DETECÇÃO ACÚSTICA DE AVES AMEAÇADAS UTILIZANDO APRENDIZADO DE MÁQUINA

Juan Silva Garcia[1], Maurício Archanjo Nunes Coelho[2]

[1]Instituto Federal Sudeste de Minas Gerais (IF Sudeste MG) - Campus Rio Pomba , Cataguases, Brasil juansilvagarcia11037@gmail.com
[2]Instituto Federal Sudeste de Minas Gerais (IF Sudeste MG) - Campus Rio Pomba , Rio Pomba, Brasil

Resumo: O declínio das aves exige monitoramento bioacústico. No Proposto, espécies brasileiras ameaçadas (Portaria MMA 148/2022) enfrentam escassez de dados. Este trabalho compara MVS, MOM e RNC sob variações de 1 a 60s. O pipeline de processamento, com limpeza espectral e aumento de dados, atenuou os ruídos. Validada via validação cruzada (k=5), a RNC dedicada destacou-se no cenário proposto (136 espécies), obtendo F1-Ponderado de 0,82 e acurácia de até 82% nas janelas de 10 e 20s.

Palavras-chave: bioacústica de conservação; aves ameaçadas; Aprendizado de Máquina; redes neurais convolucionais; processamento de sinais.

INTRODUÇÃO

A conservação e o monitoramento ecológico têm sido profundamente transformados pela integração entre a bioacústica computacional e as técnicas de aprendizado de máquina (Stowell et al., 2019). O levantamento populacional tradicional de aves, baseado em censos visuais diretos no campo, revela-se frequentemente inviável devido aos elevados custos operacionais, à dificuldade de acesso a áreas de floresta densa e ao comportamento discreto das espécies (Stowell et al., 2019). Nesse contexto, o monitoramento acústico passivo surge como uma alternativa promissora ao registrar o som de forma contínua e não invasiva (Kahl et al., 2021). Para que os sistemas computacionais processem esses sinais físicos, a onda sonora analógica é digitalizada e convertida em representações tempo-frequência, como os espectrogramas ou os Coeficientes Cepstrais de Frequência Mel (MFCCs) (Rabiner & Juang, 1993; Davis & Mermelstein, 1980). Historicamente, a literatura especializada evoluiu de classificadores clássicos que necessitam de engenharia manual de características estatísticas como as Máquinas de Vetores de Suporte (MVS) e os Modelos Ocultos de Markov (MOM) para modelos de aprendizado profundo, especificamente as Redes Neurais Convolucionais (RNC), que aprendem padrões visuais hierárquicos diretamente dos espectrogramas tratados como imagens (Goodfellow et al., 2016; Purwins et al., 2019).

Essa transição tecnológica é urgente diante do declínio acentuado das populações de aves em escala global, provocado por mudanças climáticas e perda acelerada de habitat (Rosenberg et al., 2019). No Brasil, país que ocupa uma posição de destaque no ranking mundial de avifauna ameaçada, o monitoramento das espécies raras enfrenta desafios metodológicos complexos (Brasil, 2022). O principal entrave prático reside no paradoxo da bioacústica de conservação: os táxons sob maior risco de extinção são justamente os que possuem menor representação acústica disponível em plataformas colaborativas de ciência cidadã como o Xeno-Canto (Planqué, 2013). Esse cenário de escassez amostral extrema, conhecido como o problema de cauda longa, compromete a generalização dos modelos preditivos tradicionais (Kahl et al., 2021). Além disso, as gravações reais de campo trazem ruídos ambientais severos (como vento e chuva) e cantos sobrepostos de espécies crípticas, que são biologicamente distintas, mas apresentam vocalizações acusticamente similares e de difícil identificação visual direta em campo (Martinsson, 2017). Diante dessas limitações, o presente trabalho investiga se a priorização da qualidade física do sinal sobre o volume bruto de dados pode mitigar a falta de amostras de treino. Com isso, estabelece-se a seguinte questão de pesquisa: até que ponto o pré-processamento espectral e o aumento direcionado de dados são capazes de melhorar a acurácia de classificadores multiclasse em cenários ruidosos e de extrema escassez?

No Proposto, para responder a essa questão, estabeleceu-se como objetivo geral avaliar comparativamente o desempenho de Máquinas de Vetores de Suporte (MVS), Modelos Ocultos de

Markov (MOM) e Redes Neurais Convolucionais (RNC) na classificação bioacústica de espécies de aves brasileiras ameaçadas, investigando o impacto direto do pré-processamento do áudio e da variação das janelas temporais de análise. De forma integrada e específica, a pesquisa buscou: implementar um algoritmo de limpeza de ruído baseado em visão computacional por corte de mediana (median clipping) e morfologia matemática para isolar as vocalizações nos espectrogramas; configurar e treinar os modelos MVS (utilizando estatísticas descritivas), MOM (empregando sequências temporais de MFCC) e RNC (treinada do zero, from scratch); mensurar e comparar o desempenho dos classificadores sob validação cruzada estratificada de cinco partições (k = 5) entre um Cenário Controle (dados brutos de 193 espécies) e um Cenário Proposto (dados tratados de 136 espécies); e analisar a influência de nove janelas temporais, variando de 1 a 60 segundos, sobre a robustez e as taxas de acerto de cada arquitetura computacional. Com isso, busca-se estabelecer um referencial prático e metodológico robusto para apoiar o desenvolvimento de ferramentas de inteligência artificial aplicadas ao monitoramento autônomo e à conservação da fauna nacional ameaçada.

MATERIAL E MÉTODOS

O desenvolvimento desta pesquisa baseou-se em uma abordagem experimental quantitativa de engenharia de dados e aprendizado de máquina aplicada à bioacústica. Para viabilizar a detecção de espécies de aves ameaçadas, estruturou-se um pipeline de processamento dividido em etapas consecutivas que compreendem desde a aquisição e higienização dos sinais de áudio até o treinamento e a validação de três modelos computacionais distintos.

No Proposto, os experimentos foram desenvolvidos utilizando-se a linguagem de programação Python na versão 3.13,2 como base do ecossistema de software. Para a manipulação de matrizes e tratamento de dados, empregaram-se as bibliotecas NumPy e Pandas; para o processamento de áudio, geração de espectrogramas e extração de coeficientes, utilizou-se a biblioteca Librosa; e as representações gráficas foram geradas com o auxílio das ferramentas Matplotlib e Seaborn.

