Uso de redes neurais convolucionais para a identificação de estresse em vocalizações de gado bovino

Bryan Teixeira Paiva[1][0000−0003−0031−9970], Ana Paula Lüdtke Ferreira[1][0000−0001−7057−9095], and Naylor Bastiani Perez[1,2][0000−0002−4667−783X]

[1] Programa de Pós-graduação em Computação Aplicada
Universidade Federal do Pampa, Bagé-RS, Brasil
{bryanpaiva.aluno,anaferreira}@unipampa.edu.br
[2] Empresa Brasileira de Pesquisa Agropecuária, Bagé-RS, Brasil
naylor.perez@embrapa.br

Resumo O controle do estresse animal nos sistemas de produção de gado de corte assegura o bem-estar animal, a qualidade da carne e os ganhos financeiros em toda a cadeia produtiva. As diversas etapas da produção que envolvem a interação humano-animal fazem com que os animais sejam submetidos a situações de estresse, causando aumento do pH da carne e agitação que pode causar hematomas e ferimentos, tornando a carne imprópria para consumo humano. Este artigo objetiva avaliar o uso de três arquiteturas de Redes Neurais Convolucionais para identificação de estresse em vocalizações de bovino. A implementação consiste em uma arquitetura básica, uma intermediária e uma robusta, utilizando Coeficientes Cepstrais de Frequência-MEL para extração de características acústicas. Os resultados mostram que as arquiteturas básica, intermediária e robusta alcançaram *F1-score* de 96,97%, 97,90% e 98,74%, respectivamente. A análise estatística mostrou uma diferença significativa entre as arquiteturas básica e robusta.

Keywords: Análise acústica · Aprendizado de máquina · Bem-estar animal

1 Introdução

A pecuária de corte bovino figura como uma das principais atividades econômicas no Brasil, com uma produção estimada em 213 milhões de cabeças de gado, contribuindo com cerca de 6% do Produto Interno Bruto (PIB) nacional [7] e fazendo com que o país seja um dos principais produtores, consumidores e exportadores de carne a nível global. Mesmo com esse volume de produção, o setor de bovinocultura de corte busca maior produtividade, com foco principalmente no aumento na qualidade da carne produzida, uma vez que a exportação para mercados de maior valor agregado exige padrões rígidos quanto aos procedimentos adotados na produção e criação dos animais [3].

O perfil dos consumidores também tem mudado nos últimos anos, com uma maior exigência quanto à qualidade da carne produzida. Além dos aspectos específicos de qualidade da carne como aspecto, sabor, textura, maciez e aroma, um crescente número de consumidores passou a se preocupar com a forma de criação e tratamento dos animais. Mercados específicos, como o de países da União Europeia, consideram pagar um maior valor em carnes quando há a garantia de bem-estar animal durante a produção [12].

O bem-estar animal está relacionado com os conceitos de necessidades básicas: liberdade, felicidade, adaptação, controle, sentimentos, sofrimento, dor, ansiedade, medo, tédio, estresse e saúde. As diferentes fases de produção da bovinocultura exigem o manejo dos animais, que usualmente tem um impacto negativo no bem-estar e pode ter um efeito nocivo à qualidade final da carne produzida. A não adoção de boas práticas de manejo resultam em estresse, contusões, e até mesmo em mortes de animais [1,3]. Como consequência do estresse, o pH da carne aumenta, adquirindo as propriedades DFD (escura, dura e seca), o que reduz a qualidade da carne produzida e acarreta em prejuízos financeiros para os produtores, visto que uma carne com essas características não pode ser vendida para consumo humano [9]. Essa situação demanda uma reformulação dos métodos utilizados no manejo dos animais, tanto para melhores condições de bem-estar do animal, quanto para uma maior lucratividade dos produtores. A identificação automatizada de estresse do gado é desejável para que um sistema de alerta possa ser construído, avisando as pessoas mais próximas de situações que violem o bem-estar dos animais.

O objetivo deste trabalho foi realizar um estudo comparativo sobre a capacidade de discernimento de estresse em vocalizações bovinas usando três arquiteturas de redes neurais convolucionais. O trabalho foi conduzido com base nas seguintes hipóteses de pesquisa: (i) os sons emitidos por animais estressados são diferentes daqueles emitidos por animais mais tranquilos; (ii) a análise dos sons emitidos pelos animais pode ser executada por redes neurais para verificação de situações de estresse.

O restante do texto está estruturado como se segue: a Seção 1 apresenta uma contextualização sobre o problema de pesquisa e os objetivos propostos. A Seção 2 traz as ferramentas, tecnologias e procedimentos empregados o desenvolvimento do trabalho, incluindo a coleta de dados, a preparação da base de dados e a classificação de vocalizações. A Seção 3 presenta os principais resultados da pesquisa, incluindo teste de hipóteses do desempenho das arquiteturas implementadas e a comparação dos resultados com os trabalhos correlatos encontrados na literatura. A Seção 4 apresenta a conclusão do trabalho, com indicação de trabalhos futuros.

