UFC

UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS JARDINS DE ANITA - ITAPAJÉ

PROCESSAMENTO DE SINAIS APLICADA A REDE NEURAL CONVOLUCIONAL E RECORRENTE PARA RECONHECIMENTO E CLASSIFICAÇÃO DE AVES NATIVAS DE ITAPAJÉ

Autores:
Pedro Vinicius Félix Rosa Viana
Francisca Marília Oliveira
José Mario Oliveira Patrício

Disciplina: APRENDIZADO PROFUNDO

Itapajé, Janeiro, 2026

Conteúdo

Resumo do Projeto

Este trabalho apresenta o desenvolvimento de um sistema de classificação automática de espécies de aves a partir de seus cantos, utilizando técnicas de processamento digital de sinais e aprendizado profundo. Inicialmente, os áudios são pré-processados e transformados em representações no domínio do tempo-frequência, como espectrogramas, a fim de extrair características relevantes dos sinais sonoros. Em seguida, essas representações são utilizadas para treinar um modelo híbrido CNN-LSTM, que combina redes neurais convolucionais para extração espacial de padrões acústicos e redes recorrentes do tipo LSTM para modelar dependências temporais. O desempenho do modelo é avaliado por meio de métricas de classificação, demonstrando o potencial da abordagem para auxiliar na identificação automática de espécies, com aplicações em monitoramento ambiental e conservação da biodiversidade.

Palavras-chave: Processamento de sinais, Aprendizado profundo, Classificação de áudio, Canto de aves, Redes neurais convolucionais (CNN), LSTM, Bioacústica, Reconhecimento de padrões, Espectrogramas.

1 Introdução

A identificação automática de espécies de aves a partir de seus cantos tem se tornado uma área de grande interesse na bioacústica e no monitoramento ambiental, especialmente devido ao avanço das técnicas de processamento digital de sinais e aprendizado profundo. O canto das aves contém informações ricas e discriminativas que permitem diferenciar espécies, mesmo em ambientes naturais complexos e com altos níveis de ruído.

Tradicionalmente, a classificação de espécies baseada em áudio dependia da extração manual de características e da análise especializada de pesquisadores. No entanto, esse processo é demorado, sujeito a erros e de difícil escalabilidade. Nesse contexto, modelos de redes neurais profundas surgem como uma alternativa eficiente, capazes de aprender automaticamente representações relevantes diretamente dos dados.

Entre essas abordagens, arquiteturas híbridas que combinam Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes do tipo Long Short-Term Memory (LSTM) têm se destacado. As CNNs são eficazes na extração de padrões locais em representações tempo-frequência, como espectrogramas, enquanto as LSTMs capturam dependências temporais presentes nos sinais acústicos.

Dessa forma, este trabalho tem como objetivo aplicar técnicas de processamento de sinais em gravações de canto de aves para treinar um modelo CNN-LSTM voltado à classificação automática de espécies. A proposta busca contribuir para o desenvolvimento de ferramentas computacionais que auxiliem o monitoramento da biodiversidade, a pesquisa ecológica e a conservação ambiental.

2 Proposta do Trabalho e Objetivos

Este trabalho propõe o desenvolvimento de um sistema computacional capaz de classificar automaticamente espécies de aves a partir de seus cantos, utilizando técnicas de processamento digital de sinais e aprendizado profundo. A abordagem adotada consiste na conversão dos sinais de áudio em representações tempo-frequência e no treinamento de um modelo híbrido baseado em redes neurais convolucionais (CNN) e redes recorrentes do tipo LSTM.

A proposta visa explorar a capacidade dessas arquiteturas em extrair características acústicas relevantes e modelar a dinâmica temporal dos sinais sonoros, buscando obter um desempenho satisfatório mesmo diante de variações naturais dos cantos e da presença de ruídos ambientais.

2.1 Objetivo Geral

Desenvolver e avaliar um modelo CNN-LSTM para a classificação automática de espécies de aves a partir de áudios de seus cantos, alcançando uma acurácia mínima de 75% no conjunto de testes.

2.2 Objetivos Específicos

- Aplicar técnicas de pré-processamento e extração de características em sinais de áudio de canto de aves;
- Gerar representações adequadas, como espectrogramas ou matrizes de features (.npy), para entrada no modelo;
- Projetar e implementar uma arquitetura CNN-LSTM adequada ao problema;
- Treinar e validar o modelo utilizando uma base de dados rotulada;
- Avaliar o desempenho por meio de métricas como acurácia, precisão, revocação e matriz de confusão;
- Analisar os resultados obtidos e verificar se o objetivo mínimo de 75% de acurácia foi atingido.

3 Fundamentação Teórica

3.1 Processamento Digital de Sinais de Áudio

O processamento digital de sinais (PDS) consiste na manipulação computacional de sinais amostrados com o objetivo de extrair informações relevantes ou melhorar sua representação. No contexto de sinais de áudio, essa etapa é fundamental para transformar ondas sonoras contínuas em formas adequadas para análise por modelos de aprendizado de máquina.