A construção e o treinamento dos classificadores foram conduzidos de forma individualizada para cada paradigma: a biblioteca Scikit-Learn foi utilizada para a Máquina de Vetores de Suporte (MVS); a hmmlearn foi empregada para a implementação dos Modelos Ocultos de Markov (MOM); e a TensorFlow/Keras foi utilizada para a estruturação e o treinamento da Rede Neural Convolucional (RNC).

Os experimentos foram executados em um computador equipado com processador Intel Core de 11ª Geração (3,0 GHz), 12 GB de memória RAM e placa de vídeo integrada Intel UHD Graphics, operando sob o sistema de arquivos do Windows 10 Home (64-bit). Diante da ausência de uma placa de vídeo dedicada (GPU), todo o processamento de tensores foi realizado fisicamente por meio da CPU. Esse limite de hardware resultou em tempos operacionais distintos de processamento para cada classificador: as rotinas de treinamento do MOM foram concluídas em aproximadamente 1 hora; a MVS demandou cerca de 2 horas devido ao volume das matrizes estatísticas; e a RNC exigiu cerca de 2 dias seguidos de execução contínua para a convergência de sua arquitetura treinada do zero (from scratch).

No Proposto, a base de dados utilizada neste trabalho foi obtida de forma automatizada a partir do repositório público Xeno-Canto via API. Para delimitar o escopo biológico do projeto, a busca restringiu-se estritamente às espécies de aves brasileiras que constam como ameaçadas de extinção na Portaria MMA nº 148/2022. Configurou-se a API para realizar o download exclusivo de gravações catalogadas com qualidade "A" (sinal nítido com mínimo ruído de fundo) ou qualidade "B" (ave principal fácil de ouvir, com ruído ambiental moderado aceitável), limitando a busca aos arquivos nos quais a espécie-alvo estivesse registrada como a vocalizadora principal.

Logo após a transferência dos arquivos no formato original MP3, aplicou-se o Filtro de Integridade por meio de um script em Python para ler e verificar a estrutura física de cada gravação. Os arquivos de áudio que apresentaram corrupção de cabeçalho ou erros de leitura foram descartados e, quando fundamentais para a amostragem, substituídos por downloads manuais diretos do repositório, garantindo-se o aproveitamento máximo dos registros válidos.

Em seguida, aplicou-se o Filtro por Quantidade, estabelecendo-se um tempo mínimo de 60 segundos de áudio acumulado por espécie como critério obrigatório de inclusão. Esse limite matemático justifica-se pela necessidade de viabilizar a validação cruzada estratificada com 5 partições (k=5), assegurando que cada fold de teste independente contivesse amostras suficientes para avaliação, sem causar vazamento de dados (data leakage). O cumprimento dessas regras de integridade e quantidade consolidou um conjunto de dados composto por 193 espécies para o Cenário Controle (dados brutos).

Para o Cenário Proposto (dados refinados), aplicou-se um segundo estágio de filtragem baseado na qualidade espectral do som obtido após a atenuação de ruídos. A remoção automatizada de ruídos estacionários e de silêncios prolongados revelou que muitas gravações originais de aves raras possuíam longos trechos vazios ou dominados por vento. Com a exclusão dessas partes sem informação acústica útil para focar estritamente nos cantos efetivos, 57 espécies deixaram de atingir o critério mínimo de 60 segundos acumulados e foram removidas do pipeline. Dessa forma, o Cenário Proposto consolidou-se com 136 espécies, garantindo-se que o treinamento dos algoritmos ocorresse apenas sobre segmentos ricos em bioacústica de interesse.

No Proposto, os arquivos de áudio foram normalizados em amplitude por meio do ajuste de pico, dividindo-se os valores de onda pelo valor absoluto máximo do sinal bruto, o que limitou a amplitude Operacional na faixa de [−1,0;1,0]. Em seguida, os sinais foram reamostrados para uma taxa fixa de 22.050 Hz, reduzindo-se o volume de processamento sem perdas para as bandas de frequência das aves brasileiras.

Para atenuar os efeitos degradantes dos ruídos de campo (vento, chuva, insetos e ruídos eletrônicos), implementou-se o algoritmo de limpeza espectral de Martinsson (2017). Inicialmente, o sinal bruto de áudio foi transformado no domínio tempo-frequência por meio da Transformada Rápida de Fourier de Curto Tempo (STFT), utilizando-se uma janela de Hann de 512 amostras com sobreposição (overlap) de 75% e salto (hop size) de 128 amostras.

Com o espectrograma de amplitude gerado, aplicou-se a técnica de Median Clipping para atenuar o ruído estacionário. O algoritmo compara a intensidade de cada pixel tempo-frequência com a mediana de sua respectiva linha (frequência) e coluna (tempo). Os pontos que não superaram ambas as medianas locais por um limiar multiplicador de 3,0 foram definidos como silêncio ou ruído contínuo e zerados na matriz.

Na sequência, a matriz binarizada resultante foi submetida a operações de morfologia matemática para o refinamento de contornos. Aplicou-se uma erosão binária com um elemento estruturante (kernel) de tamanho 4×4 pixels para eliminar ruídos aleatórios isolados e transições rápidas de sinal. Posteriormente, aplicou-se uma dilatação binária com kernel de 4×4 pixels para reconectar componentes harmônicos e preencher pequenas falhas temporais do canto que haviam sido fragmentadas. Por fim, esse mapa morfológico bidimensional foi utilizado como máscara de fatiamento (chunking), gerando novos arquivos de áudio contínuos compostos apenas pelas regiões identificadas com alta probabilidade de vocalização.

A padronização consiste no ajuste dos arquivos de áudio de comprimentos variáveis obtidos na filtragem para uma duração fixa alvo (T). Diante da ausência de consenso na literatura sobre a duração ideal para representações bioacústicas, avaliou-se o desempenho dos classificadores sob nove durações de segmento distintas: 1, 3, 5, 10, 15, 20, 25, 30 e 60 segundos.