2 Material e métodos

A metodologia dividiu o trabalho em cinco etapas, a saber: (i) revisão de escopo da literatura sobre uso de redes neurais convolucionais para identificação de vocalizações de animais; (ii) coleta de dados de vocalizações em condições

de estresse e não estresse; (iii) tratamento dos dados e preparação da base de vocalizações; (iv) processamento dos dados e treinamento de redes neurais; (v) análise dos resultados.

2.1 Coleta de dados

A coleta de dados de vocalizações de bovinos foi realizada com 48 animais na raça Brangus (*Bos taurus indicus*). As vocalizações foram registradas em dois contextos psicologicamente distintos: durante o confinamento e durante o manejo. Esses ambientes representam diferentes momentos nos quais os animais podem apresentar diferentes níveis de estresse.

A coleta de vocalizações no ambiente de confinamento foi conduzida na zona rural de Bagé-RS (31°18'56"S 53°59'54"W), com espaço total de cerca de 1500m^2, dividido em potreiros medindo 30m x 25m. Nesse espaço, os animais desfrutavam de liberdade para socializar e se alimentar, com interações humanas limitadas ao fornecimento de alimentação, garantindo assim um ambiente controlado. Os sons produzidos no ambiente de confinamento foram coletados por três câmeras instaladas nas extremidades direta e esquerda e no centro do ambiente. Para uma melhor qualidade na coleta das vocalizações, as câmeras foram posicionadas em proximidade aos cochos de alimentação, áreas onde os animais passavam a maior parte do tempo durante o confinamento. A disposição das câmeras durante esse processo é ilustrada na Figura 1.

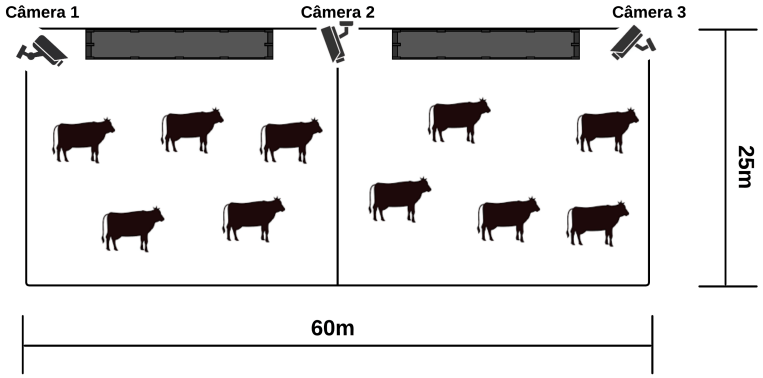

Figura 1. Esquema de posicionamento das câmeras

As imagens e sons dos animais no ambiente de confinamento foram coletadas durante duas semanas, resultando em cerca de 1000 horas de filmagem. Devido ao contexto de tranquilidade dos animais e à ausência de interações estressantes, as vocalizações capturadas foram consideradas como indicativas de um estado não estressado ou normal. A Figura 2 retrata os animais nesse contexto de confinamento.

Figura 2. Animais em confinamento

Para a coleta das vocalizações durante o manejo, foram registradas duas sessões de pesagem dos animais, cada uma com duração média de uma hora. Durante essas sessões, os animais eram conduzidos para um curral estreito que os levava até a balança de pesagem, onde eram contidos dentro de uma gaiola até a leitura do peso ser feita. No decorrer desse processo, observou-se um manejo mais vigoroso dos animais, caracterizado por interações intensas que envolviam gritos, gestos e cutucões ao longo do trajeto até a balança. A Figura 3 apresenta o local de manejo dos animais.

Figura 3. Espaço de manejo dos animais (gaiola de pesagem)

O monitoramento e o registro desse processo foram conduzidos utilizando câmeras digitais. Durante o manejo, especialistas em comportamento animal observaram agitação e aumento nos níveis de estresse dos animais. Houve também uma maior frequência na produção de vocalizações em comparação com os períodos em que os animais estavam no ambiente de confinamento.

2.2 Preparação da base de dados

A base de dados das vocalizações foi construída a partir das filmagens e gravações feitas durante a fase de coleta de dados. Os registros de confinamento e manejo foram analisados e rotulados com o intuito de extrair as características acústicas das vocalizações.