Uma das representações mais utilizadas é o espectrograma, que descreve a variação do conteúdo espectral de um sinal ao longo do tempo. Ele é obtido por meio da Transformada de Fourier de Curto Prazo (STFT), permitindo visualizar simultaneamente informações temporais e frequenciais. Essa representação é particularmente eficaz para a análise de sons biológicos, como cantos de aves, pois evidencia padrões acústicos característicos de cada espécie.

Além disso, os espectrogramas podem ser convertidos em matrizes de características (features), que são armazenadas em formatos numéricos, como arquivos .npy, possibilitando leitura eficiente e processamento direto pelos modelos de redes neurais. Essa conversão reduz a complexidade do sinal original e facilita o treinamento de modelos profundos.

3.1.1 Espectrograma

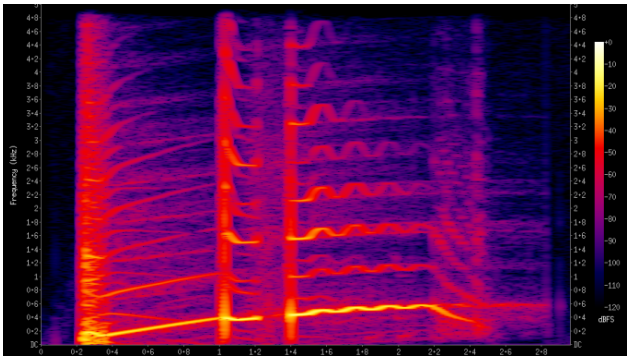

Figura 1: Exemplo: Espectrograma

Um espectrograma é uma representação do sinal de áudio no domínio tempo-frequência, isto é, mostra como a distribuição de energia (ou magnitude) do sinal varia ao longo do tempo para diferentes frequências. Em geral, ele é visualizado como uma "imagem" em que:

- o eixo x representa o tempo,
- o eixo y representa a frequência,
- a intensidade (cor) representa a magnitude ou a potência do sinal naquela frequência e instante.

Essa representação é muito utilizada em tarefas de classificação de áudio, pois evidencia padrões característicos (harmônicos, modulações, ataques) que são difíceis de observar diretamente na forma de onda.

3.1.2 STFT (Short-Time Fourier Transform)

A Transformada de Fourier tradicional fornece apenas informações de frequência "globais" do sinal inteiro, perdendo a localização temporal. Para resolver isso, utiliza-se a Transformada de

Fourier de Curto Prazo (STFT), que calcula a Transformada de Fourier em janelas curtas do sinal, permitindo analisar como o espectro muda com o tempo.

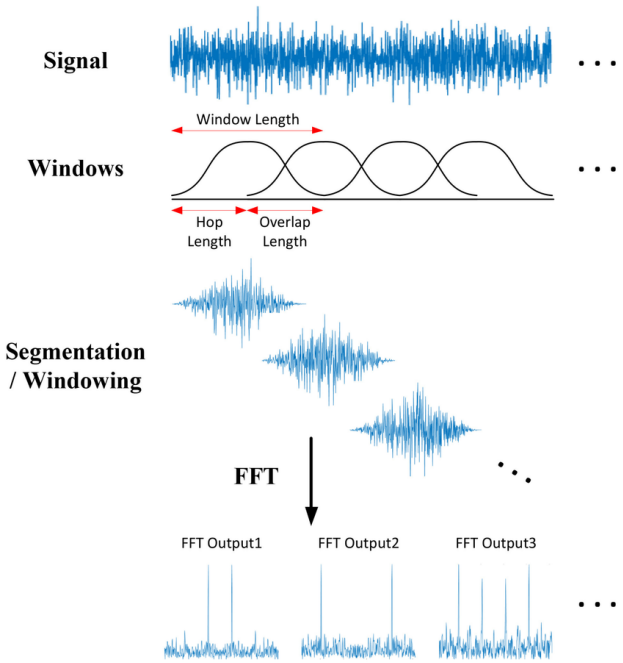

Figura 2: Processo da STFT

A STFT de um sinal discreto $x[n]$ pode ser definida como:

$$X(m, w) = \sum_{n=-\infty}^{\infty} x[n]w[n-m]e^{-jwm}$$

onde:

- $x[n]$ é o sinal no tempo discreto,
- $w[\cdot]$ é uma função janela (ex.: Hamming, Hann),
- m indica o deslocamento temporal da janela,
- w é a frequência angular,
- j é a unidade imaginária.

Na prática computacional, calcula-se a STFT para valores discretos de frequência via FFT. O espectrograma é obtido a partir do módulo (ou módulo ao quadrado) da STFT:

- Espectrograma de magnitude: $|X(m, w)|$
- Espectrograma de potência: $|X(m, w)|^2$

Muitas implementações ainda aplicam escala logarítmica (ex.: dB) para realçar detalhes.

