UNIVERSIDADE FEDERAL DO PAMPA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

BRYAN TEIXEIRA PAIVA

UM ESTUDO SOBRE MODELOS DE REDES NEURAIS PARA IDENTIFICAÇÃO DE ESTRESSE EM VOCALIZAÇÕES DE GADO BOVINO

Bagé
2024

BRYAN TEIXEIRA PAIVA

UM ESTUDO SOBRE MODELOS DE REDES NEURAIS PARA IDENTIFICAÇÃO DE ESTRESSE EM VOCALIZAÇÕES DE GADO BOVINO

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada como requisito parcial para a obtenção do título de Mestre em Computação Aplicada.

Orientadora: Ana Paula Lüdtke Ferreira

Coorientador: Naylor Bastiani Perez

Bagé
2024

Ficha catalográfica elaborada automaticamente com os dados fornecidos pelo(a) autor(a) através do Módulo de Biblioteca do Sistema GURI (Gestão Unificada de Recursos Institucionais).

B915u Paiva, Bryan Teixeira

Um estudo sobre modelos de redes neurais para identificação de estresse em vocalizações de gado bovino / Bryan Teixeira Paiva. – 2024.

125 f.: il.

Orientadora: Ana Paula Lüdtke Ferreira
Coorientador: Naylor Bastiani Perez
Dissertação (Mestrado) – Universidade Federal do Pampa, Campus Bagé, Programa de Pós-Graduação em Computação Aplicada, 2024.

1. Análise acústica. 2. Bem-estar animal. 3. Pecuária de precisão. 4. Redes neurais artificiais. I. Ana Paula Lüdtke Ferreira. II. Título.

BRYAN TEIXEIRA PAIVA

Um estudo sobre modelos de redes neurais para identificação de estresse em vocalizações de gado bovino

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada da Universidade Federal do Pampa, como requisito parcial para obtenção do Título de Mestre em Computação Aplicada.

Dissertação defendida e aprovada em: 15 de março de 2024.

Banca examinadora:

_____

Prof.ª Dr.ª Ana Paula Lüdtke Ferreira

Orientadora

(UNIPAMPA)

_____

Prof.ª Dr.ª Isabella Dias Barbosa Silveira

(UFPel)

_____

Prof. Dr. Rodrigo Schramm

(UFRGS)

Prof. Dr. Fernando Santos Osório

(USP)

Assinado eletronicamente por **ANA PAULA LUDTKE FERREIRA**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 25/04/2024, às 14:59, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **Fernando Santos Osório**, **Usuário Externo**, em 26/04/2024, às 16:24, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **Rodrigo Schramm**, **Usuário Externo**, em 06/11/2024, às 15:32, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

Assinado eletronicamente por **Isabella Barbosa**, **Usuário Externo**, em 07/11/2024, às 16:43, conforme horário oficial de Brasília, de acordo com as normativas legais aplicáveis.

A autenticidade deste documento pode ser conferida no site https://sei.unipampa.edu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1388928** e o código CRC **947F17D6**.

Dedico este trabalho à minha esposa, aos meus pais e aos meus irmãos.

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão a todas as pessoas que, de alguma maneira, contribuíram para a realização deste trabalho de pesquisa e para a conclusão bem-sucedida deste mestrado.

Em primeiro lugar, sou imensamente grato aos meus orientadores, Ana Paula e Naylor, pela orientação, apoio e paciência ao longo deste trabalho. Enfrentamos inúmeros contratempos, mas com o apoio e dedicação de vocês, conseguimos superá-los, o que foi fundamental para a conclusão deste trabalho.

Agradeço também aos membros da banca examinadora, os professores Isabella Silveira, Fernando Osório e Rodrigo Schramm, pela avaliação cuidadosa e pelas valiosas sugestões que contribuíram para a qualidade final desta dissertação.

Sou grato à Universidade Federal do Pampa pela oportunidade de ensino gratuito e de qualidade, ao PPGCAP e a todos os professores do programa pelos valiosos ensinamentos e aprendizados durante o curso. Expresso também minha gratidão à CAPES pelo suporte financeiro concedido por meio de bolsa de estudos, que tornou possível a realização deste projeto de pesquisa.

Por fim, gostaria de expressar minha profunda gratidão aos meus pais, familiares, amigos e colegas pelo apoio e suporte ao longo da minha trajetória acadêmica. Um agradecimento especial à minha esposa Juliana por estar ao meu lado em todos os momentos, fornecendo o apoio emocional necessário para enfrentar os desafios deste mestrado.

Este trabalho não teria sido possível sem o auxílio e contribuição de cada um de vocês. Sintam-se todos parte integrante desta conquista. Muito obrigado.

RESUMO

A pecuária bovina de corte desempenha um papel fundamental na economia brasileira, sendo uma das principais fontes de renda do país. Durante as várias etapas da produção, como procedimentos veterinários, pesagens e transporte, os animais são frequentemente submetidos a diferentes níveis de manejo, cada um com seu potencial de causar estresse nos animais. O estresse desenvolvido pelos animais pode ter um impacto significativo nas propriedades da carne, desde a redução da sua qualidade até torná-la inadequada para o consumo humano. Adicionalmente, um manejo mal conduzido pode resultar em lesões nos animais, que variam de leves hematomas a situações que levam à morte do animal, resultando em perdas econômicas para produtores e frigoríficos, além do sofrimento dos próprios animais. O objetivo deste trabalho foi de avaliar diferentes abordagens para a identificação da ocorrência de estresse nos animais por meio dos sons emitidos durante o manejo. A avaliação foi desenvolvida de duas formas: (i) a implementação de diferentes arquiteturas de redes neurais, incluindo MLP (*Multilayer Perceptron*), LSMT (*Long Short-Term Memory*) e CNN (*Convolutional Neural Network*), utilizando MFCC (*Mel Frequency Cepstral Coefficients*) para extração das características acústicas e (ii) a análise estatística de parâmetros acústicos de vocalizações, como a frequência fundamental (F0), os formantes do espectro (F1 - F4), *jitter*, *shimmer*, harmonia e intensidade. Os resultados obtidos revelaram que as redes neurais convolucionais atingiram melhores taxas de classificação de estresse (F1-score = 98,76%), seguidas pelas redes MLP (F1 = 90,07%) e LSTM (F1 = 85,61%). Na análise acústica observou-se diferenças significativas ($p < 0,001$) entre os parâmetros das vocalizações de estresse e não estresse em grande parte dos casos, reforçando a viabilidade de utilizar características acústicas para identificar o estado emocional dos animais durante o manejo.

Palavras-chave: Análise acústica; Bem-estar animal; Pecuária de precisão; Redes neurais artificiais.

ABSTRACT

Beef cattle farming plays a fundamental role in the Brazilian economy, being one of the main sources of income for the country. During various stages of production, such as veterinary procedures, weighing, and transportation, animals are often subjected to different levels of handling, each with its potential to cause stress in the animals. Stress developed by the animals can have a significant impact on the properties of the meat, ranging from reducing its quality to making it unsuitable for human consumption. Additionally, poorly conducted handling can result in injuries to the animals, ranging from minor bruises to situations that lead to the death of the animal, resulting in economic losses for producers and slaughterhouses, as well as the suffering of the animals themselves. The aim of this study was to evaluate different approaches to identifying the occurrence of stress in animals through the sounds emitted during handling. The evaluation was carried out in two ways: (i) the implementation of different neural network architectures, including MLP (Multilayer Perceptron), LSMT (Long Short-Term Memory), and CNN (Convolutional Neural Network), using MFCC (Mel Frequency Cepstral Coefficients) for acoustic feature extraction, and (ii) the statistical analysis of acoustic parameters of vocalizations, such as fundamental frequency (F0), spectral formants (F1 - F4), jitter, shimmer, harmonicity and intensity. The results showed that convolutional neural networks achieved better stress classification rates (F1-score = 98.76%), followed by MLP networks (F1-score = 90.07%) and LSTM (F1-score = 85.61%). In the acoustic analysis, significant differences ($p < 0.001$) were observed between the parameters of stress and non-stress vocalizations in most cases, reinforcing the feasibility of using acoustic characteristics to identify the emotional state of animals during handling.

