figura 21 – Atribuição média com o algoritmo *DeepLift* para diferentes classes

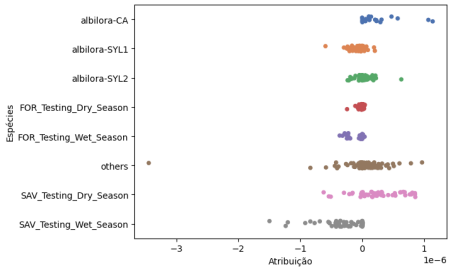
(a) Faixa 2kHz-4kHz



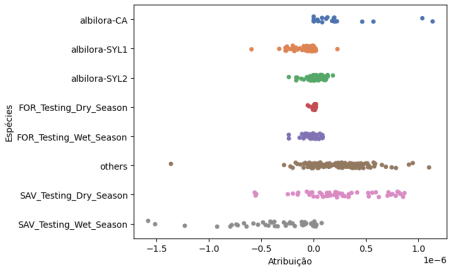

(b) Faixa 4kHz-6kHz




(c) Faixa 6kHz-8kHz



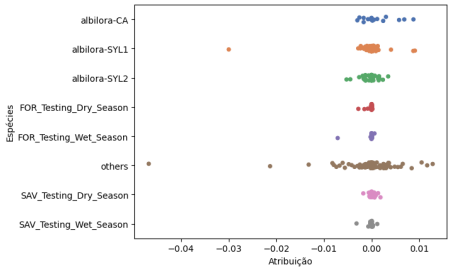

(d) Faixa 8kHz-10kHz



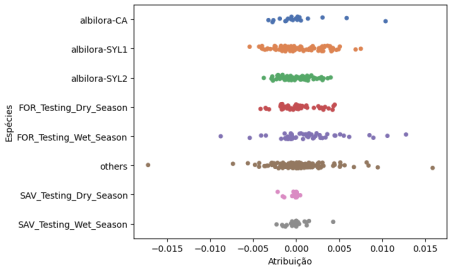

Fonte: O autor.

Figura 22 – Atribuição média com o algoritmo *DeepLift* para diferentes classes com ruídos como referência

(a) Ruído gaussiano




(b) Ruído impulsivo




Fonte: O autor.

Com os *TCAVs*, os ruídos também tiveram maior atribuição para a classificação do modelo, com principal destaque o 5º bloco. A fim de analisar melhor o impacto de cada faixa de frequência previamente definida, foram realizados testes *t* de *student* com *p* igual a 0.05 e hipótese nula $h_0$ de que a característica sob visualização não tem importância para classificar as entradas foram realizados considerando cada faixa de característica e os ruídos. Também foi realizado o teste entre faixas de frequência para determinar quais têm mais importância sobre os áudios.

A Tabela 2 mostra que o modelo captou diferentes características ao longo dos blocos convolucionais, em que todas as faixas analisadas tiveram impacto para classificar as entradas, por ao menos uma vez, em que a faixa 8kHz-10kHz teve cada vez menor influência e a faixa 4kHz-6kHz teve maior contribuição. No entanto, em comparação aos ruídos gaussianos e impulsivos, a faixa 8kHz-10kHz também teve cada vez mais influência sobre as entradas, junto com a faixa 4kHz-6kHz, possivelmente foram as faixas as quais tiveram maior atividade acústica, de modo que foi possível serem diferenciadas dos ruídos. Para as outras faixas, o comportamento permaneceu semelhante.

Tabela 2 – Teste *t* de *student* para frequências como conceitos aprendidos pelo modelo.

| Bloco | Faixa sob observação | *p* | Hipótese |
|---|---|---|---|
| 1º | 2kHz-4kHz | 0,038 | Rejeitada |
| 1º | 4kHz-6kHz | 0,429 | Aceita |
| 1º | 6kHz-8kHz | 0,027 | Rejeitada |
| 1º | 8kHz-10kHz | 0,001 | Rejeitada |
| 2º | 2kHz-4kHz | 0,085 | Aceita |
| 2º | 4kHz-6kHz | 0,139 | Aceita |
| 2º | 6kHz-8kHz | 0,528 | Aceita |
| 2º | 8kHz-10kHz | 0,015 | Rejeitada |
| 5º | 2kHz-4kHz | 0,01 | Rejeitada |
| 5º | 4kHz-6kHz | 0,026 | Rejeitada |
| 5º | 6kHz-8kHz | 0,058 | Aceita |
| 5º | 8kHz-10kHz | 0,483 | Aceita |