3.1.3 Matriz de características (features)

Uma matriz de características (features) é uma representação numérica estruturada que reúne informações relevantes extraídas do sinal, servindo como entrada para o modelo de aprendizado de máquina.

No caso deste trabalho, a matriz de features é construída a partir do áudio processado (por exemplo, a partir de um espectrograma ou Mel-espectrograma) e pode ser interpretada como:

- uma matriz $F \in \mathbb{R}^{T \times K}$ em que;
- T é o número de quadros temporais (janelas ao longo do tempo),
- K é o número de componentes espectrais (bins de frequência ou bandas Mel).

Assim, cada linha representa um instante (janela), e cada coluna representa uma frequência (ou banda), contendo valores de magnitude/potência (geralmente em escala log).

Essas matrizes podem ser armazenadas em arquivos .npy (NumPy array), o que facilita o carregamento eficiente, padroniza a entrada do modelo e reduz o custo de recalcular a extração de características durante o treinamento.

3.2 Redes Neurais Convolucionais (CNN)

As Redes Neurais Convolucionais (Convolutional Neural Networks – CNN) são amplamente empregadas em tarefas de reconhecimento de padrões visuais, como classificação de imagens e detecção de objetos. Sua principal característica é o uso de camadas convolucionais capazes de aprender filtros automaticamente, extraindo características locais relevantes, como bordas, texturas e formas.

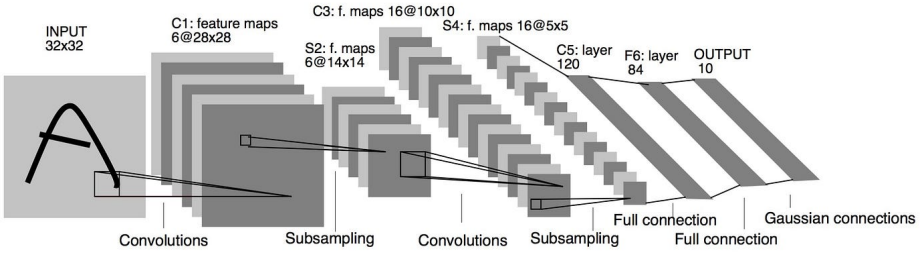

Figura 3: Funcionamento de uma CNN

Quando aplicadas a espectrogramas, as CNNs passam a identificar padrões acústicos representados visualmente, tais como harmônicos, transientes e estruturas temporais locais. Dessa forma, o modelo aprende automaticamente descritores relevantes dos cantos das aves, eliminando a necessidade de extração manual de características.

3.2.1 Operação de Convolução

A principal operação realizada por uma CNN é a convolução, na qual um pequeno filtro (ou kernel) é deslizado sobre a entrada para extrair padrões locais.

Matematicamente, para uma entrada bidimensional X e um filtro K, a convolução discreta pode ser expressa como:

$$Y(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(i+m, j+n) \cdot K(m,n)$$

onde:

- X é a matriz de entrada (por exemplo, um espectrograma),
- K é o kernel de tamanho $M \times N$,
- $Y(i, j)$ é o valor do mapa de características (feature map) na posição (i, j).

Durante o treinamento, os valores do kernel são ajustados automaticamente para maximizar a capacidade do modelo em distinguir padrões relevantes.

3.2.2 Mapas de Características (Feature Maps)

Cada filtro aplicado gera um mapa de características, que representa a ativação daquele padrão específico ao longo da entrada. Em estágios iniciais, os filtros tendem a aprender padrões simples, como transições abruptas de energia ou harmônicos, enquanto camadas mais profundas capturam estruturas acústicas mais complexas.

Quando aplicadas a espectrogramas, essas ativações correspondem a padrões temporais e espectrais típicos dos cantos das aves.

3.2.3 Camadas de Pooling

As camadas de pooling são utilizadas para reduzir a dimensionalidade espacial dos mapas de características, mantendo as informações mais relevantes. A forma mais comum é o max pooling, definido como:

$$Y(i, j) = \max_{m,n \in \Omega} X(i + m, j + n)$$

onde Ω representa a região local considerada.

Essa operação reduz o custo computacional, aumenta a invariância a pequenas variações e contribui para diminuir o risco de sobreajuste (overfitting).

3.2.4 Funções de Ativação

Após a convolução, aplica-se uma função de ativação não linear, sendo a mais comum a ReLU (Rectified Linear Unit):

$$f(x) = max(0, x)$$

Essa função permite que a rede modele relações não lineares complexas presentes nos dados acústicos.

3.2.5 Aplicação de CNN em Espectrogramas

Ao tratar o espectrograma como uma imagem, a CNN passa a identificar:

- padrões harmônicos,
- modulações de frequência,
- ataques sonoros,
- estruturas temporais locais.