Keywords: Acoustic analysis; Animal welfare; Artificial neural networks; Precision livestock farming.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACM	*Association for Computing Machinery*
ANN	*Artificial Neural Networks*
ANOVA	*Analysis of Variance*
ASABE	*American Society of Agricultural and Biological Engineers*
BLE	*Bluetooth Low Energy*
CFS	*Correlation-based Feature Selection*
dB	Decibel
DC	*Direct Current*
CNN	*Convolutional Neural Network*
DCT	*Discrete Cosine Transform*
DFT	*Discrete Fourier Transform*
FFT	*Fast Fourier Transform*
GLM	*Generalized Linear Model*
GRU	*Gated Recurrent Unit*
HMM	*Hidden Markov Models*
IEEE	*Institute of Electrical and Electronics Engineers*
IFFT	*Inverse Fast Fourier Transform*
IP	*Internet Protocol*
LPC	*Linear Prediction Coding*
LSD	*Least Significant Difference*
LSTM	*Long Short-Term Memory*
MFCC	*Mel Frequency Cepstrum Coefficients*
MLP	*Multilayer Perceptron*
pH	Potencial de Hidrogênio

PIB	Produto Interno Bruto
PSD	*Power Spectral Density*
RAM	*Random Access Memory*
ReLU	*Rectified Linear Unit*
RMS	*Root Mean Square*
RNN	*Recurrent Neural Network*
STFT	*Short-Time Fourier Transform*
SVDD	*Support Vector Data Description*
SVM	*Support Vector Machine*
UNIPAMPA	Universidade Federal do Pampa
UR	Umidade Relativa
USB	*Universal Serial Bus*

SUMÁRIO

1 INTRODUÇÃO

1.1 Produção, transporte e bem-estar animal

A pecuária bovina de corte é uma das principais atividades geradoras de renda no Brasil, com a produção estimada em 213 milhões de cabeças de gado (IBGE, 2018). Este setor é responsável por 6% do PIB brasileiro, sendo o maior produtor, consumidor e exportador de carne do mundo. 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 (COSTA, 2000).

O perfil dos consumidores tem mudado nos últimos anos, com uma maior exigência quanto à qualidade da carne produzida. Além dos aspectos específicos de qualidade da carne como sabor, maciez, aroma, entre outros, um crescente número de consumidores passou a se preocupar com os aspectos relacionados à forma de criação e tratamento dos animais. Mercados específicos, como o de países da União Europeia, consideram pagar um valor elevado em carnes quando há a garantia de bem-estar animal durante a produção (MOLENTO, 2005).

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. Alterações fisiológicas como aumento da frequência cardíaca, doenças, resposta imunológica reduzida e ferimentos, bem como alterações comportamentais como automutilação e agressividade são sinais que indicam que o animal está em condições precárias de bem-estar (MOLENTO, 2005). O bem-estar é uma característica individual do animal, variando dentro de uma escala de muito bom a muito ruim como resultado do que lhe é oferecido pelo homem. Em outras palavras, o homem não concede ou retira o bem-estar do animal, mas sim fornece condições que podem aumentar ou diminuir esse estado individual (COSTA, 2000).

Dentre as fases de produção da bovinocultura, o processo de transporte dos animais entre as fazendas e frigoríficos usualmente tem um impacto negativo no bem-estar e pode ter um efeito nocivo à qualidade final da carne produzida. O bem-estar do animal nessa etapa de pré-abate está diretamente relacionado com o manejo adequado nas fases de embarque, deslocamento e desembarque, onde a não adoção de boas práticas resultam

em estresse, contusões, e até mesmo em mortes de animais (COSTA, 2000). A ocorrência de estresse no animal durante esta etapa altera as propriedades da carne, tornando-a mais seca, escura, dura e, consequentemente, de menor qualidade. Além disso, partes lesionadas da carne necessitam ser removidas, o que gera prejuízos financeiros para os produtores ou para os frigoríficos (SCHWARTZKOPF-GENSWEIN et al., 2012).

A sua condução de animais para ambientes desconhecidos é um indutor importante de estresse (COSTA, 2000). Com a intenção de acelerar o processo de embarque e desembarque, comumente são utilizados estímulos: cutucões, choques elétricos e gritos são exemplos. Essas ações estressam ainda mais o animal, aumentando a sua agressividade e, consequentemente, o risco de acidentes. Outro fator que agrava a situação é a capacidade do animal de reconhecer pessoas e estabelecer ligações com determinadas situações e sensações. Ao vivenciar uma experiência negativa, o animal procura evitar as situações relacionadas a ela, fugindo, lutando e dificultando o manejo. Outros fatores que podem afetar o bem-estar animal durante o transporte são a experiência e treinamento do motorista, o manejo no embarque e no desembarque, a densidade de carregamento, a duração e a distância do transporte, o tipo de caminhão utilizado e a idade, o tamanho e a condição do animal (SCHWARTZKOPF-GENSWEIN; GRANDIN et al., 2014). Além desses fatores, as estradas em más condições, o clima extremo, a umidade, a ventilação insuficiente, a vibração, o barulho, o transporte de grupos de animais desconhecidos e a privação de alimento e água também são fatores geradores de estresse (DAMTEW et al., 2018). A Tabela 1 apresenta as causas e consequências de estresse durante o transporte.

Tabela 1 – Tipos de estresse e suas causas e consequências durante o transporte

Estresse	**Causa**	**Consequências**
Comportamental	Ambiente desconhecido, restrição, barulho, superlotação, grupos desconhecidos	Medo, interação agressiva
Nutricional	Jejum	Desidratação e fome
Físico	Superlotação, grupos desconhecidos, condições da estrada, habilidade do motorista, chifres	Hematomas e lesões
Infectuoso	Poeira, exposição	Doença respiratória

Fonte: Adaptado de (DAMTEW et al., 2018)

Como consequência do estresse, o pH da carne aumenta, adquirindo as propriedades DFD (do inglês, *dark, firm, dry* – ou escura, dura e seca), o que reduz

significativamente 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 (JORQUERA-CHAVEZ et al., 2019).

Hopkins, Bruce e Li (2016) apontam que, na Austrália, a incidência de carne escura em carcaça de bovinos é de 10%, causando prejuízos financeiros estimados em aproximadamente AU$ 36 milhões por ano para a indústria bovina (R$ 116,69 milhões, em valores de janeiro de 2024). No Brasil, Neto et al. (2015) verificaram a ocorrência de lesões em 42,4% das carcaças de bovinos devido ao transporte e manejo, acarretando em perdas financeiras que podem ultrapassar R$ 200 mil por ano para um frigorífico de porte médio. Essa situação demanda uma reformulação dos métodos utilizados atualmente no manejo e transporte dos animais, tanto para melhores condições de bem-estar do animal, quanto para maior lucratividade dos produtores.

A identificação de estresse em um grupo de animais pode não ser uma tarefa simples para pessoas que não estejam acostumadas ao manejo desses animais ou que não considerem que essa identificação seja importante. Processos de transporte de gado bovino são feitos em diversos horários e não há seres humanos capazes de identificar situações de extremo estresse, pois os animais não são acompanhados o tempo todo, seja no processo de transporte ou em outras situações. Dessa forma, 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.