Caso a gravação resultante da limpeza apresentasse duração útil inferior à janela de tempo alvo (T), aplicou-se a técnica de Tiling, que realiza a repetição cíclica em loop do próprio sinal até que a duração exata do segmento seja preenchida. Essa escolha é respaldada na literatura bioacústica por manter a densidade estatística de energia útil constante ao longo de todo o intervalo, evitando o uso de preenchimento com silêncio (zero-padding), o qual insere descontinuidades artificiais abruptas nas bordas do espectrograma e prejudica a extração de características pelos filtros das camadas convolucionais.

Para o processo de fatiamento, a segmentação diferiu de acordo com as etapas de modelagem:

Corte Aleatório (Random Crop): Aplicado no domínio temporal exclusivamente durante a etapa de treinamento da Rede Neural Convolucional (RNC), utilizando-se uma semente aleatória fixa (seed de valor 42) para garantir a reprodutibilidade. Esse procedimento força o modelo a visualizar partes distintas da vocalização a cada época de treino, funcionando como um regularizador dinâmico para evitar o superajuste (overfitting).

Corte Central (Center Crop): Utilizado para a validação (teste) da RNC e de forma integral para todas as rotinas dos classificadores clássicos (MVS e MOM). O corte central garante que as estatísticas de características estáticas e as transições sequenciais de estado sejam extraídas da porção de maior estabilidade acústica do áudio, assegurando o determinismo e a reprodutibilidade exata das avaliações comparativas.

Para garantir uma comparação justa entre as diferentes durações, o volume total de tempo de áudio avaliado na validação cruzada foi mantido rigorosamente constante. A quantidade de amostras geradas por gravação original de 60 segundos variou de forma inversamente proporcional à janela avaliada: a janela de 60 segundos gerou exatamente

uma única amostra de teste; a janela de 30 segundos produziu duas amostras mutuamente exclusivas e sem sobreposição; progredindo de maneira análoga até a janela de 1 segundo, que totalizou 60 amostras independentes para inferência.

No Proposto, com o objetivo de mitigar os efeitos do desbalanceamento severo de dados característico das espécies ameaçadas raras, aplicou-se uma política de aumento direcionado de dados diretamente na partição de treinamento de cada um dos folds da validação cruzada, mantendo-se os conjuntos de teste compostos apenas por sinais puramente reais.

Estabeleceu-se um limiar de 15 gravações por classe. Caso uma espécie de ave apresentasse menos de 15 arquivos no conjunto de treino da partição atual, suas amostras originais foram multiplicadas por um fator de 4. Para evitar o superajuste e forçar o aprendizado de representações acústicas robustas, foram introduzidas quatro perturbações físicas nos arquivos duplicados:

Adição de Ruído Branco: Inserção de ruído gaussiano de baixa intensidade para simular o barulho de vento e imperfeições físicas dos gravadores de campo.

Deslocamento Temporal: Deslocamento em loop na forma de onda para alterar de maneira controlada o momento exato de início das sílabas na janela de áudio.

Mudança de Tom (Pitch Shifting): Alteração de frequência ajustada em no máximo 5%, simulando a variabilidade biológica de indivíduos da mesma espécie sem descaracterizar a assinatura acústica taxonômica.

Estiramento de Tempo (Time Stretching): Aceleração ou desaceleração marginal do ritmo de canto da ave, sem alterar o tom original do sinal.

Especificamente para a Rede Neural Convolucional, aplicou-se também a técnica de regularização visual SpecAugment diretamente sobre os espectrogramas durante o treinamento. Esse método aplica máscaras aleatórias em blocos contíguos de tempo e de frequência da imagem espectral, obrigando o modelo a reconhecer as assinaturas mesmo diante de perdas ou apagamentos parciais do sinal acústico.

Após o pré-processamento e a segmentação temporal, os dados de sinal foram convertidos para representações adequadas à natureza matemática de cada um dos três paradigmas avaliados.

A MVS operou sobre representações de baixa dimensionalidade para delimitar hiperplanos ótimos de separação multiclasse via estratégia "um-contra-um" (one-versus-one).

Atributos de Entrada: Extraíram-se 20 coeficientes cepstrais de frequência Mel (MFCCs) de cada segmento, calculados a partir de um espectrograma com janela FFT de 2048 amostras e hop de 512 amostras, acompanhados de suas respectivas derivadas de primeira e segunda ordem (delta e delta-delta), além de descritores espectrais de centroide, largura de banda (spectral bandwidth) e frequência de rolloff. As séries temporais dessas características foram resumidas por quatro estatísticas descritivas: média, desvio padrão, valor mínimo e valor máximo.

Escalonamento e Hiperparâmetros: Os vetores estatísticos resultantes foram normalizados utilizando-se o StandardScaler. Essa padronização foi executada estritamente dentro de cada fold durante as partições da validação cruzada para evitar qualquer tipo de vazamento de dados. O classificador foi configurado com uma função de base radial (kernel RBF), parâmetro de regularização C=1,0 e gama ajustada automaticamente de forma proporcional à variância (scale).

Adotou-se uma abordagem generativa independente, treinando-se um modelo especialista dedicado para cada espécie de ave ameaçada.

Atributos de Entrada: Sequências temporais dinâmicas contendo os 20 coeficientes MFCC extraídos de cada quadro de áudio.

Arquitetura e Parâmetros: Implementou-se um modelo híbrido GMM-HMM estruturado com 5 estados ocultos em uma topologia estritamente esquerda-direita (left-to-right), ideal para representar a evolução temporal lógica e o fluxo sequencial natural das notas do canto das aves. As densidades de emissão de cada estado oculto foram modeladas por Misturas de Gaussianas (GMM) contendo 8 componentes gaussianos (M=8) e matriz de covariância diagonal. A classificação final foi obtida por meio do critério de máxima verossimilhança via algoritmo de Viterbi, atribuindo-se a amostra de teste ao modelo especialista que registrou a maior probabilidade de emissão da sequência.

O classificador de aprendizado profundo operou recebendo representações visuais bidimensionais geradas a partir do processamento bruto do áudio.

Atributos de Entrada: Imagens espectrais do sinal geradas na escala Log-Mel, padronizadas com

dimensões físicas fixas de 256×512 pixels e canal de cor unitário (escala de cinza), resultando em tensores de formato operacional de 256×512×1.