Esses padrões são fundamentais para distinguir diferentes espécies de aves, pois cada uma apresenta assinaturas acústicas próprias.

Além disso, a CNN elimina a necessidade de projetar manualmente descritores acústicos, aprendendo automaticamente as características mais discriminativas diretamente dos dados.

3.3 Redes Neurais Recorrentes e LSTM

As Redes Neurais Recorrentes (RNN) são projetadas para lidar com dados sequenciais, nos quais a ordem temporal possui grande importância. No entanto, RNNs tradicionais apresentam limitações relacionadas ao desaparecimento ou explosão do gradiente, dificultando o aprendizado de dependências de longo prazo.

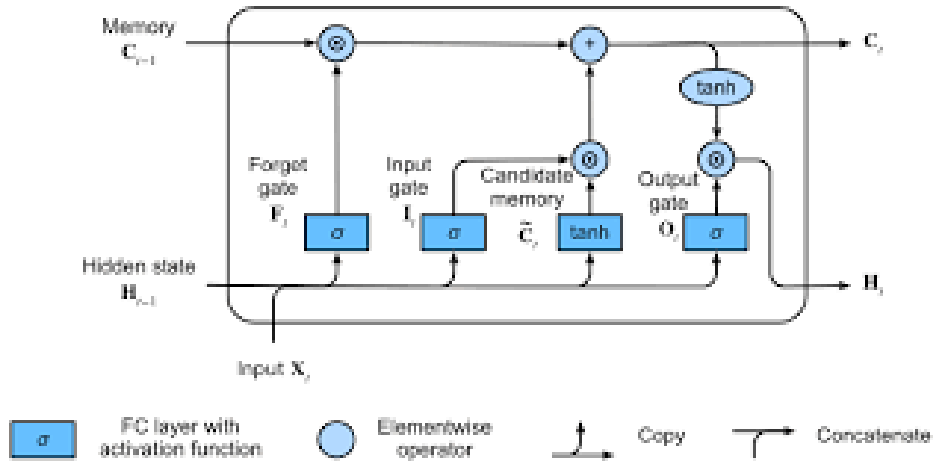

Figura 4: Funcionamento de uma LSTM

Para contornar esse problema, foram propostas as redes Long Short-Term Memory (LSTM), que utilizam mecanismos de portas (gates) para controlar o fluxo de informações ao longo do tempo. Isso permite que o modelo memorize padrões relevantes por períodos maiores, sendo especialmente útil para sinais de áudio, nos quais a informação discriminativa pode estar distribuída ao longo de toda a gravação.

3.3.1 Estrutura de uma célula LSTM

Uma célula LSTM é composta por um conjunto de mecanismos denominados portas (gates), responsáveis por controlar o fluxo de informações:

- Porta de esquecimento (forget gate)
- Porta de entrada (input gate)
- Porta de saída (output gate)
- Estado da célula (cell state)

Esses componentes permitem que a rede decida quais informações devem ser armazenadas, atualizadas ou descartadas ao longo do tempo.

3.3.2 Formulação matemática

Dado um vetor de entrada x_t, o estado oculto anterior h_{t-1} e o estado da célula anterior C_{t-1}, as operações da célula LSTM são definidas como:

Porta de Esquecimento:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

Porta de Entrada:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_t)$$

$$\tilde{C} = tanh(W_C \cdot [h_{t-1}, x_t] + b_C$$

Atualização do estado da célula:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

Porta de saída:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \odot tanh(C_t)$$

onde:

- $\sigma(\cdot)$ é a função sigmoid.
- $tanh(\cdot)$ é a tangente hiperbólica,
- Web são pesos e vieses aprendidos,
- $\odot$ representa multiplicação elemento a elemento.

3.3.3 Aplicação de LSTM em sinais de áudio

Em problemas de classificação de áudio, como o reconhecimento de cantos de aves, a informação relevante não está concentrada em um único instante, mas distribuída ao longo do tempo.

As redes LSTM são capazes de:

- modelar variações temporais dos padrões espectrais,
- capturar sequências rítmicas e melódicas,
- integrar informações acústicas ao longo de toda a gravação.

Isso torna as LSTMs especialmente adequadas para complementar CNNs, que extraem características locais, mas não modelam explicitamente dependências temporais.

3.4 Arquitetura Híbrida CNN-LSTM

A arquitetura híbrida CNN-LSTM combina as vantagens das duas abordagens: a capacidade das CNNs de extrair características espaciais e a habilidade das LSTMs de modelar dependências temporais.