1.2 Objetivos e delineamento da proposta

O objetivo deste trabalho foi realizar um estudo sobre a capacidade de redes neurais artificiais discernirem entre vocalizações de gado bovino em situações de estresse e de não estresse.

O equipamento para coleta dos dados, construído para os fins deste trabalho, contém câmeras, microfones e sensores para captação de dados do ambiente. A ideia inicial deste trabalho era embarcar o equipamento em um caminhão de transporte de gado para coletar dados de transporte e, posteriormente, analisá-los para a identificação dos principais fatores que afetam a qualidade de vida dos animais durante a etapa de transporte. Esse objetivo não pode ser alcançado em virtude de questões próprias das empresas de transporte contactadas. Dessa forma, o método de coleta foi alterado

para outro ambiente, com maior controle, e somente os sons foram usados nas análises realizadas. Considera-se que essa análise é importante, pois a coleta de imagens no processo de transporte é dificultada por condições de iluminação e trepidação.

As hipóteses de pesquisa que serão trabalhadas ao longo do texto são as seguintes:

1. Os sons emitidos por animais estressados são diferentes daqueles emitidos por animais mais tranquilos.
2. A análise dos sons emitidos pelos animais pode ser executada em tempo real para verificação de situações de estresse.

Considerando os objetivos do trabalho e as hipóteses de pesquisa apresentadas anteriormente, os seguintes objetivos específicos foram estabelecidos:

1. Construir um equipamento experimental de coleta de dados das variáveis ambientais do processo de transporte, imagens e sons dos animais transportados, que possa ser embarcado em um caminhão de transporte de gado.
2. Analisar diferentes arquiteturas de redes neurais artificiais (ANN) para verificação de suas capacidades de discernimento entre sons relacionados a situações de estresse e de não estresse.
3. Implementar uma solução de *software* que seja capaz de discernir os sons emitidos pelos animais, identificando situações de estresse.

1.3 Organização do trabalho

O texto do deste projeto de dissertação está organizado como se segue:

O Capítulo 1 está organizado em três seções, a saber: a Seção 1.1 traz a caracterização do problema que será abordado neste trabalho, em especial, as questões referentes ao transporte de gado de corte e seus impactos ambientais, econômicos e de bem-estar animal; a Seção 1.2 traz a descrição do objetivo geral, problemas de pesquisa e dos objetivos específicos do trabalho; esta Seção 1.3 apresenta a organização do texto, sumarizando o que é tratado em cada parte do trabalho. O objetivo é fazer uma apresentação com um resumo prévio, de forma que o leitor possa dirigir-se diretamente à seção de interesse.

O Capítulo 2 está organizado em quatro seções: a Seção 2.1 apresenta a caracterização do trabalho de pesquisa, detalhando o método utilizado para o

desenvolvimento do trabalho; a Seção 2.2 apresenta as etapas de construção do equipamento de coleta; a Seção 2.3 traz a revisão sistemática da literatura realizada para o levantamento do referencial teórico; a Seção 2.4 traz as ferramentas e tecnologias empregadas no desenvolvimento de *softwares* nas etapas de coleta, preparação e processamento de dados, bem como nas etapas de desenvolvimento, treinamento e validação das redes neurais.

O Capítulo 3 apresenta a fundamentação teórica da pesquisa, explorando os principais conceitos relacionados aos procedimentos e técnicas utilizados no desenvolvimento do estudo, sendo organizada em quatro seções: a Seção 3.1 explora os conceitos de processamento de som e análise acústica; a Seção 3.2 traz uma contextualização sobre a técnica de MFCC utilizada para a extração de características das vocalizações; a Seção 3.3 oferece uma fundamentação sobre as características e o desenvolvimento de redes neurais artificiais, destacando as redes MLP, LSTM e CNN.

O Capítulo 4 realiza uma revisão e discussão dos trabalhos correlatos encontrados na literatura, contextualizando o atual estado da arte em relação à análise e identificação de vocalizações animais.

O Capítulo 5 está organizado em quatro seções principais: a Seção 5.1 traz o projeto lógico do produto de coleta de dados de transporte construído, apresentando os requisitos do sistema e diagramas que facilitarão o entendimento da solução proposta, além de descrever o levantamento e análise dos componentes do sistema, onde são detalhados os métodos adotados para a definição dos materiais que constituíram o produto desenvolvido; a Seção 5.2 aborda os procedimentos aplicados para a coleta de vocalizações de bovinos em condições de estresse e não estresse; a Seção 5.3 descreve o desenvolvimento dos *softwares* utilizados na classificação de sons, desde a extração das características MFCC até a implementação e treinamento das redes neurais; a Seção 5.4 apresenta os procedimentos envolvidos na implementação dos *softwares* responsáveis pela extração e análise dos parâmetros acústicos das vocalizações.

O Capítulo 6 apresenta os principais resultados da pesquisa, incluindo uma comparação com os resultados encontrados na literatura. Este capítulo é dividido em duas seções: a Seção 6.1 apresenta os resultados da análise estatística dos parâmetros acústicos das vocalizações; a Seção 6.2 descreve os resultados obtidos por meio do treinamento das redes neurais.

O Capítulo 7 apresenta uma síntese do desenvolvimento da pesquisa, abordando todas as etapas realizadas, os resultados obtidos, as contribuições alcançadas e as

oportunidades de melhorias e pesquisas futuras.

2 MATERIAL E MÉTODOS

2.1 Caracterização e fases do trabalho de pesquisa

O trabalho de pesquisa realizado nesta dissertação pode ser caracterizado como experimental e exploratório, com procedimentos de pesquisa bibliográfica e de *Design Science Research*. As fases procedimentais do trabalho estão listadas a seguir:

1. Revisão de escopo da literatura sobre reconhecimento de vocalizações de animais de corte (com ênfase em gado bovino).
2. Levantamento de requisitos de construção do equipamento de transporte de animais, em relação às variáveis de ambiente e posicionamento das câmeras nos caminhões de transporte.
3. Projeto e construção do protótipo do equipamento de coleta de dados de transporte para ser embarcado em caminhões.
4. Montagem dos equipamentos de coleta de dados de vocalização nas cocheiras.
5. Construção do software de tratamento de dados coletados *in loco*.
6. Preparação dos dados, treinamento e análise dos resultados.

A ideia inicial do trabalho era embarcar o equipamento construído em um caminhão de transporte de animais, para coleta de sons originários do processo de transporte. As câmeras do equipamento seriam usadas, neste primeiro momento, somente para identificação de situações de estresse por um especialista, para rotular os dados de sons. As duas empresas de transporte que se comprometeram com este projeto acabaram não embarcando o equipamento, por razões alheias ao produto em si. Desta forma, o método foi alterado para que a coleta fosse feita em situações em que o gado estivesse garantidamente em situações de estresse identificado. A coleta de situações de não estresse foi realizada nas cocheiras de alimentação/descanso. A situação de estresse em que os dados foram coletados foi em uma mangueira quando os animais foram levados para a pesagem, situação em que o gado é movimentado e colocado muito próximo a outros animais, situação considerada estressante para eles.

Os métodos usados para cada uma dessas fases estão descritos nas seções seguintes e o processo de condução de cada um deles são apresentados nos Capítulos 5 e 6.

2.2 Construção do equipamento de coleta de dados de transporte

O método usado para o projeto e construção do equipamento de coleta de dados de transporte segue a estrutura da *Design Science Research* (SIMON, 1996). O objetivo desse método é produzir artefatos que ainda não existem e para os quais há potencialmente infinitas possibilidades de escolha. O método é pensado para a construção de sistemas como objetos artificiais, com propriedades não necessariamente bem definidas, que devem servir a um propósito específico. As questões de pesquisa associadas a esse método apresentam-se na forma de soluções alternativas para a classe de problemas que o sistema produzido se propõe a resolver (LACERDA et al., 2013).