A fim de validar a interpretabilidade realizada, foram determinados novos experimentos à partir da atribuição. Somente serão listados os experimentos cujos resultados permitiram identificação da ave de estudo. No primeiro experimento, foram removidas todas as entradas da classe *others* cuja atribuição foi menor ou igual a 0 no experimento passado. Além disso, a classe *SAV_Testing_Wet_Season* foi desconsiderada o treinamento. Foi possível observar melhor capacidade de generalização do modelo, pois apesar da acurácia diminuir para 20,7%, a precisão aumentou para 8,7%, o *recall* para 15,7% e $F1$ para 7,5%. Essas métricas além de variarem, mostram como foi importante a remoção das entradas, visto que as classes das sílabas também foram detectadas. Ademais, como o modelo não convergiu, também foi realizado esse experimento com 200 iterações, o

que permitiu os sons da savana em período seco também serem reconhecidos, nesse caso a acurácia, precisão, *recall* e $F1-score$ foram 26,5%, 20%, 20,1% e 14,6%. Em vista da atribuição dada principalmente ao "meio"do áudio, também foi realizado um experimento como os anteriores com cada áudio carregado 0.5 segundos. Com 100 iterações os resultados foram 32,8%, 11,2%, 24% e 14,5% para a acurácia, precisão, *recall* e $F1$; e 18,9%, 3,5%, 14,3% e 5,5% para 200 iterações. A Tabela 3 mostra os resultados de todos os experimentos, na ordem que foi descrito anteriormente, para a identificação de cada uma das classes de S. albilora, com destaque em negrito para o melhor resultado das métricas precisão, *recall* e $F1-score$.

Diante dos resultados apresentados, foi possível observar melhoria no desempenho do modelo treinado. Desta forma, com a escolha de modelo para este trabalho e dos dados do Pantanal para a classificação de S. Albilora, remover entradas as quais não afetaram ou afetaram negativamente o modelo tornou capaz a identificação da ave *Synallaxis Albilora*, com destaque para a vocalização das sílabas, as quais detêm maior quantidade de dados em comparação ao canto.

Tabela 3 – Desempenho dos experimentos iniciais e propostos com interpretabilidade para identificação da ave *Synallaxis Albilora*

| Experimento | Precisão (%) | | | *Recall* (%) | | | *F1-Score* (%) | | |
|---|---|---|---|---|---|---|---|---|---|
| | CA | SYL1 | SYL2 | CA | SYL1 | SYL2 | CA | SYL1 | SYL2 |
| Inicial | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1º | 0 | 14 | 22 | 0 | 12 | 96 | 0 | 13 | 36 |
| 2º | 0 | 20 | 42 | 0 | 72 | 54 | 0 | 32 | 47 |
| 3º | 0 | **26** | **52** | 0 | **99** | 69 | 0 | **43** | **57** |
| 4º | 0 | 1 | 23 | 0 | 1 | **99** | 0 | 1 | 37 |

# CAPÍTULO 5

# CONCLUSÕES

Este trabalho teve como objetivo descobrir os impactos da interpretabilidade em RNP na bioacústica. Especificamente, este trabalho utilizou um modelo previamente treinado com dados acústicos e realizou *Transfer Learning* para dados de aves e de ambientes do Pantanal, partindo de métodos de interpretabilidade *post-hoc* para entendimento dos resultados.

Um modelo das PANNs (KONG *et al.*, 2020) foi selecionado para treinamento dos dados. O treinamento inicial demonstrou *performance* extremamente baixa, com somente a detecção de uma classe. À partir das interpretações realizadas com os métodos previamente definidos, foram identificados pontos os quais causaram problemas para que o modelo pudesse distinguir entre diferentes classes. Com a aplicação de ajustes baseado no entendimento do aprendizado do modelo, o modelo conseguiu distinguir diferentes classes e obteve melhor desempenho na identificação da espécie alvo. Assim, a interpretabilidade é viável e importante para modelos de bioacústica, pois o conhecimento do domínio permite melhorias nos resultados e entendimento do aprendizado do modelo.