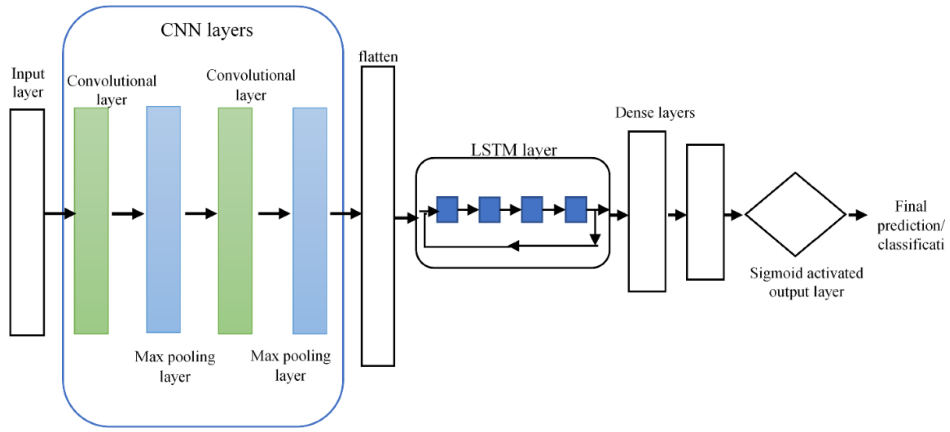

Figura 5: Arquitetura de um modelo híbrido (CNN-LSTM)

Nesse modelo, as camadas convolucionais atuam inicialmente sobre os espectrogramas, aprendendo representações acústicas discriminativas. Em seguida, essas representações são organizadas como sequências e fornecidas às camadas LSTM, que capturam a evolução temporal dos padrões sonoros.

Essa combinação é particularmente adequada para tarefas de classificação de áudio, pois considera simultaneamente a estrutura espectral e a dinâmica temporal dos sinais, características fundamentais nos cantos de aves.

3.4.1 Motivação

Embora as CNNs sejam altamente eficazes na identificação de padrões locais em espectrogramas, elas não modelam explicitamente a relação temporal entre diferentes segmentos do sinal. Por outro lado, as LSTMs capturam dependências ao longo do tempo, mas não são ideais para lidar diretamente com estruturas espaciais bidimensionais.

A integração dessas arquiteturas permite:

- extração robusta de padrões acústicos locais (CNN),
- modelagem da sequência temporal desses padrões (LSTM),
- maior capacidade discriminativa para sinais complexos, como cantos de aves.

3.4.2 Estrutura geral do modelo

Em uma arquitetura típica CNN-LSTM, o fluxo de processamento ocorre da seguinte forma:

1. A matriz .npy é fornecido como entrada para a CNN;
2. As camadas convolucionais extraem mapas de características de alta relevância;
3. Esses mapas são reorganizados em uma sequência temporal de vetores de características;
4. A sequência é processada por uma ou mais camadas LSTM;
5. A saída da LSTM é conectada a camadas densas para realizar a classificação final.

Matematicamente, esse processo pode ser representado como:

$$X = CNN(X)$$

$$H = LSTM(Z)$$

$$\hat{y} = Softmax(WH + b)$$

onde:

- X é a matriz .npy de entrada,
- Z são as características extraídas pela CNN,
- H representa a saída temproal da LSTM,
- $\hat{y}$ é o vetor de probabilidade das classes.

3.4.3 Vantagens da abordagem híbrida

O uso de CNN-LSTM apresenta diversas vantagens:

- captura simultânea de padrões espectrais e temporais;
- maior robustez a variações no canto das aves;
- melhor generalização em ambientes ruidosos;
- desempenho superior em comparação com modelos isolados em diversas aplicações de classificação de áudio.

Essas características tornam a arquitetura híbrida uma escolha adequada para sistemas automáticos de reconhecimento acústico em contextos reais.

3.5 Aumento de Dados (Data Augmentation) para Áudio

O aumento de dados (data augmentation) é uma técnica utilizada para expandir artificialmente o conjunto de treinamento, gerando novas amostras a partir das existentes. Essa estratégia é especialmente importante em problemas onde a quantidade de dados rotulados é limitada.