As vocalizações no ambiente de confinamento foram analisadas e organizadas com um software desenvolvido um *software* em linguagem Python (https://www.python.org/). Como as vocalizações de bovinos geralmente têm entre 1,3 e 2,1 segundos de duração [19], o software desenvolvido detecta picos de amplitude sonora em janelas de 3 segundos. Esses picos indicam a presença de sons discrepantes e, se identificados, os trechos correspondentes a essas janelas de tempo são armazenados separadamente para análises manuais posteriores. A análise manual foi conduzida com o software Movavi Video Editor (https://www.movavi.com/pt/videoeditor/). O processo descartou os fragmentos que não correspondiam a vocalizações de bovinos ou cujos sons apresentavam qualidade insatisfatória devido à baixa amplitude ou sobreposição de sons. Posteriormente, os sons foram extraídos dos arquivos de vídeo resultantes, no formato wav (*Waveform Audio Format*). As cerca de 1000 horas de filmagens registradas durante o confinamento resultaram na extração de 357 vocalizações individuais dos animais. Considerando o contexto de controle, esses sons foram classificados como vocalizações normais.

A análise das vocalizações de manejo verificou os registros de vídeo produzidos, sendo feita integralmente de maneira manual, também com o software Movavi Video Editor, tendo sido possível identificar um total de 186 vocalizações individuais emitidas pelos animais. Devido à natureza desse contexto, que envolveu interações intensas entre animais e humanos, esses sons foram classificados como vocalizações de estresse.

Para manter a consistência dos dados, as vocalizações com duração inferior a 3 segundos foram ajustadas para atender a esse padrão, garantindo que todos os arquivos de vocalizações tivessem o mesmo comprimento padrão de 3 segundos, o que também foi aplicado às vocalizações capturadas no ambiente de confinamento. Para assegurar a qualidade das amostras de áudio, todas as vocalizações coletadas, tanto em ambiente de confinamento quanto durante o manejo, passaram por um processo de filtragem digital para redução de ruídos. Essa etapa de filtragem foi importante para garantir a clareza e a precisão das vocalizações registradas, possibilitando uma análise mais precisa dos dados acústicos.

2.3 Classificação de vocalizações

A tarefa de classificação e identificação de som animal é parte central do trabalho desenvolvido. As principais tarefas envolvidas no reconhecimento de som são: (i) construir uma base de dados confiável contendo um repertório de vocalizações rotuladas de animais; (ii) calcular, ou definir apropriadamente as características para representar, ou classificar as vocalizações; (iii) comparar as vocalizações desconhecidas com os padrões que são conhecidos para encontrar a combinação correta e identificar o significado da vocalização [8].

O módulo de identificação de estresse animal foi implementado com softwares desenvolvidos na linguagem Python e com a utilização das bibliotecas Librosa (https://librosa.org/), Scikit-Learn (https://scikit-learn.org/), TensorFlow (https://www.tensorflow.org/) e Keras (https://keras.io/).

Com base na revisão da literatura, as técnicas escolhidas para a análise e classificação de som foram o MFCC (*Mel-frequency cepstral coefficients*) e redes neurais convolucionais. A viabilidade da utilização dessas técnicas para a identificação de sons animais foi evidenciada em trabalhos na literatura [2,5,8,11,10,14].

A construção do módulo de classificação foi composta por duas etapas, sendo a primeira responsável pela extração das features de MFCC dos arquivos de som e a segunda na criação dos modelos de identificação e classificação de som, a partir do treinamento de diferentes redes neurais. A Figura 4 mostra as duas etapas do módulo de classificação. A extração das características acústicas das vocalizações é feita na primeira etapa do módulo de classificação de vocalizações e faz uso do MFCC. Essa técnica é uma das mais populares para extração de *features* de som para a realização de reconhecimento de voz no domínio da frequência [4].

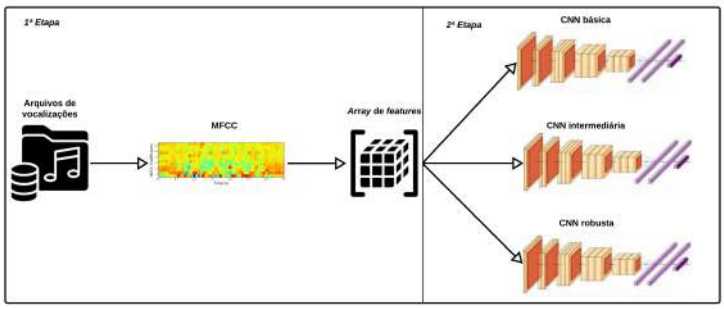

Figura 4. Módulo de classificação

O MFCC é a representação do *cepstrum* real de um sinal janelado em tempo curto derivado da transformada rápida de Fourier (FFT), em escala de frequências não lineares, denominada escala Mel. A utilização da escala Mel visa simular o comportamento do sistema auditivo humano [18]. A extração das *features* de MFCC é realizada a partir dos seguintes passos: *(i) pre emphasis*; *(ii) framing*

and windowing; *(iii)* FFT/DFT; *(iv) mel filter bank*; *(v)* IFFT/DCT. As características MFCC são amplamente utilizadas em sistemas de reconhecimento de fala, identificação de locutor, processamento de linguagem natural e sistemas de controle de voz [18]. Devido à sua capacidade de capturar características discriminativas da fala, os MFCC são uma escolha popular em uma variedade de aplicações relacionadas ao áudio, sendo comumente usadas como entrada para modelos de aprendizado de máquina, como redes neurais artificiais.