Em acordo com o método escolhido, as fases do processo de projeto foram estruturadas conforme a Tabela 2.

Tabela 2 – Etapas da *Design Science* aplicadas a este trabalho

Etapa da *Design Science*	**Atividade correspondente do trabalho**
Conscientização/ Proposta	Revisão da literatura Projeto lógico do produto
Sugestão/ Tentativa	Levantamento e análise dos componentes Descrição da capacidade e limitações Projeto físico do produto
Desenvolvimento/ Artefato	Construção do sistema de coleta de dados Construção do sistema de identificação de estresse Testes unitários Testes de integração Testes de produção*
Avaliação/ Medidas de desempenho	Coleta de dados de transporte* Verificação da qualidade dos dados obtidos* Treinamento e teste do sistema de alerta*
Conclusão/ Resultados	Análise dos resultados* Conclusão da pesquisa*

Em virtude das dificuldades já relatadas em relação ao embarque do equipamento construído no caminhão de transporte, as atividades marcadas com * não foram desenvolvidas com o equipamento construído. A alternativa para validação da questão de pesquisa relacionada à capacidade de redes neurais discernirem entre situações de estresse e de não estresse foi realizar a coleta dos dados em situações de confinamento e manejo, que caracterizam as duas situações de forma clara.

A descrição das atividades da Tabela 2 que foram executadas são apresentadas na Seção 5.1.

2.3 Revisão da literatura

A etapa de análise e revisão bibliográfica foi realizada por meio de uma revisão de escopo (ARKSEY; O'MALLEY, 2005), sendo composta pelas seguintes passos: (*i*) definição das fontes de pesquisa; (*ii*) busca por palavras-chaves; (*iii*) filtragem dos resultados encontrados; (*iv*) classificação dos trabalhos relevantes encontrados; (*v*) revisão e análise dos trabalhos selecionados.

Note-se que a revisão escopo possui essencialmente as mesmas atividades de uma revisão sistemática (KITCHENHAM, 2004), mas com algumas diferenças importantes, relacionadas tanto ao domínio do problema como ao tipo de material pesquisado. Revisões sistemáticas são oriundas da área da Saúde e dizem respeito a intervenções (cirurgias, medicações, tratamentos de forma geral) que podem gerar diferentes resultados. Revisões sistemáticas usualmente buscam as melhores práticas em contextos que são predominantemente empíricos e, por essa razão, usualmente são organizadas sobre a tríade população-intervenção-resultado e buscam registros em todo tipo de fonte, não somente em referenciais bibliográficos com revisão por pares. A revisão de escopo, por outro lado, busca determinar a abrangência da literatura relacionada a um tópico específico, com indicação do volume de achados e principais focos de trabalho (MUNN et al., 2018) sendo portanto, mais adequada para encontrar lacunas no corpo de conhecimento que possam identificar oportunidades de pesquisa de desenvolvimento científico e tecnológico.

As fontes de pesquisa escolhidas foram divididas em quatro tipos: serviços de indexação, repositórios digitais, periódicos indexados e anais de eventos científicos. As escolhas foram determinadas pela chance de que os repositórios pudessem conter trabalhos relacionados ao tema deste trabalho de Mestrado. As bases da ACM, IEEE e Springer contém boa parte dos trabalhos publicados relacionados à Computação. A área das Ciências Agrárias aparece com mais frequência nas bases ASABE, SciELO e ScienceDirect. Essa última contém os periódicos mais importantes relacionados à Agricultura Digital. Google Sholar é um repositórios genérico de trabalhos já publicados e poderia conter trabalhos publicados em anais de eventos que não foram explicitamente considerados. O ResearchGate é uma rede social de pesquisadores e muitas pessoas referenciam seus trabalhos nessa base, sendo também uma forma de encontrar trabalhos relacionados. A Tabela 3 mostra, para cada categoria, as fontes de dados utilizadas.

No segundo passo, foi realizada a busca, em Português e Inglês, por

Tabela 3 – Fontes de dados para revisão da literatura

Tipo	**Fonte pesquisada**
Serviços de indexação	Google
Bibliotecas digitais	ACM, IEEE
Periódicos indexados	ASABE, ScienceDirect, SciELO, Springer
Anais de eventos	SBIAgro
Outras bases	Google Scholar, ResearchGate

palavras-chave que se relacionam com a proposta em desenvolvimento nas fontes de pesquisa apresentadas na Tabela 3. A Tabela 4 apresenta as palavras-chave utilizadas em relação às fontes pesquisadas.

Tabela 4 – Palavras-chave utilizadas na revisão da literatura

Fonte pesquisada	**Palavras-chave pesquisadas**
Google, Google Scholar	*cattle transport*, *cattle stress*, *cattle audio analysis*, *cattle welfare*, *neural networks cattle welfare* ,transporte bovino, bem-estar animal, estresse animal, bovinocultura
ACM, IEEE, ResearchGate	*cattle transport*, *cattle stress*, *cattle welfare*, *cattle audio analysis*, *neural networks cattle welfare*, *neural networks cattle vocalizations*
ASABE, SciELO, Springer, ScienceDirect	*cattle transport*, *cattle stress*, *cattle welfare*, *cattle audio analysis*, *neural networks cattle welfare*, *neural networks vocalizations*
SBIAgro	bem-estar animal, bovinocultura, estresse animal, transporte bovino

No terceiro passo foi realizada a leitura dos resumos, introduções, conclusões e referências dos trabalhos selecionados com a finalidade de identificar a relevância dos trabalhos para a realização pesquisa, como também para identificar potenciais trabalhos relevantes que não foram encontrados pelos métodos de pesquisa utilizados.

O quarto passo foi compreendido na classificação dos trabalhos relevantes, onde foi atribuída uma nota de 1 a 5 conforme a relevância e similaridade do trabalho encontrado com a pesquisa em desenvolvimento. A nota 1 foi atribuída para trabalhos com pouca similaridade com a presente pesquisa, quanto ao problema de pesquisa e as técnicas utilizadas. A nota 2 foi atribuída para trabalhos com pouca similaridade quanto ao problema de pesquisa, porém com alguma similaridade quanto às técnicas utilizadas. Os trabalhos com problemas similares, mas que utilizaram técnicas diferentes, foram classificados com a nota 3. Para os trabalhos com o mesmo problema de pesquisa e com similaridade quanto as técnicas utilizadas, foi atribuída a nota 4. Os trabalhos

com o mesmo problema de pesquisa e que utilizaram as mesmas técnicas que a presente pesquisa, receberam a nota 5.

Por último, no quinto passo foi realizada uma análise abrangente dos trabalhos, por ordem de relevância. Na análise dos trabalhos, novas fontes puderam ser identificadas e foram incorporadas ao métodos. O processo de busca terminou quando nenhum novo trabalho relacionado com índice de relevância maior que 2 foi encontrado no processo.

2.4 Coleta de dados, treinamento e validação das redes

A pesquisa foi conduzida utilizando a linguagem Python (https://www.python.org/), uma escolha devido à sua versatilidade, sendo amplamente empregada em diversas aplicações, desde desenvolvimento web até análise de dados, aprendizado de máquina, automação de tarefas, jogos, entre outros (MATTHES, 2023). Os programas desenvolvidos foram executados no ambiente do Visual Studio Code (https://code.visualstudio.com/), um editor de código-fonte desenvolvido pela Microsoft. Reconhecido por sua interface de usuário leve, rápida e altamente personalizável, o Visual Studio Code é popular entre os desenvolvedores de software e é conhecido por sua integração com uma variedade de ferramentas e linguagens de programação, incluindo o Python.