Para trabalhos futuros, outros métodos de interpretabilidade podem ser abordados, comparando as características aprendidas de acordo com cada método. Além disso, outros modelos, seja com *Transfer Learning* ou não, podem ser testados, a fim de ratificar os impactos de interpretabilidade *post-hoc*. Essas metodologias também podem ser feitas

com outras bases de dados cujo domínio é a bioacústica, a fim de entender cada aspecto específico da área e verificar limitações dos métodos. Por último, também podem ser abordadas arquiteturas de Redes Neurais Interpretáveis, inicialmente apresentadas neste trabalho.

# REFERÊNCIAS

ADADI, Amina; BERRADA, Mohammed. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). **IEEE Access**, v. 6, p. 52138–52160, 2018. DOI: doi.org/10.1109/ACCESS.2018.2870052.

BADSHAH, Abdul Malik *et al.* Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In: p. 1–5. ISBN 978-1-5090-5140-3. DOI: 10.1109/PlatCon.2017.7883728.

BALDI, Pierre. Autoencoders, Unsupervised Learning, and Deep Architectures. In: GUYON, Isabelle *et al.* (Ed.). **Proceedings of ICML Workshop on Unsupervised and Transfer Learning**. Bellevue, Washington, USA: PMLR, jul. 2012. v. 27. (Proceedings of Machine Learning Research), p. 37–49. Disponível em: <https://proceedings.mlr.press/v27/baldi12a.html>.

CAKIR, Emre *et al.* Convolutional recurrent neural networks for bird audio detection. In: p. 1744–1748. ISBN 978-0-9928626-7-1. DOI: 10.23919/EUSIPCO.2017.8081508.

DIMENT, Aleksandr; VIRTANEN, Tuomas. Transfer learning of weakly labelled audio. In: p. 6–10. ISBN 978-1-5386-1632-1. DOI: 10.1109/WASPAA.2017.8169984.

DOSHI-VELEZ, Finale; KIM, Been. Towards a rigorous science of interpretable machine learning. **arXiv preprint arXiv:1702.08608**, 2017.

DUFOURQ, Emmanuel *et al.* Passive acoustic monitoring of animal populations with transfer learning. **Ecological Informatics**, v. 70, p. 101688, set. 2022. ISSN 15749541. DOI: 10.1016/j.ecoinf.2022.101688.

FORTELNY, Nikolaus; BOCK, Christoph. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. **Genome Biology**, v. 21, p. 190, 1 dez. 2020. ISSN 1474-760X. DOI: 10.1186/s13059-020-02100-5.

HARCOURT, Rob *et al.* Animal-Borne Telemetry: An Integral Component of the Ocean Observing Toolkit. **Frontiers in Marine Science**, v. 6, jun. 2019. ISSN 2296-7745. DOI: 10.3389/fmars.2019.00326.

HIDAYAT, Alam Ahmad; CENGGORO, Tjeng Wawan; PARDAMEAN, Bens. Convolutional Neural Networks for Scops Owl Sound Classification. **Procedia Computer Science**, v. 179, p. 81–87, 2021. ISSN 18770509. DOI: 10.1016/j.procs.2020.12.010.

JADHAV, Yogesh; PATIL, Vishal; PARASAR, Deepa. Machine Learning Approach to Classify Birds on the Basis of Their Sound. In: 2020 International Conference on Inventive Computation Technologies (ICICT). IEEE, 2020. P. 69–73. DOI: 10.1109/ICICT48043.2020.9112506.

JI, Xunsheng; JIANG, Kun; XIE, Jie. LBP-based bird sound classification using improved feature selection algorithm. **International Journal of Speech Technology**, v. 24, p. 1033–1045, 4 dez. 2021. ISSN 1381-2416. DOI: 10.1007/s10772-021-09866-4.

KAHL, Stefan *et al.* **Large-Scale Bird Sound Classification using Convolutional Neural Networks.** 2017. v. 1866.

KHAN, Abdullah Ayub; LAGHARI, Asif Ali; AWAN, Shafique Ahmed. Machine Learning in Computer Vision: A Review. **EAI Endorsed Transactions on Scalable Information Systems**, EAI, v. 8, n. 32, abr. 2021. DOI: 10.4108/eai.21-4-2021.169418.

KIM, Been *et al.* Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In: DY, Jennifer; KRAUSE, Andreas (Ed.). **Proceedings of the 35th International Conference on Machine Learning**. PMLR, jul. 2018. v. 80. (Proceedings of Machine Learning Research), p. 2668–2677.