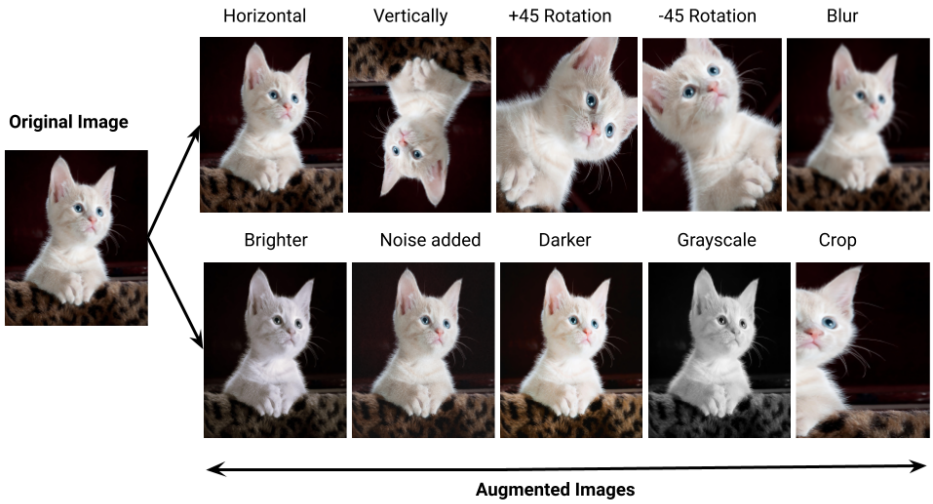

Figura 6: Funcionamento do Data Augmentation

No contexto de áudio, técnicas comuns de data augmentation incluem adição de ruído, variação de velocidade (time stretching), alteração de tom (pitch shifting) e deslocamento temporal. Essas

transformações preservam o conteúdo semântico do sinal, mas introduzem variabilidade suficiente para tornar o modelo mais robusto e menos suscetível a overfitting.

A aplicação dessas técnicas contribui significativamente para melhorar a capacidade de generalização do modelo CNN-LSTM, resultando em melhor desempenho na classificação de espécies em cenários reais.

4 Metodologia

Esta seção descreve as etapas adotadas para o desenvolvimento do sistema de classificação automática de espécies de aves, abrangendo a aquisição dos dados, o pré-processamento dos sinais de áudio, a extração de características, a construção do modelo CNN-LSTM, o processo de treinamento e os critérios de avaliação.

4.1 Base de Dados

A construção da base de dados foi realizada a partir de uma pesquisa em plataformas especializadas em bioacústica, como os sites WikiAves e eBird, amplamente utilizados para o compartilhamento e catalogação de registros sonoros de aves. Além disso, foi incorporado material disponibilizado pela Universidade Federal do Ceará (UFC), proveniente do Projeto Fênix, voltado à identificação automática de aves por meio do canto.

A partir da integração dessas fontes, foi possível reunir gravações correspondentes a 140 espécies de aves, incluindo espécies nativas e espécies migratórias que ocorrem na região de estudo. Essa diversidade contribui para a robustez do modelo proposto, permitindo avaliar seu desempenho em um cenário realista, caracterizado por grande variabilidade acústica interespécies.

Tabela 1: Espécies consolidadas obtidas a partir do Projeto Fênix (UFC), WikiAves e eBird

Família	**Espécies**
TINAMIDAE	*Crypturellus tataupa*, *Crypturellus parvirostris*
ANATIDAE	*Dendrocygna viduata*
CRACIDAE	*Penelope jacucaca*
COLUMBIDAE	*Leptotila verreauxi*, *Claravis pretiosa*, *Columbina talpacoti*, *Columbina picui*, *Columbina minuta*, *Columbina squammata*
CUCULIDAE	*Crotophaga ani*, *Piaya cayana*, *Coccyzus melacoryphus*, *Coccyzus euleri*, *Guira guira*
APODIDAE	*Streptoprocne biscutata*, *Tachornis squamata*
TROCHILIDAE	*Phaethornis ruber*, *Polytmus guainumbi*, *Chrysolampis mosquitus*, *Heliomaster squamosus*, *Calliphlox amethystina*, *Chlorostilbon lucidus*, *Thalurania furcata*, *Chrysuronia versicolor*, *Chionomesa fimbriata*, *Anopetia gounellei*
RALLIDAE	*Neocrex erythrops*
RECURVIROSTRIDAE	*Himantopus mexicanus*
JACANIDAE	*Jacana jacana*
CATHARTIDAE	*Sarcoramphus papa*, *Coragyps atratus*, *Cathartes aura*, *Cathartes burrovianus*