A exceção ao ambiente de execução foi na fase de desenvolvimento do módulo de identificação de estresse animal, onde foi utilizado o Google Colaboratory. Esta plataforma baseada na nuvem oferece a conveniência de escrever, executar e compartilhar código em Python diretamente no navegador, eliminando a necessidade de configurar um ambiente de desenvolvimento local (BISONG; BISONG, 2019). Além disso, o Google Colaboratory proporciona acesso a recursos computacionais robustos, como GPUs (*Graphics Processing Units*) e TPUs (*Tensor Processing Units*), que aceleram significativamente o treinamento de modelos de machine learning e o processamento de grandes volumes de dados. A utilização do Google Colaboratory foi crucial para aproveitar tais recursos e facilitar o desenvolvimento e experimentação com diversas arquiteturas de redes neurais.

No desenvolvimento dos códigos Python para o processamento de áudio, foram utilizadas as bibliotecas *PyAudio* (https://pypi.org/project/PyAudio/), *Librosa* (https://librosa.org/), *Matplotlib* (https://matplotlib.org/) e *Parselmouth* (https://pypi.org/project/praat-parselmouth/). A biblioteca *PyAudio* oferece

funcionalidades para gravação e reprodução de áudio, enquanto a *Librosa* é amplamente reconhecida por suas capacidades de análise e processamento de áudio, sendo utilizada para a aplicação da técnica de MFCC. A biblioteca *Matplotlib* foi empregada para a visualização de dados e gráficos resultantes do processamento de áudio. Por fim, a A biblioteca *Parselmouth* oferece uma interface Python que possibilita a interação com as funcionalidades do software *Praat* (BOERSMA, 2011), uma ferramenta amplamente reconhecida para análise acústica e frequentemente empregada em estudos sobre vocalizações animais (TORRE et al., 2015; GAVOJDIAN et al., 2023; YAJUVENDRA et al., 2013; GREEN, 2020; LEE et al., 2014; YEON et al., 2006).

Para a tarefa de treinamento de redes neurais, foram empregadas as bibliotecas *Scikit-Learn* (https://pypi.org/project/scikit-learn/), *TensorFlow* (https://pypi.org/project/tensorflow/) e *Keras* (https://pypi.org/project/keras/). O *Scikit-Learn* fornece uma variedade de algoritmos de aprendizado de máquina, incluindo aqueles utilizados para classificação e regressão. O *TensorFlow* e o *Keras*, por sua vez, são ferramentas essenciais para a construção e treinamento de redes neurais, fornecendo uma API (*Application Programming Interface*) de alto nível que simplifica o processo de desenvolvimento.

Na implementação das análises estatísticas foram utilizadas as bibliotecas *Scipy* (https://pypi.org/project/scipy/) e *Statsmodels* (https://pypi.org/project/statsmodels/). A *Scipy* oferece uma ampla gama de funcionalidades para computação científica, incluindo ferramentas estatísticas para análise de dados. Já o *Statsmodels* é uma biblioteca especializada em estatística, proporcionando recursos para realizar uma variedade de testes de hipóteses, estimativas de modelos estatísticos, entre outros. Essas ferramentas foram fundamentais para a análise estatística dos resultados obtidos na pesquisa.

Na etapa de preparação da base de dados, utilizou-se o *software* Movavi Video Editor, uma ferramenta versátil que oferece recursos para visualização, recorte, edição, montagem e conversão de vídeos. Além disso, a ferramenta possibilita a aplicação de filtros digitais para redução de ruídos, o que é crucial para aprimorar a qualidade do áudio e garantir a eficácia das análises subsequentes. A escolha do Movavi Video Editor proporcionou uma abordagem completa para o processamento dos vídeos, garantindo uma preparação adequada dos dados para as etapas seguintes da pesquisa.

A análise estatística empregada para validar os resultados do treinamento foi realizada utilizando a Análise de Variância (ANOVA), juntamente com desdobramentos por meio de testes de Tukey. A ANOVA é uma técnica estatística empregada para

comparar as médias de três ou mais grupos independentes. Seu propósito é verificar se existe uma diferença significativa entre as médias dos grupos, sendo comumente utilizada em experimentos nos quais o pesquisador busca comparar as médias de diferentes grupos ou condições (ST; WOLD et al., 1989).

Para a análise acústica dos parâmetros vocais, foi utilizado o teste t de Student. Este teste é uma versão específica da ANOVA aplicada quando há apenas dois grupos. Ele permite avaliar se as médias dos dois grupos são estatisticamente diferentes entre si, possibilitando a verificação de uma diferença significativa entre eles.

3 FUNDAMENTAÇÃO TEÓRICA

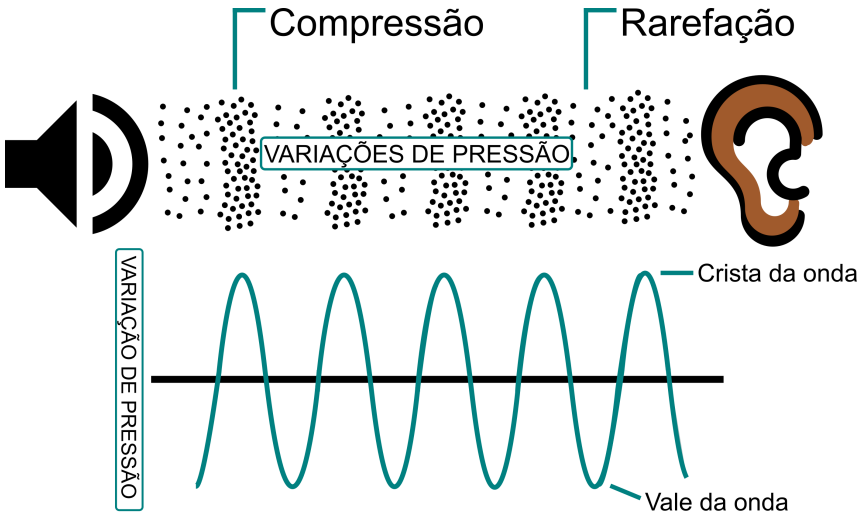
3.1 Processamento de som e análise acústica

O som desempenha um papel fundamental em nossa existência, servindo como meio primário de comunicação e expressão para organismos vivos. Fisicamente, o som consiste em ondas mecânicas que se propagam através de um meio, geralmente o ar, e podem ser percebidas e mensuradas. (CHRISTENSEN, 2019).

As ondas sonoras surgem quando há vibração em objetos, variações no fluxo de ar ou mesmo fontes de calor. No ar, essas vibrações criam oscilações na pressão, alternando entre regiões de alta e baixa pressão, conhecidas como compressão e rarefação. Quando essas variações de pressão atingem nossos ouvidos, são convertidas em impulsos elétricos que o cérebro interpreta como sons (MCLOUGHLIN, 2009).

A frequência do som é determinada pela rapidez com que a pressão varia de baixa para alta e novamente para baixa. Por outro lado, a amplitude do som é determinada pela diferença entre os pontos de pressão alta e baixa. (MCLOUGHLIN, 2009). A Figura 1 apresenta a representação dos elementos de uma onda sonora.