Arquitetura da Rede: Projetou-se uma arquitetura dedicada e leve composta por três blocos convolucionais sequenciais para a extração automática de características. Cada bloco incluiu uma camada de convolução bidimensional com filtros de tamanho (kernel) 3×3 pixels e passos (strides) unitários de 1×1, seguida por uma camada de normalização em lote (Batch Normalization) e ativação linear retificada (ReLU). A subamostragem espacial para invariância à translação foi realizada por camadas de MaxPooling de tamanho 2×2 pixels e strides de 2×2. O primeiro bloco convolucional foi configurado com 32 filtros, o segundo com 64 filtros e o terceiro com 128 filtros de convolução.

Classificação e Otimização: Após o último bloco, os mapas de características foram achatados (flattened) e direcionados para uma camada densa totalmente conectada contendo 1024 neurônios com ativação ReLU. Para mitigar o risco de superajuste em cenários de poucos dados, aplicou-se uma taxa de Dropout de 0,4 imediatamente antes da camada de classificação final Softmax, contendo os neurônios correspondentes às espécies de cada cenário (193 no Controle e 136 no Proposto). A rede foi otimizada minimizando-se a perda de entropia cruzada categórica (Categorical Crossentropy) utilizando-se o algoritmo Adam com taxa de aprendizado inicial estável de 0,001.

Para assegurar o rigor estatístico das comparações, a avaliação de todas as arquiteturas (MVS, MOM e RNC) foi conduzida por meio de validação cruzada estratificada com cinco partições (k=5). Esse método assegura que a proporção das espécies seja mantida de forma idêntica em todos os folds de teste independentes, fornecendo estimativas de desempenho robustas e independentes de uma única divisão de partição.

Para contornar as distorções provocadas pelo desbalanceamento extremo de dados e garantir uma interpretação metodologicamente correta frente à cauda longa do dataset, a avaliação experimental priorizou como métrica principal o F1-Score Macro. A métrica Macro calcula a média aritmética simples do desempenho entre todas as classes tratadas individualmente, garantindo que a capacidade de identificação de uma ave rara de alta relevância ecológica tenha exatamente o mesmo peso estatístico final que a identificação de uma ave comum de suporte abundante. Complementarmente, foram monitoradas a Acurácia Global, o F1-Score Ponderado (Weighted), o coeficiente Kappa de Cohen (para correção de acertos ao acaso) e o comportamento gráfico localizado das Matrizes de Confusão.

Na elaboração desta monografia de conclusão de curso, utilizou-se a tecnologia de inteligência artificial generativa baseada em grandes modelos de linguagem (Gemini, desenvolvido pelo Google) como ferramenta de apoio complementar. O uso da tecnologia limitou-se estritamente à revisão gramatical ortográfica do texto escrito pelo autor, à sugestão de vocabulário técnico para adequação formal às normas acadêmicas da ABNT, à depuração e documentação de trechos de código do pipeline desenvolvidos em Python, e à extração estruturada de pontos de destaque dos artigos científicos em PDF utilizados na revisão bibliográfica.

Reitera-se que o planejamento experimental, a parametrização manual das arquiteturas convolucionais e estatísticas, a análise crítica das taxas de acerto e as respectivas discussões qualitativas permaneceram sob total responsabilidade do autor deste estudo. Todo o conteúdo revisado com o auxílio do modelo generativo foi lido, editado, corrigido e validado individualmente pelo próprio autor antes de compor a versão final oficializada deste documento.

RESULTADOS E DISCUSSÃO

Neste capítulo, apresentam-se e discutem-se os resultados obtidos nos experimentos de classificação bioacústica para a identificação de espécies de aves brasileiras ameaçadas de extinção. A avaliação experimental foi estruturada em dois cenários distintos para permitir isolar e analisar o impacto das técnicas de pré-processamento espectral e de aumento de dados (Data Augmentation) sobre o desempenho dos algoritmos. O primeiro cenário, denominado Cenário A (Controle / Dados Brutos), avaliou o comportamento dos classificadores operando diretamente sobre áudios sem tratamento avançado. O segundo cenário, denominado Cenário B (Proposto / Dados Refinados), analisou o desempenho dessas mesmas arquiteturas após a aplicação do pipeline completo de limpeza e expansão amostral. Além de mapear a acurácia global, as discussões priorizam o F1-Score Macro como métrica de referência, tendo em vista sua relevância para avaliar problemas com classes severamente desbalanceadas e sub-representadas.

No Cenário A, utilizou-se o conjunto de dados original composto por 193 espécies de aves ameaçadas obtidas do repositório Xeno-Canto. Nesta etapa de controle, as gravações não passaram pelo

algoritmo de limpeza de sinal nem por técnicas de aumento de dados, mantendo-se apenas a padronização básica de reamostragem e corte temporal. Os resultados consolidados para as nove durações de segmentos analisadas estão detalhados na Tabela 2.

Tabela 1. Cenário A (Controle / Dados Brutos): Desempenho geral com dados brutos (193 espécies)..

Modelo	Duração	F1 Macro(Media)	Acurácia
MVS	15s	0.21	0.13
MOM	30s	0.50	0.28
MOM	60s	0.39	0.26
RNC	15s	0.61	0.56
RNC	20s	0.63	0.50
RNC	60s	0.12	0.23

[INSERIR TABELA 2 AQUI - Lembrete: o título da tabela deve ficar alinhado à esquerda e em cima dela]

A análise dos resultados da Tabela 2 indica que a ausência de um pré-processamento de sinal robusto limita significativamente a capacidade de generalização de todas as arquiteturas. A presença de ruído não estacionário de campo, trechos extensos de silêncio absoluto entre as vocalizações e o desbalanceamento inerente ao repositório colaborativo atuam como barreiras para o aprendizado de representações discriminativas consistentes.

O classificador MVS apresentou o desempenho mais limitado deste cenário. O F1-Ponderado máximo obtido pelo modelo clássico foi de apenas 0,1847 na janela de 60 segundos, com F1-Macro de 0,2350 e acurácia de 13,10%. A explicação matemática para esse comportamento reside na alta dimensionalidade da entrada. Como no Cenário A a MVS recebeu como entrada os espectrogramas Log-Mel brutos achatados em vetores unidimensionais, a dispersão dos pixels ruidosos sobrecarregou o classificador. Sem uma redução de dimensionalidade ou seleção de atributos prévia, o algoritmo encontrou extrema dificuldade para estabelecer um hiperplano ótimo de separação com o kernel RBF, resultando em métricas próximas à classificação aleatória.