Continua na próxima página

Família	**Espécies**
ARDEIDAE	*Egretta thula*, *Tigrisoma lineatum*, *Ardea alba*
ACCIPITRIDAE	*Chondrohierax uncinatus*, *Rostrhamus sociabilis*, *Accipiter bicolor*, *Geranospiza caerulescens*, *Rupornis magnirostris*, *Geranoaetus melanoleucus*, *Geranoaetus albicaudatus*, *Buteo brachyurus*, *Buteo albonotatus*, *Buteogallus meridionalis*
STRIGIDAE	*Megascops choliba*, *Glaucidium brasilianum*, *Aegolius harrisii*
TROGONIDAE	*Trogon curucui*
PICIDAE	*Picumnus limae*, *Veniliornis passerinus*, *Celeus ochraceus*, *Piculus chrysochloros*
FALCONIDAE	*Herpetotheres cachinnans*, *Micrastur ruficollis*, *Caracara plancus*
PSITTACIDAE	*Brotogeris chiriri*, *Forpus xanthopterygius*, *Pyrrhura griseipectus*, *Eupsittula cactorum*
THAMNOPHILIDAE	*Formicivora grisea*, *Formicivora melanogaster*, *Sakesphoroides cristatus*, *Herpsilochmus atricapillus*, *Thamnophilus capistratus*, *Thamnophilus pelzelni*, *Thamnophilus doliatus*, *Thamnophilus caerulescens*, *Taraba major*, *Radinopsyche sellowi*
DENDROCOLAPTIDAE	*Sittasomus griseicapillus*, *Dendroplex picus*
RHYNCHOCYCLIDAE	*Tolmomyias flaviventris*, *Hemitriccus margaritaceiventer*
TYRANNIDAE	*Hirundinea ferruginea*, *Myiopagis viridicata*, *Phaeomyias murina*, *Phyllomyias fasciatus*, *Myiarchus swainsoni*, *Myiarchus ferox*, *Myiarchus tyrannulus*, *Casiornis fuscus*, *Myiodynastes maculatus*, *Myiozetetes similis*, *Tyrannus melancholicus*, *Empidonomus varius*, *Knipolegus nigerrimus*, *Todirostrum cinereum*, *Camptostoma obsoletum*, *Nesotriccus murinus*, *Elaenia spectabilis*, *Myiophobus fasciatus*, *Pitangus sulphuratus*, *Megarynchus pitangua*
FURNARIIDAE	*Cranioleuca semicinerea*, *Synallaxis scutata*, *Synallaxis frontalis*, *Lepidocolaptes angustirostris*
VIREONIDAE	*Hylophilus amaurocephalus*, *Vireo chivi*, *Cyclarhis gujanensis*
CORVIDAE	*Cyanocorax cyanopogon*
POLIOPTILIDAE	*Polioptila plumbea*
TROGLODYTIDAE	*Troglodytes musculus*, *Cantorchilus longirostris*
TURDIDAE	*Turdus leucomelas*, *Turdus rufiventris*, *Turdus amaurochalinus*
ESTRILDIDAE	*Estrilda astrild*
FRINGILLIDAE	*Euphonia chlorotica*
PASSERIDAE	*Passer domesticus*
PASSERELLIDAE	*Arremon taciturnus*
ICTERIDAE	*Cacicus solitarius*, *Icterus pyrrhopterus*, *Icterus jamacaii*, *Chrysomus ruficapillus*, *Molothrus rufoaxillaris*, *Molothrus bonariensis*
PARULIDAE	*Setophaga pitiayumi*, *Myiothlypis flaveola*, *Basileuterus culicivorus*
CARDINALIDAE	*Piranga flava*

Continua na próxima página

Família	Espécies
THRAUPIDAE	*Nemosia pileata*, *Dacnis cayana*, *Volatinia jacarina*, *Coryphospingus pileatus*, *Tachyphonus rufus*, *Sporophila albogularis*, *Thlypopsis sordida*, *Conirostrum speciosum*, *Schistochlamys melanopis*, *Thraupis sayaca*, *Stilpnia cayana*, *Paroaria dominicana*, *Coereba flaveola*

4.2 Pré-processamento dos Áudios

No presente trabalho, o pré-processamento dos sinais de áudio foi realizado com o objetivo de padronizar as entradas do modelo e facilitar a extração eficiente de características relevantes. Foram adotadas duas abordagens distintas para representação dos dados: a utilização de matrizes numéricas armazenadas no formato .npy e a utilização de imagens de espectrogramas.

Na primeira abordagem, os sinais de áudio foram transformados em espectrogramas por meio da STFT, e posteriormente convertidos em matrizes de características, que foram armazenadas como arquivos NumPy (.npy). Essa estratégia possibilita o carregamento rápido dos dados e reduz o custo computacional durante o treinamento, uma vez que a etapa de extração de características não precisa ser repetida a cada execução.

Na segunda abordagem, os espectrogramas gerados foram salvos diretamente como imagens, permitindo sua utilização como entrada para arquiteturas convolucionais convencionais, explorando técnicas comuns de processamento de imagens.

Em ambas as abordagens, os áudios foram segmentados utilizando janelas temporais de tamanho fixo, garantindo uniformidade dimensional entre as amostras e compatibilidade com a arquitetura CNN-LSTM. Essa segmentação também contribui para o aumento do número de amostras disponíveis para treinamento e para a captura de padrões acústicos locais relevantes ao longo do tempo.

Essas estratégias de pré-processamento asseguram consistência nos dados de entrada e possibilitam a comparação do desempenho do modelo sob diferentes formas de representação do sinal acústico.

4.3 Aumento de Dados (Data Augmentation) - NÃO APLICADO!

Para ampliar a diversidade do conjunto de treinamento e reduzir o risco de sobreajuste, foram aplicadas técnicas de aumento de dados específicas para áudio, tais como:

- adição de ruído branco;
- variação de velocidade;
- alteração de tom (pitch shifting);
- deslocamento temporal.