KIM, Phil. Convolutional Neural Network. In: MATLAB Deep Learning: With Machine Learning, Neural Networks and Artificial Intelligence. Berkeley, CA: Apress, 2017. P. 121–147. ISBN 978-1-4842-2845-6. DOI: 10.1007/978-1-4842-2845-6_6.

KOKHLIKYAN, Narine *et al.* **Captum: A unified and generic model interpretability library for PyTorch**. 2020. arXiv: 2009.07896 [cs.LG].

KONG, Qiuqiang *et al.* PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 28, p. 2880–2894, 2020. DOI: 10.1109/TASLP.2020.3030497.

KOO, Peter K.; PLOENZKE, Matt. Deep learning for inferring transcription factor binding sites. **Current Opinion in Systems Biology**, v. 19, p. 16–23, fev. 2020. ISSN 2452-3100. DOI: 10.1016/j.coisb.2020.04.001.

LI, Oscar *et al.* Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 32, n. 1, abr. 2018. DOI: 10.1609/aaai.v32i1.11771.

LINARDATOS, Pantelis; PAPASTEFANOPOULOS, Vasilis; KOTSIANTIS, Sotiris. Explainable AI: A Review of Machine Learning Interpretability Methods. **Entropy**, v. 23, n. 1, 2021. ISSN 1099-4300. DOI: 10.3390/e23010018.

MATLEY, Jordan K. *et al.* Global trends in aquatic animal tracking with acoustic telemetry. **Trends in Ecology & Evolution**, v. 37, n. 1, p. 79–94, 2022. ISSN 0169-5347. DOI: 10.1016/j.tree.2021.09.001.

MCCULLOCH, Warren S; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, Springer, v. 5, p. 115–133, 1943.

MILLER, Tim. Explanation in artificial intelligence: Insights from the social sciences. **Artificial Intelligence**, Elsevier, v. 267, p. 1–38, 2019. ISSN 0004-3702. DOI: 10.1016/j.artint.2018.07.007.

MURPHY, John. An overview of convolutional neural network architectures for deep learning. **Microway Inc**, p. 1–22, 2016.

NOUMIDA, A; RAJAN, Rajeev. Multi-label bird species classification from audio recordings using attention framework. **Applied Acoustics**, v. 197, p. 108901, 2022. ISSN 0003-682X. DOI: 10.1016/j.apacoust.2022.108901.

PASZKE, Adam *et al.* PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: WALLACH, H. *et al.* (Ed.). **Advances in Neural Information Processing Systems 32**. Curran Associates, Inc., 2019. P. 8024–8035. Disponível em: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.

PÉREZ-GRANADOS, Cristian; SCHUCHMANN, Karl-L. Automated signal recognition as a useful tool for monitoring little-studied species: The case of the Band-tailed Nighthawk. **Ecological Informatics**, v. 72, p. 101861, dez. 2022. ISSN 15749541. DOI: 10.1016/j.ecoinf.2022.101861.

PÉREZ-GRANADOS, Cristian; SCHUCHMANN, Karl-L. The sound of the illegal: Applying bioacoustics for long-term monitoring of illegal cattle in protected areas. **Ecological Informatics**, v. 74, p. 101981, mai. 2023. ISSN 1574-9541. DOI: 10.1016/j.ecoinf.2023.101981.

PÉREZ-GRANADOS; SCHUCHMANN. PÉREZ-GRANADOS AND SCHUCHMANN 2020 BIOACOUSTICS - Computational Bioacoustics Research Unit, 2020. Disponível em: <https://cobra.ic.ufmt.br/?page_id=642>. Acesso em: 21 nov. 2022.

PINHEIRO SARAVY, Fábio; SCHUCHMANN, Karl-L.; MARQUES, Marinez I. Diversity of Insect Flower Visitors of Xylopia aromatica (Magnoliales, Annonaceae) in a Brazilian Savanna. **Diversity**, v. 13, n. 12, 2021. ISSN 1424-2818. DOI: 10.3390/d13120661.

POTAPOV, Anton M *et al.* Global monitoring of soil animal communities using a common methodology. **bioRxiv**, Cold Spring Harbor Laboratory, p. 2022–01, 2022.

RAI, Pallavi *et al.* An automatic classification of bird species using audio feature extraction and support vector machines. In: 2016 International Conference on Inventive Computation Technologies (ICICT). 2016. v. 1, p. 1–5. DOI: 10.1109/INVENTIVE.2016.7823241.