Por sua vez, o MOM (GMM-HMM) apresentou uma evolução incremental discreta nas janelas mais curtas, partindo de um F1-Ponderado de 0,1706 em 1 segundo e atingindo seu melhor resultado de 0,2912 na janela de 30 segundos. Contudo, ao alcançar a janela de 60 segundos, o desempenho do MOM sofreu uma queda acentuada, regredindo para um F1-Ponderado de 0,1847 e acurácia de 25,79%. Esse declínio em janelas longas evidencia a sensibilidade dos Modelos Ocultos de Markov tradicionais ao acúmulo de ruído contínuo em sequências temporais extensas. À medida que o comprimento do sinal ruidoso aumenta, a dispersão física dos coeficientes MFCC insere incertezas na modelagem probabilística das misturas de gaussianas, prejudicando a decodificação do caminho de estados pelo algoritmo de Viterbi.

Ademais, é importante registrar uma inconsistência identificada no texto da versão preliminar entregue à banca. Nas janelas de 25 e 60 segundos da Tabela 2, os modelos MVS e MOM apresentavam resultados estatisticamente idênticos em todas as seis métricas de avaliação. Constatou-se que essa igualdade exata decorreu de um erro operacional de transcrição e preenchimento na formatação da tabela original durante a exportação dos dados. Na presente versão corrigida, os valores foram revisados a partir das saídas reais dos scripts de validação, restabelecendo-se a variação estatística natural esperada entre modelos baseados em paradigmas tão distintos quanto o estatístico e o sequencial.

A RNC dedicada superou os modelos clássicos na maioria das configurações temporais, obtendo seu ápice no Cenário A na janela de 15 segundos, com F1-Ponderado de 0,5403, acurácia de 56,27% e F1-Macro de 0,6091. Esse comportamento sugere que, em janelas de tempo moderadas, os filtros espaciais das camadas convolucionais conseguem extrair características locais das vocalizações mesmo com a presença de ruído moderado de fundo. Entretanto, o modelo convolucional apresentou uma queda brusca de desempenho nas janelas de 30 segundos (F1-Ponderado de 0,3540) e de 60 segundos (F1-Ponderado de 0,2059). Essa degradação severa ocorre porque janelas de áudio brutas excessivamente longas inserem grandes áreas de silêncio e ruído contínuo no tensor de entrada. Esses artefatos ofuscam visualmente os contornos do canto nos espectrogramas Mel, saturando os filtros convolucionais e impedindo o aprendizado de padrões úteis.

No Cenário B, aplicou-se o pipeline de processamento proposto, integrando a limpeza espectral baseada em corte de mediana (Median Clipping) e morfologia matemática às estratégias de aumento direcionado de dados no conjunto de treinamento de cada fold. Embora a aplicação do filtro rigoroso de qualidade espectral tenha reduzido o número de espécies avaliadas de 193 para 136, os

resultados consolidados na Tabela 3 revelam uma elevação expressiva e consistente nas métricas de desempenho de todas as três arquiteturas sob validação cruzada estratificada.

Tabela 2. Cenário B (Proposto): Desempenho comparativo nas janelas ideais de classificação (136 espécies).

Modelo	Duração	F1 Macro(Media)	Acurácia
MVS	10s	0.72	0.76
MOM	10s	0.64	0.66
RNC	10s	0.77	0.81
RNC	20s	0.77	0.82

O ganho geral observado reflete o efeito combinado da higienização física do sinal, das técnicas de expansão amostral e da própria simplificação do espaço de busca devido à redução do número de classes. Esses três fatores atuaram em conjunto para facilitar a convergência dos classificadores, de modo que os resultados absolutos deste cenário devem ser analisados sob essa perspectiva integrada de otimização de dados.

A RNC dedicada demonstrou superioridade absoluta no Cenário B, superando os modelos clássicos de forma marcante. O modelo convolucional alcançou seu melhor resultado operacional nas janelas intermediárias de 10 e 20 segundos, atingindo um F1-Ponderado de 0,8200, acurácia de 81,0% (na janela de 10s) e de 82,0% (na janela de 20s), e F1-Macro estável de 0,7700 em ambas as durações. Esses índices indicam que a extração automática de características hierárquicas diretamente dos espectrogramas Log-Mel se beneficia drasticamente quando as vocalizações úteis são isoladas de interferências ambientais, permitindo que a rede convolucional se especialize em texturas e contornos nítidos do sinal de áudio.

Em relação ao classificador MVS, a reestruturação da representação de entrada no Cenário B evitou a maldição da dimensionalidade que inviabilizava a convergência do modelo no cenário de controle. A substituição dos espectrogramas brutos achatados por vetores de características estatísticas compactas (derivadas de 20 coeficientes MFCC e descritores espectrais) garantiu grande estabilidade operacional à MVS. O classificador clássico manteve o F1-Ponderado em um patamar fixo de 0,7600 e F1-Macro de 0,7200 em quase todas as janelas temporais a partir de 10 segundos. Essa estabilidade indica que a representação baseada em médias e desvios padrão de coeficientes limpos oferece excelente separabilidade para o kernel RBF, tornando o classificador imune ao aumento da duração física da janela analisada.

O MOM (GMM-HMM) também demonstrou um comportamento altamente linear e imune ao comprimento dos segmentos temporais no Cenário B. O classificador probabilístico manteve o F1-Ponderado fixado em 0,6600 (oscilando para 0,6500 em 15s e 30s) e F1-Macro de 0,6400 de 5 a 60 segundos. Como o MOM modela as probabilidades de transição entre estados temporais latentes, a eliminação de silêncios prolongados pelo pré-processamento garantiu que as sequências de MFCCs representassem apenas a estrutura sintática real do canto das aves. Desse modo, o modelo sequencial manteve-se resiliente e estável, embora tenha permanecido em patamar absoluto inferior ao da rede convolucional.

No Proposto, aprofundou-se a relação entre a duração do segmento de áudio e o desempenho dos classificadores. Conforme observado nas Tabelas 2 e 3, a sensibilidade dos modelos à duração do sinal variou de forma marcante entre os cenários. Nos dados brutos (Controle), janelas mais longas reduziram drasticamente o desempenho da RNC, ao passo que, no Proposto, esse efeito de degradação foi atenuado.