Cada vocalização coletada foi submetida individualmente ao algoritmo MFCC para a extração de suas características. As características e a classe do som foram armazenadas em um *array* dinâmico, que foi preenchido à medida em que as características de MFCC eram extraídas dos arquivos de vocalizações. Para implementar o extrator de características MFCC, foi empregada a biblioteca Librosa, que permitiu a extração de características a partir de um arquivo de áudio e da definição da frequência de amostragem e do número desejado de coeficientes de MFCC. No software desenvolvido, a frequência de amostragem foi definida como 44,1 kHz, o número de coeficientes MFCC como 13, o tamanho da janela para a transformada de Fourier foi configurado como 2048, além do espaçamento entre amostras que foi estabelecido em 512. A definição dessas configurações foram baseadas nas características dos sinais de áudio, na convenção comum de uso desses parâmetros com a biblioteca Librosa, e também em estudos relevantes na literatura [16,21,18,10].

Para a aplicação do extrator desenvolvido, foram selecionadas 200 das 357 vocalizações de confinamento, enquanto todas as 186 vocalizações de manejo foram utilizadas. Essa definição foi feita para manter um equilíbrio entre as classes de estresse e normal, evitando que os classificadores fossem treinados com classes desbalanceadas, o que poderia prejudicar a etapa de treinamento, em que os modelos podem acabar favorecendo a classe majoritária e não aprendendo adequadamente a distinguir a classe minoritária.

Sobre as características MFCC extraídas, foi aplicada a normalização conhecida como *z-score*. Essa técnica de normalização é utilizada para padronizar os dados e facilitar a comparação entre diferentes características. O *z-score* transforma os valores originais de cada característica, subtraindo a média e dividindo pelo desvio padrão. A Equação 1 apresenta o cálculo do *z-score*, onde x é o valor individual, μ é a média das amostras e σ é o desvio padrão das amostras.

$$z - score = \frac{x - \mu}{\sigma} \tag{1}$$

A normalização dos dados evita que características com escalas diferentes influenciem negativamente no treinamento dos modelos, contribuindo para a estabilidade e convergência eficiente do algoritmo. Por fim, após a extração das características e tratamento dos dados, o array resultante, juntamente com as categorias correspondentes de todas as vocalizações, foram armazenadas em um *dataframe* para servirem como entrada dos modelos de classificação de estresse animal.

A fase de classificação foi feita sobre três arquiteturas distintas de CNN, cada uma representando um nível diferente de complexidade: uma estrutura

básica, uma intermediária e uma versão mais robusta. Essa abordagem teve como objetivo examinar o desempenho das redes sob várias perspectivas, questionando se o aumento ou a redução da complexidade dessas redes teriam um impacto significativo na eficácia de classificação das vocalizações indicativas de estresse em bovinos.

As redes neurais convolucionais (CNN) emergiram do estudo do córtex visual do cérebro e têm sido utilizadas no reconhecimento de imagens desde os anos 1980. As CNN não estão restritas à percepção visual, sendo também bem-sucedidas em outras tarefas, como reconhecimento de voz ou processamento de linguagem natural [6]. A arquitetura típica de uma CNN consiste em uma camada de entrada, camadas convolucionais, camadas de agrupamento (ou *pooling*) e camadas totalmente conectadas [17]. Enquanto as camadas convolucionais são utilizadas para aprender as características de baixo nível, as camadas de *pooling* são utilizadas para reduzir o tamanho espacial das características convolucionais, diminuindo assim o custo computacional [17]. Por fim, em arquiteturas CNN, a etapa de classificação dos dados é geralmente realizada por camadas totalmente conectadas [6].