REN, Zhao; NGUYEN, Thanh Tam; NEJDL, Wolfgang. Prototype Learning for Interpretable Respiratory Sound Analysis. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2022. P. 9087–9091. DOI: 10.1109/ICASSP43922.2022.9747014.

ROJAS, Raúl. The Backpropagation Algorithm. In: NEURAL Networks: A Systematic Introduction. Berlin, Heidelberg: Springer Berlin Heidelberg, 1996. P. 149–182. ISBN 978-3-642-61068-4. DOI: 10.1007/978-3-642-61068-4_7.

ROSENBLATT, Frank. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, American Psychological Association, v. 65, n. 6, p. 386, 1958.

RUBIO, Tatiana Colombo; PINHO, João Batista de. Biologia reprodutiva de Synallaxis albilora (aves: Furnariidae) no Pantanal de Poconé, Mato Grosso. **Papéis Avulsos de Zoologia**, Museu de Zoologia da Universidade de São Paulo, v. 48, n. 17, p. 181–197, 2008. ISSN 0031-1049. DOI: 10.1590/S0031-10492008001700001.

RUMELHART, David E; HINTON, Geoffrey E; WILLIAMS, Ronald J. **Learning internal representations by error propagation**. 1985.

SCHUCHMANN, Karl-L. *et al.* Reproduction and agonistic behavior of black skimmers (Rynchops niger) in a mixed-species colony in the Brazilian Pantanal. **Studies on Neotropical Fauna and Environment**, Taylor & Francis, v. 53, n. 3, p. 219–232, 2018. DOI: 10.1080/01650521.2018.1479951.

SENIČ, Martin *et al.* Activity Patterns, Sex Ratio, and Social Organization of the Bare-Faced Curassow (Crax fasciolata) in the Northern Pantanal, Brazil. **Birds**, v. 4, n. 1, p. 117–137, 2023. ISSN 2673-6004. DOI: 10.3390/birds4010010.

SHRIKUMAR, Avanti; GREENSIDE, Peyton; KUNDAJE, Anshul. Learning Important Features Through Propagating Activation Differences. In: PRECUP, Doina; TEH, Yee Whye (Ed.). **Proceedings of the 34th International Conference on Machine Learning**. PMLR, ago. 2017. v. 70. (Proceedings of Machine Learning Research), p. 3145–3153.

SNELL, Jake; SWERSKY, Kevin; ZEMEL, Richard. Prototypical Networks for Few-shot Learning. In: GUYON, I. *et al.* (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf>.

SPRINGENBERG, Jost Tobias *et al.* Striving for Simplicity: The All Convolutional Net. In: ARXIV preprint. Dez. 2014.

STEWART, James. **Cálculo - Volume 2**. 5. ed. São Paulo: Thomson Learning, 2007.

TABUR, Mehmet Ali; AYVAZ, Yusuf. Ecological importance of birds. **Second International Symposium on Sustainable Development Conference**, p. 560–565, jun. 2010.

UNWIN, Mike. **The Atlas of Birds**: Diversity, Behavior, and Conservation. Princeton: Princeton University Press, 2011. ISBN 9781400838257. DOI: doi:10.1515/9781400838257.

VENTURA, Thiago M. *et al.* Audio parameterization with robust frame selection for improved bird identification. In: v. 42, p. 8463–8471. DOI: 10.1016/j.eswa.2015.07.002.

YUAN, Bo *et al.* CellBox: Interpretable Machine Learning for Perturbation Biology with Application to the Design of Cancer Combination Therapy. **Cell Systems**, v. 12, 128–140.e4, 2 fev. 2021. ISSN 24054712. DOI: doi.org/10.1016/j.cels.2020.11.013.

ZEILER, Matthew D.; FERGUS, Rob. Visualizing and Understanding Convolutional Networks. In: FLEET, David *et al.* (Ed.). **Computer Vision – ECCV 2014**. Cham: Springer International Publishing, 2014. P. 818–833. DOI: 10.1007/978-3-319-10590-1_53.

ZINEMANAS, Pablo *et al.* An Interpretable Deep Learning Model for Automatic Sound Classification. **Electronics**, v. 10, n. 7, 2021. ISSN 2079-9292. DOI: 10.3390/electronics10070850.