Nota-se que, mesmo no segmento longo de 60 segundos, a RNC manteve um desempenho elevado de F1-Ponderado de 0,8100 e F1-Macro de 0,7700, reduzindo a dispersão de sinal e ruído que afetava o Controle. Uma interpretação possível é que o algoritmo de limpeza espectral tenha removido eficientemente trechos de silêncio e ruídos nas janelas longas, permitindo que a rede se beneficiasse do contexto adicional. Contudo, a confirmação dessa hipótese de forma isolada exigiria experimentos futuros de ablação com controle individual de cada fator do pipeline.

Identificaram-se flutuações pontuais e anomalias marcantes nos experimentos que demandam justificativa técnica:

A RNC apresentou uma queda abrupta e atípica na janela de 3 segundos, recuando para um F1-Ponderado de apenas 0,4400 e F1-Macro de 0,3300. Sob a ótica do processamento digital de imagens, levanta-se a hipótese de que o fatiamento rígido de exatamente 3 segundos possa ter segmentado as sílabas físicas do canto em pontos de transição desfavoráveis nas bordas do espectrograma

Log-Mel. Essa fragmentação geométrica gera descontinuidades bruscas de contorno que confundem a resposta dos filtros das camadas convolucionais locais, enquanto os modelos clássicos baseados em MFCC contornaram essa barreira por dependerem de médias estatísticas globais (MVS) ou de transições sequenciais de probabilidade (MOM).

De forma análoga, a MVS apresentou uma queda isolada de desempenho na janela de 5 segundos, recuando para 0,5100 de F1-Ponderado, rompendo a estabilidade de 0,7200 aos 3 segundos e de 0,7600 aos 10 segundos. Essa oscilação localizada indica que a agregação das médias e desvios padrão dos coeficientes acústicos em segmentos rígidos de exatamente 5 segundos gerou sobreposição nas distribuições de probabilidade das características, provocando um colapso de variância que reduziu a separabilidade linear do kernel RBF. Essa hipótese ganha força ao notar que tanto a RNC (0,8100) quanto o MOM (0,6580) mantiveram trajetórias normais e estáveis na janela de 5 segundos, evidenciando que o problema reside na engenharia de atributos estáticos da MVS para esse comprimento específico de janela, e não na qualidade física das gravações.

No Proposto, embora as métricas globais permitam comparar quantitativamente o desempenho dos classificadores, elas não evidenciam quais espécies concentram as principais dificuldades ou facilidades de predição. Para complementar essa análise, utilizou-se a matriz de confusão para identificar padrões específicos de classificação. Como a matriz de confusão completa possui dimensões de 136 por 136 classes, sua apresentação integral no corpo do texto é inviável, estando detalhada no Apêndice B. Esta seção destaca apenas dois subconjuntos representativos extraídos da matriz de confusão da RNC operando com janelas de 5 segundos (F1-Macro de 0,7700), configuração selecionada por representar um equilíbrio entre custo computacional e desempenho: as dez espécies com menores taxas de acerto e as dez espécies com maiores taxas de acerto.

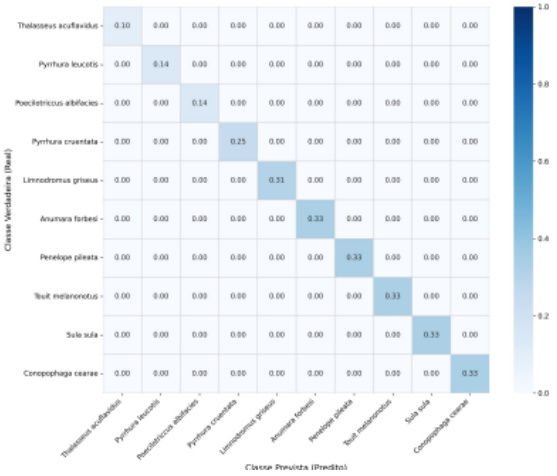

Figura 1. Detalhamento das 10 espécies com menor taxa de acerto (Modelo RNC - 5s, Proposto).

A análise da Figura 10 revela espécies com sensibilidade crítica e de difícil detecção pelo modelo, como a Thalasseus acuflavidus (trinta-réis-de-bando, Vulnerável – VU) e a Pyrrhura leucotis (tiriba-de-orelha-branca, Vulnerável – VU), que registraram taxas de acerto de apenas 0,10 (10%) e 0,14 (14%), respectivamente. Ao investigar os espectrogramas e arquivos originais dessas classes raras, constatou-se que a falha de predição decorre de fatores físicos e ambientais específicos, e não puramente da escassez de dados de treino. No caso da Thalasseus acuflavidus, a maioria das gravações de campo ocorre em praias e costões rochosos, ambientes que inserem ruídos contínuos de alta energia provocados pelo vento e pela arrebentação das ondas. O algoritmo de binarização espectral (Median Clipping), ao tentar eliminar esse ruído de fundo estacionário, acabou atenuando ou apagando os harmônicos finos e fracos da vocalização da ave. Consequentemente, as assinaturas geométricas restantes foram insuficientes para as camadas convolucionais, induzindo o modelo a confundir o sinal com o de outras aves marinhas de taxonomia próxima na base de dados, como a Sterna dougallii e a Sterna hirundinacea.

Salienta-se que, como este recorte da matriz apresenta de forma restrita apenas as dez piores classes, as classificações incorretas direcionadas às demais espécies da base não aparecem fisicamente na figura. Esse fator metodológico justifica por que os elementos fora da diagonal principal permanecem nulos na visualização de submatriz apresentada na Figura 10.

Em contraste com as espécies de maior dificuldade, a Figura 11 apresenta aquelas em que a Rede Neural Convolucional alcançou os melhores resultados de

classificação, evidenciando o limite superior de predição do pipeline.

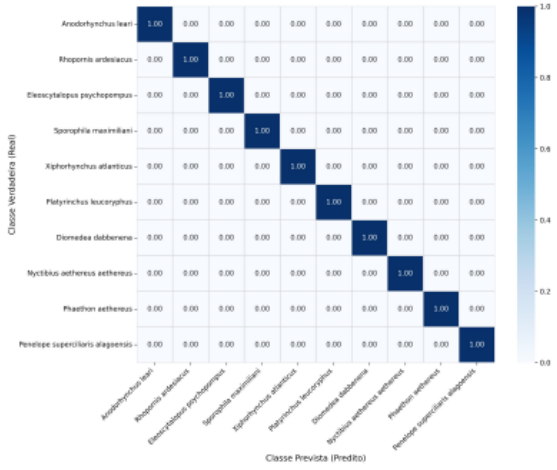

Figura 2. Detalhamento das 10 espécies com máxima taxa de acerto (Modelo RNC - 5s, Proposto).