Figura 1 – Representação da onda sonora

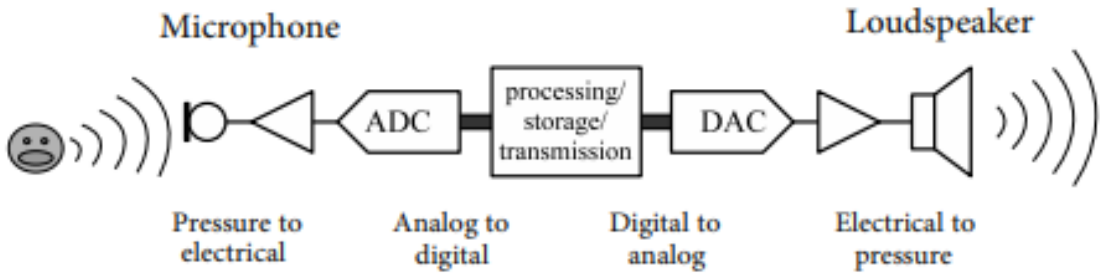


Fonte: Adaptado de McLoughlin (2009)

Assim como nosso ouvido converte as ondas sonoras em sinais elétricos, os microfones desempenham um papel semelhante ao capturar essas ondas e convertê-las em sinais elétricos, que podem ser medidos em termos de tensão ou corrente. Esses sinais elétricos são então transformados em dados digitais por meio de um conversor

analógico-digital (ADC). Em seguida, para reproduzir o som original, os dados digitais são convertidos de volta para um sinal analógico através de um conversor digital-analógico (DAC). Combinado com um amplificador conectado a um alto-falante, o sinal analógico é amplificado e convertido novamente em ondas sonoras audíveis. Esse processo permite a reprodução fiel do som original, completando o ciclo de captura, processamento e reprodução do áudio (CHRISTENSEN, 2019). A Figura 2 apresenta o diagrama de um sistema digital de áudio.

Figura 2 – Diagrama de blocos de um sistema de áudio digital



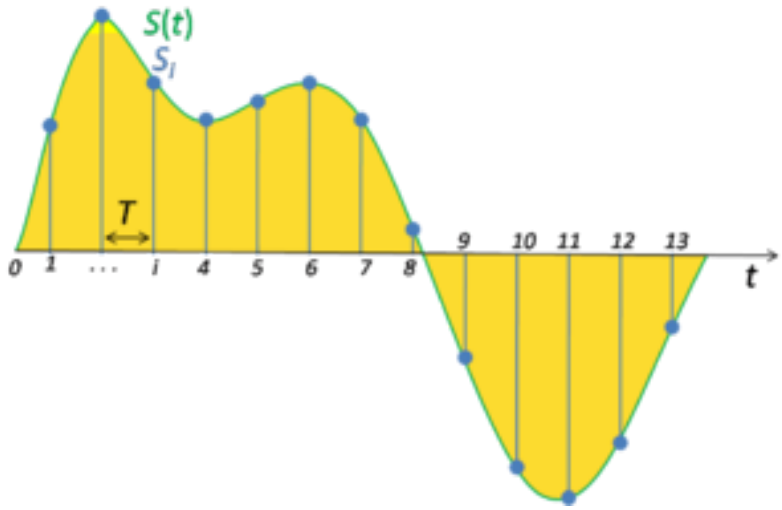
Fonte: McLoughlin (2009)

Para realizar o processamento digital de sinais de áudio, é necessário converter os sinais contínuos no tempo em amostras discretas. Esse processo, conhecido como amostragem, envolve medir o valor do sinal contínuo $S(t)$ em intervalos regulares de tempo, chamados de intervalo de amostragem T (ROCCHESSO, 2003). As amostras resultantes formam uma sequência ordenada de números que representa o sinal original. Quanto menor o intervalo de amostragem T, mais precisa será a representação do sinal contínuo. A frequência com que as amostras são obtidas por segundo é chamada de frequência de amostragem, denotada por f_s, e é definida pela equação $f_s = 1/T_s$ (CHRISTENSEN, 2019). A Figura 3 ilustra o processo de amostragem de um sinal contínuo.

Contudo, para que seja possível recuperar o sinal original, a frequência de amostragem deve ser o dobro da frequência máxima do sinal, evitando a perda de informação pela sobreposição do sinal (*aliasing*). Assim, o teorema fundamental da amostragem, ou teorema de Nyquist, é definido pela equação $f_s > 2f_{max}$ (PROAKIS, 2001).

A representação digital de um sinal consiste em uma sequência finita de números, onde cada número é expresso por um conjunto finito de dígitos binários. No

Figura 3 – Processo de amostragem de um sinal contínuo no tempo



Fonte: Christensen (2019)

entanto, sistemas computacionais possuem um número limitado de bits disponíveis para representar esses números, resultando em uma quantidade finita de valores diferentes que podem ser utilizados para representar o sinal. Esse processo de mapear um conjunto infinito de valores para um conjunto finito é conhecido como quantização (CHRISTENSEN, 2019). Por exemplo, se um sinal está contido dentro de um intervalo de $-\alpha$ a α, e o sistema possui β bits disponíveis, então é possível representar 2^β valores distintos. A técnica de quantização mais comum para sinais de áudio é a quantização uniforme, na qual o intervalo entre $-\alpha$ e α é dividido em 2^β partes iguais, cada uma com tamanho $\Delta = \frac{\alpha}{2^{\beta-1}}$. Quanto maior o número de bits β, menor será o valor de Δ, que é conhecido como tamanho do passo de quantização. Geralmente, os sinais de áudio são quantizados com 16, 24 ou 32 bits por amostra (CHRISTENSEN, 2019).

O erro que ocorre ao representar valores contínuos de um sinal em valores discretos finitos é conhecido como erro de quantização, ou ruído de quantização. A quantização de sinais analógicos inevitavelmente leva a perdas de informação, devido à imprecisão introduzida pelo tamanho dos passos de quantização, o que torna esse processo irreversível (PROAKIS, 2001).

Uma das operações fundamentais no processamento de áudio é a filtragem, que permite realizar diversas manipulações nos sinais de áudio. Um sinal digital é essencialmente uma sequência ordenada de números. Um filtro digital opera sobre essa sequência de entrada, representada como x_n, onde n é o índice de tempo, e a transforma para produzir uma nova sequência de números, a saída y_n. Essa saída é uma

versão modificada do sinal de entrada, com as alterações desejadas realizadas pelo filtro (CHRISTENSEN, 2019). A Figura 4 ilustra um filtro digital.

Figura 4 – Exemplo de filtro digital

Fonte: Christensen (2019)

Dentre os diversos tipos de filtros, os mais clássicos e simples são o passa-baixa, passa-alta e passa-banda. O filtro passa-baixa permite a passagem de frequências abaixo de uma determinada frequência de corte, enquanto atenua ou remove aquelas acima dela. Este tipo de filtro é comumente empregado em situações onde é desejável suavizar ou atenuar componentes de alta frequência de um sinal, como na filtragem de ruído de alta frequência ou em sistemas de áudio para eliminar componentes de alta frequência indesejados (CHRISTENSEN, 2019).

Por outro lado, os filtros passa-alta eliminam as frequências abaixo da frequência de corte, enquanto preservam todas aquelas acima dela. Na área de áudio, esses filtros são frequentemente utilizados para suprimir os graves de alguns instrumentos antes da mixagem, sendo também conhecidos como filtros de corte baixo (CHRISTENSEN, 2019).

Já o filtro passa-banda permite a passagem apenas das frequências do sinal de entrada entre duas frequências de corte especificadas, enquanto atenua ou remove as demais. Ele pode ser obtido pela combinação em série de um filtro passa-baixa e um filtro passa-alta, onde a frequência de corte do filtro passa-alta é menor que a do filtro passa-baixa. A diferença entre essas duas frequências de corte é denominada largura de banda (CHRISTENSEN, 2019).