Para a realização do estudo de arquiteturas CNN para a tarefa de classificação de vocalizações de estresse em bovinos, foram propostas as seguintes configurações:

1. Arquitetura Básica:
   - Número de camadas: 5 (1 camada convolucional, 1 camada de *max pooling*, 1 camada *flatten*, 1 camada totalmente conectada, 1 camada de saída)
   - Camada convolucional: 64 filtros de tamanho 3 x 3, ativação leaky ReLU
   - *Pooling*: Max *pooling* de tamanho 2 x 2
   - Camada totalmente conectada: 32 neurônios, ativação leaky ReLU
   - Camada de saída: 1 neurônio, ativação sigmoid
2. Arquitetura Intermediária
   - Número de camadas: 7 (2 camadas convolucionais, 2 camadas de *max pooling*, 1 camada *flatten*, 1 camada totalmente conectada, 1 camada de saída)
   - Camadas convolucionais: 2 camadas, 64 filtros de tamanho 3 x 3 e 32 filtros de tamanho 3 x 3, ativação leaky ReLU
   - *Pooling*: Max *pooling* de tamanho 2 x 2 após cada convolução
   - Camada totalmente conectada: 64 neurônios, ativação leaky ReLU
   - Camada de saída: 1 neurônio, ativação sigmoid
3. Arquitetura Robusta
   - Número de camadas: 10 (3 camadas convolucionais, 3 camadas de *max pooling*, 1 camada *flatten*, 2 camada totalmente conectada, 1 camada de saída)
   - Camadas convolucionais: 3 camadas, 256 filtros de tamanho 3 x 3, 128 filtros de tamanho 3 x 3 e 64 filtros de tamanho 3 x3, ativação leaky ReLU
   - *Pooling*: Max *pooling* de tamanho 2 x 2 após cada convolução

- Camadas totalmente conectadas: 2 camadas, 128 neurônios e 64 neurônios, ativação leaky ReLU
- Camada de saída: 1 neurônio, ativação sigmoid

A Arquitetura Básica consiste em uma estrutura simples de rede convolucional, composta por uma única camada convolucional seguida por uma camada de *pooling* e uma camada totalmente conectada. Seu objetivo principal é oferecer uma abordagem direta para classificar as vocalizações de estresse em bovinos, focando em capturar características fundamentais das vocalizações. Suas vantagens incluem simplicidade e eficiência computacional, tornando-a fácil de entender e rápida de treinar. No entanto, sua capacidade de aprendizado pode ser limitada devido à falta de camadas adicionais para extrair representações mais complexas.

A Arquitetura Intermediária, por sua vez, foi projetada para superar as limitações da arquitetura básica, incorporando camadas convolucionais adicionais. Com duas camadas convolucionais e duas camadas de *pooling*. Esta arquitetura visa melhorar a capacidade de representação da rede, permitindo que aprenda características mais abstratas das vocalizações. Suas vantagens incluem uma maior capacidade de aprendizado e generalização devido à inclusão de camadas adicionais.

Por fim, a Arquitetura Robusta é a mais complexa das três, apresentando um maior número de camadas convolucionais e totalmente conectadas. Com três camadas convolucionais, três camadas de pooling e duas camadas totalmente conectadas, esta arquitetura visa capturar representações ainda mais detalhadas e abstratas das vocalizações. Seus pontos fortes incluem uma capacidade de aprendizado superior e uma melhor capacidade de generalização devido à sua profundidade e complexidade. No entanto, sua maior complexidade também pode tornar o treinamento mais demorado e exigir recursos computacionais, além de aumentar o risco de *overfitting*.

Na configuração dos parâmetros para o treinamento de redes neurais, uma série de escolhas precisam ser feitas para otimizar o desempenho do modelo. Esses parâmetros incluem a escolha do otimizador, a taxa de aprendizagem, o método de inicialização de pesos, número de épocas, entre outros. A seleção adequada desses parâmetros é crucial, pois pode influenciar significativamente a convergência do modelo, sua capacidade de generalização e a eficácia na resolução do problema em questão. Este processo de ajuste fino dos parâmetros visa encontrar a combinação mais adequada que maximize o desempenho da rede neural para a tarefa específica em análise. Dessa forma, no treinamento das redes neurais foram definidos os seguintes parâmetros:

- Otimizador: Adam
- Inicialização de pesos: Glorot
- Taxa de aprendizagem: 0,01
- Épocas: 200
- *Batch size*: 32

Em todas as arquiteturas implementadas, foram empregadas técnicas de regularização, como *dropout* e *batch normalization*, com o objetivo de mitigar o

sobreajuste dos modelos. A taxa de aprendizagem foi fixada em 0,01 e ajustada dinamicamente durante o treinamento, sendo reduzida a taxa pela metade (fator = 0,5) caso não houvesse melhorias no treinamento após 10 épocas.