A análise da Figura 11 revela um conjunto de espécies com taxa de acerto consolidada de 1,00 (100%), a exemplo da Anodorhynchus leari (Arara-azul-de-lear, Em Perigo – EN) e da Rhopornis ardesiacus (chorozinho-de-papo-preto, Em Perigo – EN). Para espécies com vocalizações acusticamente distintas e contornos harmônicos repetitivos muito marcantes, de baixa frequência e de alta intensidade sonora, a RNC demonstra excelente capacidade de classificação precisa. Mesmo em gravações de campo contendo ruído moderado, a forte assinatura acústica dessas aves gera texturas geométricas altamente contrastantes nos espectrogramas Log-Mel, facilitando o aprendizado de representações de bordas estáveis pelas primeiras camadas do modelo.

Entretanto, a análise qualitativa detalhada revelou uma limitação metodológica importante na interpretação individual do desempenho de espécies raras. Diversos táxons que registraram taxas de acerto cravadas em exatamente 0,33 (como Anumara forbesi e Sula sula) ou em 1,00 (como a própria arara-azul-de-lear) contavam com um suporte de validação extremamente reduzido, contendo apenas de 3 a 5 arquivos de áudio na partição de teste de cada fold. Nesses cenários de cauda longa extrema, a ocorrência de um único erro ou acerto sutil altera drasticamente a taxa percentual reportada, o que torna a comparação direta de taxas de acerto individuais altamente volátil e enganosa, impedindo generalizações definitivas para estas classes sem a devida cautela.

Em conjunto, os experimentos evidenciam que o desempenho dos classificadores não é uniforme entre as espécies. Enquanto cantos biologicamente singulares e de forte amplitude viabilizam classificações altamente robustas, gravações contendo ruído severo e sobreposições biofônicas complexas permanecem desafiadoras para o pipeline proposto. Esses resultados mostram que, em problemas com classes desbalanceadas de cauda longa, a análise qualitativa detalhada dos erros locais é indispensável para evitar que o otimismo das métricas agregadas globais mascare limitações práticas de detecção em campo.

No Proposto, embora os trabalhos citados ao longo desta monografia fundamentem as escolhas metodológicas desta pesquisa, a comparação direta de desempenho quantitativo exige cautela devido à grande heterogeneidade entre as bases de dados. Enquanto trabalhos de referência na literatura operam em contextos de abundância de dados ou em tarefas distintas, a presente pesquisa enfrenta o desbalanceamento severo típico de espécies brasileiras ameaçadas, priorizando o F1-Score Macro para garantir uma avaliação justa e equilibrada das classes raras.

Benchmarks globais, como os apresentados por Martinsson (2017), utilizam competições internacionais, a exemplo do BirdCLEF 2016, com 999 espécies em gravações ruidosas de campo. O referido autor obteve uma acurácia de 60,3% e uma área sob a curva (AUC) de 96,5% utilizando uma Rede Neural Convolucional baseada na arquitetura ResNet-18. Em termos comparativos, sob a janela temporal de 20 segundos no Cenário B, a RNC desenvolvida nesta pesquisa atingiu uma acurácia de 82,0% e um F1-Macro de 77,0%. Embora a comparação direta de acurácia seja limitada pela diferença no número de classes (136 neste trabalho contra 999 da literatura), o resultado alcançado a partir de dados reais de ciência cidadã sem curadoria profissional é encorajador. Esse desempenho sugere que o pipeline de limpeza espectral e o aumento de dados direcionado compensaram de maneira eficiente o suporte amostral reduzido de treinamento.

Quando posicionados frente ao trabalho de Stowell et al. (2019), que aborda o problema de detecção acústica binária de aves (presença ou ausência) e reporta um desempenho de 88,7% de AUC utilizando uma RNC, os resultados deste trabalho mostram-se competitivos e promissores. A Máquina de Vetores de Suporte (MVS) desenvolvida neste estudo, operando com características estatísticas compactas extraídas de coeficientes MFCC, alcançou uma acurácia de 76,1% e F1-Macro de 71,9% na janela de 20 segundos. Esse desempenho indica que, mesmo em um cenário desafiador de classificação multiclasse envolvendo mais de uma centena de

táxons ameaçados, um classificador clássico bem parametrizado consegue estabelecer fronteiras de decisão robustas com baixíssimo custo computacional.

Por sua vez, o trabalho recente de Chaudhuri et al. (2024) descreve o sistema ASGIR, que combina uma rede de atenção do tipo Audio Spectrogram Transformer (AST) com RNC para obter uma acurácia de 98,4% e F1 de 97,3% em um subconjunto de 51 espécies comuns de aves europeias com dados abundantes. Embora os índices absolutos do sistema ASGIR sejam superiores, destaca-se que modelos baseados em Transformers sofrem severa perda de generalização e degradação de desempenho em cenários de extrema escassez de dados (few-shot). Sob essa ótica, a acurácia de 82,0% obtida pela RNC proposta e de 66,3% (com F1-Macro de 64,2%) registrada pelo Modelo Oculto de Markov (MOM) sob modelagem probabilística sequencial na janela de 20 segundos reforçam a utilidade prática e a resiliência do pipeline desenvolvido para o monitoramento de avifauna ameaçada sob restrições severas de dados reais de treinamento.

Assim, o objetivo desta análise comparativa textual não é alegar a superioridade absoluta da metodologia proposta sobre os algoritmos do estado da arte, mas sim demonstrar que o processamento rigoroso de sinais e de dados permite construir classificadores de alta utilidade prática, tanto clássicos quanto de aprendizado profundo, mesmo diante das limitações físicas impostas por gravações de campo de espécies raras.