A Transformada de Fourier é uma ferramenta fundamental no processamento de sinais e na análise de sistemas lineares e invariantes no tempo. Ela desempenha um papel crucial na decomposição de sinais complexos em suas componentes de frequência. Em termos simples, a Transformada de Fourier converte um sinal de domínio do tempo em seu equivalente no domínio da frequência. Isso significa que ela nos permite analisar um sinal para determinar quais componentes de frequência estão presentes e com que intensidade (CHRISTENSEN, 2019). Para sinais discretos, como aqueles comumente encontrados em

sistemas digitais, usamos a Transformada Discreta de Fourier (DFT, do inglês *(Discrete Fourier Transform)*), que é uma versão discreta da Transformada de Fourier. A equação 1 define a equação da DFT.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}} \qquad 0 \leq k \leq N-1 \tag{1}$$

Onde:

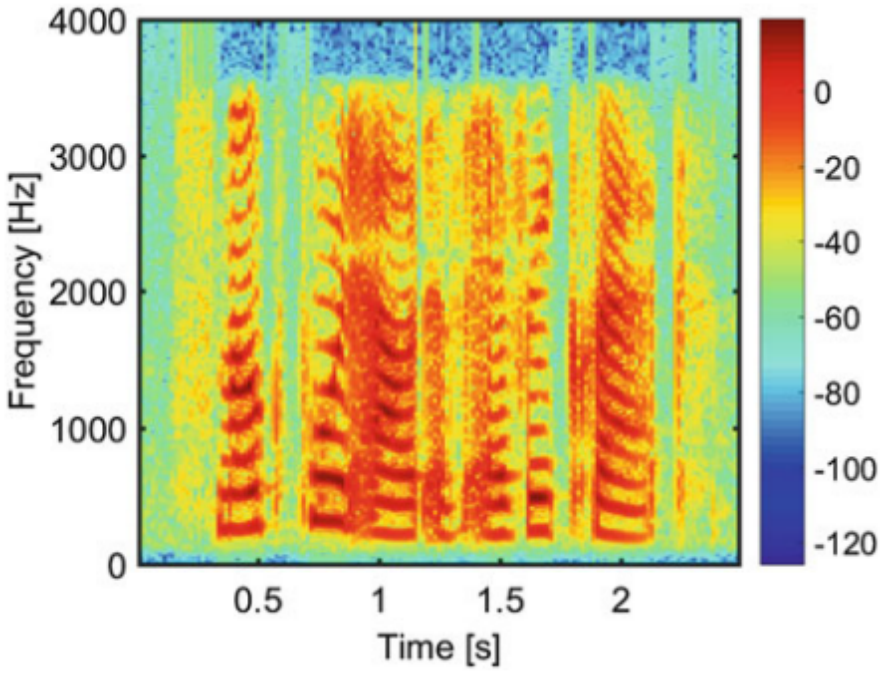
- $X(k)$: é o k-ésimo componente DFT, que representa a contribuição da frequência k ao sinal.
- $x(n)$: é o valor do sinal discreto no instante de tempo n
- N: é o comprimento total do sinal discreto
- k: é o índice de frequência, indicando a componente de frequência a ser analisada
- $e^{\frac{-j2\pi nk}{N}}$: é a exponencial complexa que representa a contribuição de uma determinada frequência k no sinal $x(n)$

Uma maneira comum de analisar e representar sinais de áudio é através do espectrograma, uma representação bidimensional que fornece uma visão detalhada dos conteúdos do sinal ao longo do tempo e da frequência. O espectrograma é uma imagem na qual o eixo horizontal representa o tempo e o eixo vertical representa a frequência. A intensidade do sinal, ou seja, a potência em diferentes frequências e momentos no tempo, é codificada por meio de cores, como vermelho para alta potência, azul para baixa potência e gradientes de amarelo entre eles. Essa representação visual é valiosa para analisar as características espectrais do sinal de áudio ao longo do tempo (CHRISTENSEN, 2019).

Para obter um espectrograma, o sinal de áudio é inicialmente subdividido em pequenos segmentos de tempo conhecidos como janelas. Em seguida, para cada janela, é aplicada a Transformada de Fourier de Curto Tempo (STFT), que calcula a transformada de Fourier do sinal dentro daquela janela específica. Isso permite analisar as características de frequência do sinal ao longo do tempo. Após o cálculo da STFT, o espectro de potência é obtido, o qual representa a distribuição de energia do sinal em diferentes frequências e ao longo do tempo (MCLOUGHLIN, 2009). Esse processo proporciona uma visualização intuitiva e informativa das características espectrais do sinal de áudio. A Figura 5 ilustra um exemplo de espectrograma de um sinal de áudio.

Além do processamento de som, é crucial compreender não apenas como os sinais de áudio são transformados em representações visuais, como o espectrograma, mas

Figura 5 – Espectrograma de um sinal de áudio



Fonte: Christensen (2019)

também os elementos fundamentais que compõem esses sinais. Dentre esses elementos, destacam-se características específicas, como a frequência fundamental, os formantes do espectro, o *jitter*, o *shimmer*, a intensidade e a harmonia, cada um desempenhando um papel vital na identificação e na análise do conteúdo sonoro.

A frequência fundamental, também chamada de *pitch* ou F0, é a componente de frequência mais baixa de um sinal sonoro, e que se relaciona harmonicamente com as outras parciais, o que significa que a frequência da maioria das parciais está relacionada à frequência da parcial mais baixa por uma pequena proporção de números inteiros (GERHARD et al., 2003). Em estudos sobre vocalizações animais, como os de bovinos, o *pitch* pode ser uma métrica relevante para avaliar variações na comunicação vocal, refletindo diferenças na emoção, comportamento ou estado de saúde dos animais (TORRE et al., 2015; GREEN, 2020).

Os formantes do espectro podem ser definidos como picos de energia em uma região do espectro sonoro. São caracterizados pela frequência do pico, pelo fator de ressonância e pelo nível de amplitude relativa do som (LEE et al., 2014). Um formante é um modo natural de vibração (ressonância) do trato vocal, e Como a maioria dos mamíferos não consegue alterar a forma ou as dimensões do seu trato vocal, pois ele é fisicamente restringido por estruturas esqueléticas, o comprimento do trato vocal

em bovinos correlaciona-se significativamente com a dispersão de formantes (GREEN, 2020).

Em bovinos, a análise dos formantes do espectro pode fornecer dicas e indícios sobre a idade, tamanho e/ou gênero do vocalizador (TORRE et al., 2015). Mesmo para animais que podem alterar sua voz formato do trato, por exemplo, veado vermelho que pode retrair sua laringe, o formante mínimo a dispersão ainda se correlaciona com a idade e o tamanho corporal (GREEN, 2020). Em bezerros de corte, foi demonstrado que as frequências dos formantes diminuíram à medida que os bezerros envelheceram, sendo um resultado direto do crescimento e desenvolvimento dos animais (TORRE et al., 2015).

O *jitter* e o *shimmer* representam variações na frequência fundamental. Enquanto o *jitter* indica a variabilidade ou perturbação na frequência fundamental, o *shimmer* refere-se à mesma perturbação, mas relacionada à amplitude da onda sonora, ou seja, à intensidade da emissão vocal. O *jitter* é influenciado principalmente pela falta de controle de vibração das pregas vocais, enquanto o *shimmer* está relacionado à redução da resistência glótica e lesões de massa nas pregas vocais, frequentemente correlacionado com a presença de ruído. Em estudos com bovinos, valores elevados para *jitter* e *shimmer* foram associados à presença de estresse (YAJUVENDRA et al., 2013).

A harmonia representa o grau de periodicidade acústica, sendo medida como a razão entre a energia harmônica e a energia não harmônica na vocalização. Valores mais altos refletem vocalizações mais tonais (GREEN, 2020). Em estudos com animais, a harmonia foi empregada na caracterização e identificação de vocalizações (GREEN, 2020; LINHART et al., 2015; MAIGROT; HILLMANN; BRIEFER, 2018).