O treinamento das redes neurais foi realizado a partir de cinco repetições de validações cruzadas (*cross-validation*) com k = 5. Essa técnica consiste em dividir o conjunto de treinamento em subconjuntos complementares e cada modelo é treinado com uma combinação diferente desses subconjuntos e validado em relação às partes restantes [6]. A ideia por trás da validação cruzada é que se o conjunto de teste for sempre o mesmo, pode ocorrer um superajuste do modelo a esse conjunto de teste, o que significa que o modelo pode estar ajustando a análise a um conjunto de dados específico a ponto de não conseguir analisar adequadamente um conjunto diferente. Assim, ao variar o conjunto de teste, evitamos o sobreajuste, garantindo uma análise mais robusta e generalizável para diferentes conjuntos de dados [17].

Vale ressaltar foram exploradas e testadas outras configurações para o treinamento das redes neurais, incluindo os otimizadores Nadam, SGD, RMSprop e AdaGrad. Além disso, foram avaliadas inicializações de pesos uniformes e normais, bem como alterações nos valores de *batch size*. No entanto, as variações nesses parâmetros não proporcionaram melhorias significativas no desempenho durante o treinamento das redes. Portanto, optou-se por adotar uma configuração de treinamento padrão para todas as redes.

3 Resultados

A análise dos resultados obtidos no treinamento das diferentes arquiteturas de redes CNN foi feita a partir das métricas de acurácia, precisão, revocação e *F1-score*. A Tabela 1 apresenta os valores médios obtidos para cada uma das métricas das três arquiteturas implementadas após a execução do treinamento por repetições de validação cruzada. Nessa tabela, pode-se observar que todas as arquiteturas alcançaram um bom desempenho na classificação das vocalizações. A arquitetura mais robusta apresentou desempenho superior em comparação com as de menor complexidade. No entanto, apesar de sua maior complexidade, evidenciada pelo aumento no número de camadas e de filtros por camada, os ganhos em acurácia, precisão, revocação e *F1-score* foram modestos. Isso sugere que os recursos adicionais de computação investidos não se traduziram em melhorias proporcionais na capacidade de classificação de vocalizações.

Tabela 1. Médias para as métricas de acurácia, precisão, revocação e *F1-score*

Architecture	**Accuracy**	**Precision**	**Recall**	***F1-score***
Basic	96,92%	95,37%	98,63%	96,97%
Intermediate	97,88%	96,87%	98,95%	97,90%
Robust	98,74%	98,37%	99,11%	98,74%

Também observa-se que as arquiteturas CNN apresentaram melhores desempenhos na classificação das vocalizações de estresse em comparação com as vocalizações normais, evidenciado pelos maiores valores de revocação. Esses resultados ressaltam a eficácia das arquiteturas CNN na tarefa de classificação de vocalizações animais, comprovando seu potencial para aplicações práticas em estudos comportamentais e de bem-estar animal.

Para avaliar o desempenho das arquiteturas implementadas e determinar se as diferenças observadas nos resultados entre as arquiteturas foram devidas às variações nas complexidades das arquiteturas ou apenas ao acaso, foi realizada uma análise de variância (ANOVA).

O teste de hipóteses foi conduzido utilizando o *F1-score* como a métrica de análise, devido à sua robustez em comparação com a acurácia, precisão e revocação. Foram comparadas as médias do *F1-score* entre as arquiteturas CNN. As hipóteses formuladas foram as seguintes:

- Hipótese nula (H_0): Não há diferença significativa entre as médias do *F1-score* entre as arquiteturas. Qualquer variação observada é atribuída ao acaso.
- Hipótese alternativa (H_A): Existe uma diferença estatisticamente significativa entre as médias do *F1-score* entre as arquiteturas. A variação observada não é devida ao acaso, indicando uma relação genuína entre a complexidade das redes e o desempenho na classificação de vocalizações normais e de estresse.

A Tabela 2 apresenta os resultados da ANOVA, para um nível de confiança de 5%, para os três modelos de arquiteturas CNN implementadas.

Tabela 2. Análise de variância por testes de Tukey entre os modelos de CNN

Architecture vs Architecture	**Mean F1-score**		**Diference**	***P-value***
Basic CNN vs Intermediate CNN	96,97	97,90	0,93	0,158
Basic CNN vs Robust CNN	96,97	98,74	1,77	0,002**
Intermediate CNN vs Robust CNN	97,90	98,74	0,84	0,22

A análise comparativa entre os modelos de redes neurais revelou diferenças significativas em relação ao desempenho das arquiteturas, conforme evidenciado pelos testes de Tukey. Os resultados apontam uma melhora estatisticamente significativa entre as arquiteturas básica e robusta (*P-value* = 0,002), contudo entre as arquiteturas básica e intermediária, e intermediária e robusta não há evidências estatísticas de que o aumento de complexidade refletiu em melhores resultados. Esses resultados destacam a influência da arquitetura e da complexidade do modelo no desempenho das redes neurais, sugerindo que, em determinados contextos, aumentar a complexidade do modelo pode resultar em melhorias significativas na capacidade de generalização e aprendizado. No entanto, para outras arquiteturas, o aumento de complexidade pode não resultar em maior poder de classificação [13].