As amostras geradas mantêm o rótulo original, aumentando a robustez do modelo frente a variações naturais do ambiente.

4.4 Arquitetura do Modelo CNN-LSTM

O modelo proposto é composto por:

- camadas convolucionais para extração de padrões espectrais;

- camadas de pooling para redução dimensional;
- camadas LSTM para modelagem temporal;
- camadas densas finais com função Softmax para classificação multiclasse.

A arquitetura foi projetada de forma a equilibrar desempenho e custo computacional.

4.5 Processo de Treinamento

O treinamento foi realizado utilizando:

- função de perda categórica (categorical cross-entropy);
- otimizador Adam;
- mini-batches;
- número fixo de épocas;
- técnica de early stopping para evitar overfitting.

Os pesos do modelo foram ajustados com base no conjunto de treinamento e monitorados pelo desempenho no conjunto de validação.

4.6 Avaliação do Desempenho

O desempenho do modelo foi avaliado no conjunto de teste por meio das seguintes métricas:

- acurácia;
- precisão;
- revocação;
- F1-score;
- matriz de confusão.

O objetivo estabelecido foi alcançar uma acurácia mínima de 75%.

5 Resultados

O modelo desenvolvido para identificação de espécies de aves por meio de seus cantos apresentou desempenho satisfatório, alcançando uma métrica de acurácia de aproximadamente 75% em um conjunto composto por diversas espécies. Os testes realizados com áudios gravados em campo e com amostras extraídas da internet demonstraram a capacidade do sistema em generalizar para diferentes condições de gravação, mantendo um nível elevado de acerto.

Os resultados obtidos evidenciam uma alta assertividade do modelo, indicando que a abordagem adotada — desde a coleta e consolidação dos dados até o treinamento e avaliação — foi adequada para o problema proposto. Dessa forma, conclui-se que o trabalho atingiu os objetivos esperados, validando a viabilidade do uso de técnicas de aprendizado de máquina para o reconhecimento automático de espécies de aves a partir de sinais acústicos.

6 Sugestões de Trabalhos Futuros

Como continuidade deste trabalho, diversas melhorias e extensões podem ser exploradas com o objetivo de aumentar a precisão do sistema, sua robustez e aplicabilidade prática. Inicialmente, propõe-se restringir o conjunto de espécies utilizadas no treinamento apenas àquelas que efetivamente ocorrem no município de Itapajé, uma vez que atualmente a extração de dados considera espécies distribuídas de forma mais ampla na região Nordeste. Essa especialização tende a reduzir ambiguidades e melhorar o desempenho do modelo em cenários locais.

Outra possibilidade consiste na ampliação das informações associadas a cada espécie, de modo a enriquecer a interface desenvolvida, permitindo apresentar ao usuário dados adicionais, como características biológicas, comportamento e imagens da ave identificada.

No âmbito metodológico, recomenda-se investigar o uso de aprendizagem por reforço, bem como a aplicação de técnicas de data augmentation sobre os sinais de áudio, com o intuito de avaliar uma possível melhora na convergência e na capacidade de generalização do modelo. Também pode ser considerada a segmentação do treinamento por grupos de espécies com características semelhantes, como porte físico (pequeno, médio e grande), visando reduzir a complexidade do problema de classificação.

Em relação à arquitetura do modelo, sugere-se testar uma abordagem CNN–LSTM em paralelo, em substituição à estrutura sequencial atualmente empregada, a fim de explorar simultaneamente padrões espaciais e temporais dos sinais acústicos.

Adicionalmente, melhorias no pré-processamento dos áudios capturados pela interface podem ser implementadas, incluindo filtragem de ruído, seleção manual ou automática do intervalo contendo o canto da ave e normalização ou amplificação do volume, aumentando assim a qualidade das entradas fornecidas ao modelo.

Como expansão funcional, destaca-se o desenvolvimento de uma interface web ou aplicativo móvel, facilitando o uso do sistema por pesquisadores e pelo público em geral. Também é pertinente explorar a identificação de espécies por meio de imagens, possibilitando um sistema multimodal (áudio e visão computacional).

Por fim, pode-se avaliar a implementação de uma classificação hierárquica em múltiplas etapas, na qual o modelo identifique inicialmente a família da ave e, posteriormente, a espécie. Embora essa abordagem apresente desafios devido à grande diversidade entre as espécies, ela representa uma alternativa promissora a ser investigada.

ATENÇÃO: Este relatório cobre apenas o projeto de identificação de espécies de aves, não se aplicando aos demais projetos que compõem o conjunto "bioitapajé", desenvolvido em 2025, pela a turma de Deep Learning do semestre 02.2025 do Curso de Ciência de Dados.

Para métricas mais precisas e referências/bibliográfias utilizadas, consultar o relatório final da disciplina.