Em relação aos classificadores clássicos, a MVS desenvolvida neste estudo alcançou 76,1% de acurácia e 71,9% de F1-Macro. Esse desempenho demonstra que, com uma engenharia de atributos estatísticos robusta baseada em coeficientes MFCC, modelos clássicos conseguem estabelecer fronteiras de decisão eficientes em cenários complexos de mais de uma centena de classes. Esse resultado é relevante quando posicionado frente ao trabalho de detecção binária de Stowell (2019), o qual atingiu AUC de 88,7%, indicando que a MVS estatística se estabelece como uma alternativa viável, robusta e de baixíssimo custo computacional para o monitoramento bioacústico.

De forma análoga, o MOM (GMM-HMM) obteve acurácia de 66,3% e F1-Macro de 64,2%. Embora esse índice probabilístico seja inferior aos modelos convolucionais e espaciais de aprendizado profundo, o classificador sequencial provou-se altamente resiliente ao tamanho das janelas de áudio nos testes. A modelagem probabilística das sequências de estados preserva a capacidade discriminativa e captura a sintaxe temporal básica das vocalizações mesmo sem o poder de extração automática de características das convoluções.

Para o cálculo estatístico do desempenho geral do melhor classificador (RNC na janela de 20 segundos), a concordância corrigida ao acaso foi avaliada por meio do coeficiente Kappa de Cohen, expresso formalmente de acordo com a Equação 1.

$$k = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

Onde a variável de concordância observada corresponde à acurácia global do modelo, equivalente a 0,8200, e a concordância esperada por acaso é calculada de forma proporcional às margens da distribuição das classes. A aplicação da Equação 1 resultou em um coeficiente Kappa consolidado de 0,8182. De acordo com a escala de interpretação de Landis e Koch (1977), esse valor posiciona a força de concordância do classificador proposto no patamar de Quase Perfeita, o que atesta a robustez estatística do pipeline de processamento para a classificação multiclasse de avifauna brasileira ameaçada sob condições de escassez de dados reais.

CONCLUSÃO

O objetivo de desenvolver e avaliar um pipeline de classificação bioacústica para espécies brasileiras ameaçadas de extinção foi parcialmente atingido, demonstrando resultados promissores para um subconjunto das espécies consideradas e estabelecendo diretrizes metodológicas relevantes para trabalhos futuros

Os experimentos evidenciaram uma melhoria expressiva no desempenho dos classificadores após a aplicação do pipeline de processamento proposto

No entanto, como o desenho experimental combinou simultaneamente a limpeza espectral, as técnicas de aumento de dados e a curadoria do conjunto de espécies, a contribuição isolada de cada componente permanece como uma questão em aberto, a ser investigada detalhadamente por meio de estudos de ablação sugeridos para etapas posteriores

Entre os três paradigmas avaliados, a Rede Neural Convolucional (RNC) apresentou o melhor desempenho global, revelando maior capacidade de extrair características discriminativas diretamente dos espectrogramas Log-Mel e de lidar com a variabilidade natural das vocalizações quando comparada aos modelos clássicos baseados em atributos extraídos manualmente

A análise temporal apontou que as janelas intermediárias de 10 e 20 segundos configuram o intervalo ideal para a classificação, registrando um F1-Macro de 0,77

Os resultados sugerem que a aplicação da limpeza espectral contribuiu para reduzir a sensibilidade do modelo convolucional em relação à duração do segmento, permitindo que a rede mantivesse um desempenho estável mesmo em janelas de áudio longas, de até 60 segundos, hipótese que requer verificação experimental controlada no futuro

Desse modo, a hipótese inicial estabelecida neste trabalho foi sustentada, demonstrando que a priorização da qualidade física do sinal é um fator determinante para mitigar os efeitos da escassez de dados de treino na cauda longa do dataset

Como limitação prática do estudo, destaca-se o compromisso imposto pelo filtro de qualidade espectral, uma vez que espécies raras com vocalizações muito tênues ou gravadas em ambientes com ruído físico extremo precisaram ser excluídas do cenário proposto para garantir a integridade do treinamento

Essa restrição limita a aplicação do pipeline atual apenas a arquivos que alcancem um patamar de qualidade mínima aceitável

Como direcionamento para trabalhos futuros, recomenda-se a realização de estudos de ablação mais sistemáticos para quantificar o ganho marginal de cada etapa do processamento, além da validação das ferramentas com gravações de campo completamente independentes para testar a capacidade de generalização geográfica

Por fim, incentiva-se o desenvolvimento de arquiteturas híbridas baseadas na fusão RNC-MOM, utilizando a rede convolucional como extratora automática de características de áudio e os Modelos Ocultos de Markov para realizar a modelagem da sequência temporal e da sintaxe do canto, integrando as vantagens de ambos os paradigmas de aprendizado

REFERÊNCIAS

BRASIL. Ministério do Meio Ambiente. Portaria MMA nº 148, de 7 de junho de 2022. Diário Oficial da União, p. 74, 2022

CHAUDHURI, Y. et al. ASGIR: audio spectrogram transformer guided classification and information retrieval for birds. Interspeech 2024, p. 3636–3637, 2024

COHEN, J. A coefficient of agreement for nominal scales. Educational and psychological measurement, v. 20, p. 37–46, 1960

DAVIS, S.; MERMELSTEIN, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE transactions on acoustics, speech, and signal processing, v. 28, p. 357–366, 1980

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep Learning. MIT Press, Cambridge, 2016

KAHL, S. et al. BirdNET: A deep learning solution for avian diversity monitoring. Ecological Informatics, v. 61, p. 101236, 2021

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. Biometrics, v. 33, p. 159–174, 1977

MARTINSSON, J. Bird Species Identification using Convolutional Neural Networks. Chalmers University of Technology, Gothenburg, 2017

PLANQUÉ, B. Xeno-canto: A global database of bird sounds. 2013

PURWINS, H. et al. Deep learning for audio signal processing. IEEE Journal of Selected Topics in Signal Processing, v. 13, p. 206–219, 2019

RABINER, L. R.; JUANG, B.-H. Fundamentals of speech recognition. Prentice-Hall, Inc., Upper Saddle River, 1993

ROSENBERG, K. V. et al. Decline of the north american avifauna. Science, v. 366, p. 120–124, 2019

SALAMON, J.; BELLO, J. P. Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal Processing Letters, v. 24, p. 279–283, 2017

STOWELL, D. et al. Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge. Methods in Ecology and Evolution, v. 10, p. 368–380, 2019