Realizando um comparativo com os resultados encontrados na literatura, é possível destacar a utilização de redes neurais do tipo CNN para classificação de vocalizações animais em diferentes trabalhos [14,16,15,20,10]. A Tabela 3 apresenta um comparativo entre os resultados obtidos por este trabalho e a literatura.

Tabela 3. Comparação entre os resultados obtidos e a literatura

Work	**Characteristic**	**Results**
Present work	MFCC	Accuracy = 98,74% Precision = 98,37% Recall = 99,11% F1-score = 98,74%
Sattar (2022)	MFCC	Accuracy = 84%
Jung et al. (2021)	MFCC	Accuracy = 94,18%
Sasmaz e Tek (2018)	MFCC	Accuracy = 75%
Pandeya et al. (2022)	Spectrogram	F1-score = 70,90%
Shorten (2023)	Spectrogram	Accuracy = 96,2%
Vidana-Vila et al. (2023)	Spectrogram	F1-score = 61,7%
Gavojdian et al. (2023)	23 vocal parameters	F1-score = 89,4%

Ao analisar a Tabela 3, destaca-se que a maioria dos estudos encontrados na literatura empregou os coeficientes MFCC como características de análise acústica e também observa-se o uso do espectrograma como uma característica de interesse para o modelo de classificação, como também o uso de parâmetros vocais nos domínios da frequência e amplitude. Ainda que os resultados numéricos não possam ser comparados, por diferenças nos dados e métodos usados, além de diferenças entre as raças de bovinos analisados, os resultados obtidos neste trabalho são consistentes com estudos anteriores que adotaram abordagens semelhantes. Os resultados alcançados neste trabalho não foram testados com bases independentes porque não conseguimos encontrar dados abertos rotulados na forma deste trabalho. Ainda assim, os resultados obtidos são promissores.

Os resultados deste trabalho revelaram a viabilidade em empregar as características extraídas do MFCC como base para o treinamento de redes neurais na identificação de estresse em bovinos. No estudo de arquiteturas CNN para a classificação de vocalizações, todas alcançaram bons resultados, essa fato pode ser um indício para a preferência predominante na literatura pelo uso de redes CNN na análise acústica de vocalizações animais.

Entre as diferentes complexidades de arquiteturas, a variante robusta apresentou resultado estatisticamente superior à variante básica. Contudo, é importante ressaltar que o aumento na complexidade vem acompanhado de maiores exigências computacionais, como requisitos de memória e poder de processamento. As redes robustas demandaram consideravelmente mais tempo de treinamento, além de exigir maior poder de processamento em comparação com as redes menos complexas. Portanto, ao escolher uma arquitetura de rede neural é

importante avaliar os aspectos computacionais, principalmente em sistemas embarcados e de tempo real, onde as limitações de *hardware* podem ser restritivas para o uso eficaz de redes neurais mais robustas.

4 Conclusão

O estresse bovino é um dos principais fatores causadores de perdas na qualidade final da carne. Situações estressantes podem desencadear reações fisiológicas que resultam em carne escura e dura, reduzindo o valor agregado do produto e provocando prejuízos ao setor pecuário. Os estudos sobre bem-estar animal têm se expandido, buscando alternativas para garantir uma melhor qualidade de vida para os animais durante toda a sua criação. Nesse contexto, este trabalho buscou coletar e avaliar vocalizações de bovinos em duas condições psicologicamente distintas: confinamento e manejo. Em confinamento os animais estavam livres de interações estressantes, enquanto durante o manejo foi perceptível o estresse dos animais. As vocalizações foram tratadas e filtradas, a técnica de Mel Frequency Cepstrum Coefficients (MFCC) foi empregada para a extrair suas características, e redes CNN foram utilizadas para a tarefa de classificação de vocalizações.

Os resultados obtidos alcançaram acurácias variando de 96,92% a 98,74%, atestou-se a eficácia do MFCC em capturar as características essenciais das vocalizações, possibilitando a distinção entre momentos de estresse e não estresse. Os resultados da pesquisa também confirmaram as hipóteses levantadas de que os sons emitidos por animais estressados são distintos dos emitidos por animais tranquilos, e de que redes neurais artificiais são capazes de discernir situações de estresse e não estresse. Os resultados podem contribuir com o avanço do conhecimento sobre as melhores redes e configurações para a construção de classificadores de vocalizações bovinas, agregando ao estado da arte nesse campo de estudo, além de consolidar o avanço no desenvolvimento de métodos não invasivos para monitoramento do bem-estar animal na indústria pecuária.

Como trabalhos futuros, espera-se analisar outras arquiteturas de redes neurais convolucionais, buscando aprofundar o entendimento das diferenças entre vocalizações de diferentes raças bovinas e produzindo informações sobre quais são os animais menos sujeitos ao estresse, contribuindo para sistemas de melhoramento animal. Nossas bases de dados também serão tornadas públicas, dentro dos princípios FAIR de ciência aberta.

Referências

1. Camargo, M.S., Ferreira, A.P.L., Perez, N.B.: Identificação de variáveis de relevância no índice de contusões associadas ao transporte de gado de corte. In: Anales de CAI 2018 - Congreso de AgroInformática. pp. 257–265. Sociedad Argentina de Informática, Buenos Aires (2018)
2. Chung, Y., Lee, J., Oh, S., Park, D., Chang, H., Kim, S.: Automatic detection of cow's oestrus in audio surveillance system. Asian-Australasian Journal of Animal Sciences **26**(7), 1030 (2013)

3. Costa, P., Jr, M.: Ambiência na produção de bovinos de corte a pasto. Anais de Etologia **18**, 26–42 (2000)
4. Dave, N.: Feature extraction methods LPC, PLP and MFCC in speech recognition. International Journal for Advance Research in Engineering and Technology **1**(6), 1–4 (2013)
5. Deshmukh, O., Rajput, N., Singh, Y., Lathwal, S.: Vocalization patterns of dairy animals to detect animal state. In: Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012). pp. 254–257. IEEE (2012)
6. Géron, A.: Hands-on machine learning with Scikit-Learn. Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems **1** (2019)
7. IBGE: Indicadores da produção pecuária. Tech. rep., Instituto Brasileiro de Geografia e Estatística (2018)
8. Jahns, G.: Call recognition to identify cow conditions—a call-recogniser translating calls to text. Computers and Electronics in Agriculture **62**(1), 54–58 (2008)
9. Jorquera-Chavez, M., Fuentes, S., Dunshea, F.R., Jongman, E.C., Warner, R.D.: Computer vision and remote sensing to assess physiological responses of cattle to pre-slaughter stress, and its impact on beef quality: A review. Meat science **156**, 11–22 (2019)
10. Jung, D.H., Kim, N.Y., Moon, S.H., Jhin, C., Kim, H.J., Yang, J.S., Kim, H.S., Lee, T.S., Lee, J.Y., Park, S.H.: Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering. Animals **11**(2), 357 (2021)
11. Manteuffel, G., Schön, P.C.: Measuring pig welfare by automatic monitoring of stress calls. Agrartechische Berichte **29**(1) (2002)
12. Molento, C.F.M.: Bem-estar e produção animal: aspectos econômicos-revisão. Archives of Veterinary Science **10**(1) (2005)
13. Paiva, B.T.: Um estudo sobre modelos de redes neurais para identificação de estresse em vocalizações de gado bovino. Master's thesis, Programa de Pós-graduação em Computação Aplicada, Universidade Federal do Pampa (2024)
14. Pandeya, Y.R., Bhattarai, B., Afzaal, U., Kim, J.B., Lee, J.: A monophonic cow sound annotation tool using a semi-automatic method on audio/video data. Livestock Science **256**, 104811 (2022)
15. Şaşmaz, E., Tek, F.B.: Animal sound classification using a convolutional neural network. In: 2018 3rd International Conference on Computer Science and Engineering (UBMK). pp. 625–629. IEEE (2018)
16. Sattar, F.: A context-aware method-based cattle vocal classification for livestock monitoring in smart farm. Chemistry Proceedings **10**(1) (2022). https://doi.org/10.3390/IOCAG2022-12233
17. Silaparasetty, V.: Deep Learning Projects Using TensorFlow 2. Springer (2020)
18. Tiwari, V.: MFCC and its applications in speaker recognition. International journal on emerging technologies **1**(1), 19–22 (2010)
19. de la Torre, M.P., Briefer, E.F., Reader, T., McElligott, A.G.: Acoustic analysis of cattle (bos taurus) mother–offspring contact calls from a source–filter theory perspective. Applied Animal Behaviour Science **163**, 58–68 (2015). https://doi.org/https://doi.org/10.1016/j.applanim.2014.11.017
20. Vidana-Vila, E., Malé, J., Freixes, M., Solıs-Cifré, M., Jiménez, M., Larrondo, C., Guevara, R., Miranda, J., Duboc, L., Mainau, E., et al.: Automatic detection of cow vocalizations using convolutional neural networks (2023)
21. Zheng, F., Zhang, G., Song, Z.: Comparison of different implementations of MFCC. Journal of Computer science and Technology **16**, 582–589 (2001)