A intensidade é uma medida da potência do som por unidade de área, descrevendo a quantidade de energia sonora transmitida por uma onda sonora em uma determinada região. Quanto maior a intensidade, mais energia sonora está presente, e, portanto, o som é percebido como mais alto. Em estudos com animais, a intensidade tem sido utilizada para analisar vocalizações em condições estressantes e não estressantes (MOURA et al., 2008; LINHART et al., 2015).

Uma das principais áreas de aplicação envolvendo o processamento de som e análise acústica é a classificação e reconhecimento de sons. Ao utilizar técnicas como espectrogramas, extração de parâmetros acústicos e outras ferramentas de processamento de som, é possível investigar a composição e as características dos sons e utilizá-las como entrada em sistemas de aprendizado profundo. No contexto de estudos com animais, o processamento e análise acústica das vocalizações podem fornecer informações cruciais

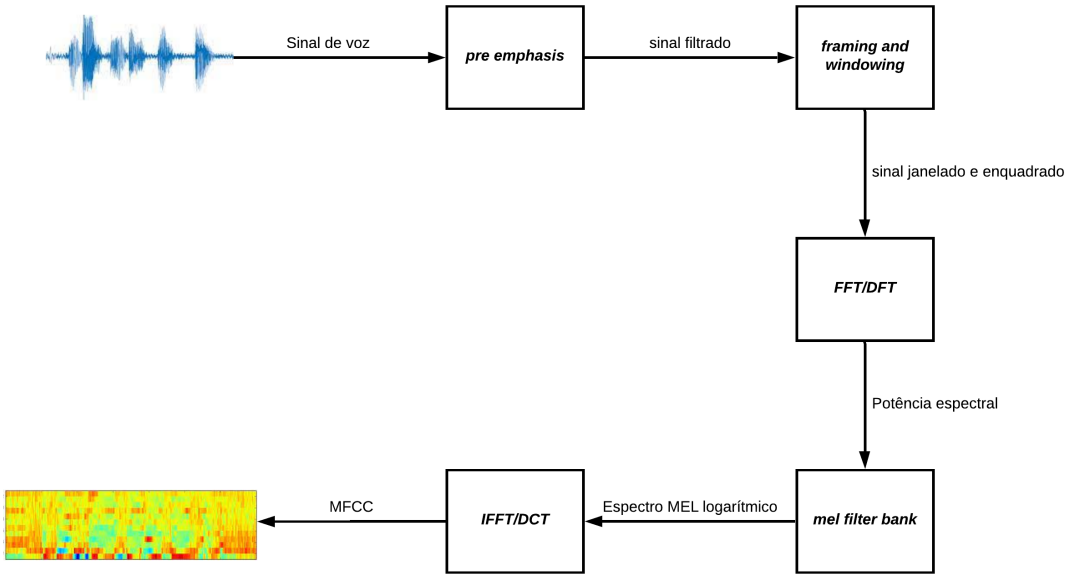
sobre o estado emocional dos animais, além de outros aspectos relevantes. Na próxima seção, a análise será aprofundada, explorando especificamente a técnica de *Mel Frequency Cepstral Coefficients* (MFCC) e seu papel na extração de características acústicas para a classificação de sinais de áudio.

3.2 *Mel Frequency Cepstrum Coefficients* (MFCC)

O MFCC é uma das técnica mais populares para extração de *features* de som para a realização de reconhecimento de voz no domínio da frequência (DAVE, 2013). 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, baseando-se na escala do ouvido humano (TIWARI, 2010).

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. A Figura 6 apresenta o processo de extração de *features* de MFCC.

Figura 6 – Processo de extração das *features* de MFCC



Fonte: Adaptado de Tiwari (2010)

A etapa de *pre emphasis* se refere à filtragem que enfatiza as frequências mais altas do sinal de voz. Seu objetivo é equilibrar o espectro sonoro, compensando a atenuação

das componentes de alta frequência causadas pelo mecanismo da produção de voz (RAO; MANJUNATH, 2017).

Como o sinal de voz varia lentamente no tempo, ele pode ser considerado estacionário se analisado em segmentos com períodos de tempo suficientemente curtos. A etapa de *framing and windowing* é necessária atenuar as descontinuidades nas bordas de cada segmento do sinal. As medições espectrais de curto prazo são normalmente realizadas ao longo janelas Hamming de 20 ms, com deslocamento entre janelas de 10ms (DELLER; PROAKIS; HANSEN, 2000). Avançando na janela de tempo a cada 10ms habilita as características temporais de sons de fala individuais para serem rastreadas, e a janela de análise de 20ms geralmente é suficiente para fornecer resolução espectral desses sons e, ao mesmo tempo, curta o suficiente para resolver características temporais significativas (RAO; MANJUNATH, 2017).

Após o sinal ser janelado, é aplicado a cada *frame* do sinal a DFT, ou a FFT (*Fast Fourier Transform*), que é basicamente um algoritmo mais eficiente para se calcular a DFT, definida na Equação 1, onde assume-se que N = 2n, reduzindo assim a complexidade do algoritmo de O(n^2) para O(n log n), facilitando a computação das funções (PROAKIS, 2001). Com a aplicação da FFT se obtém as componentes de frequência e a potência espectral de cada *frame* do sinal (RAO; MANJUNATH, 2017).

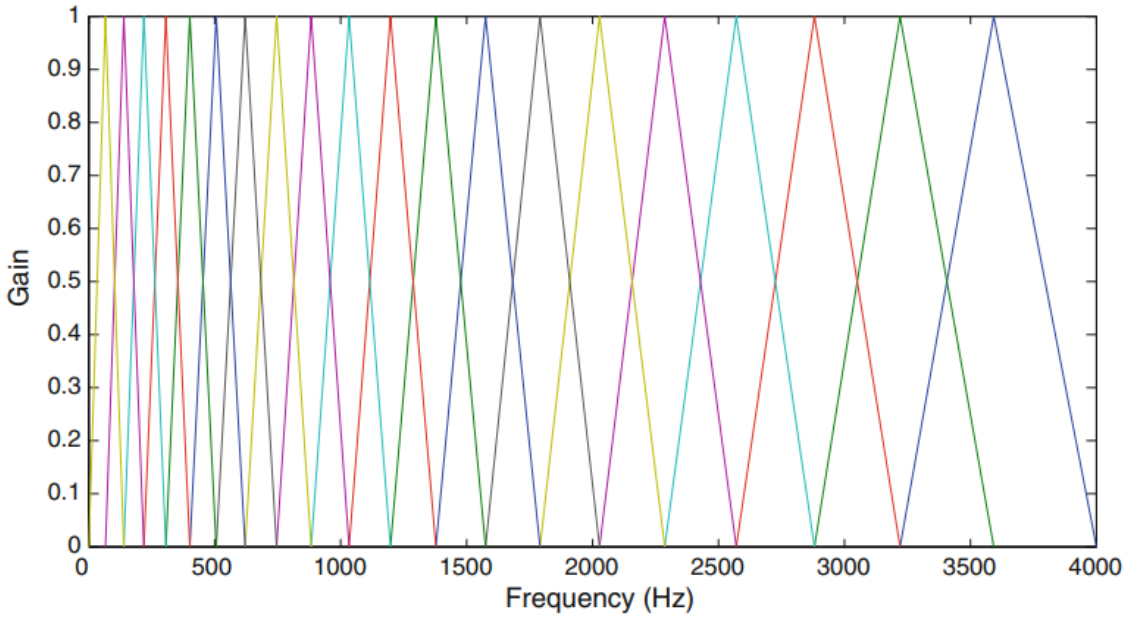
Na etapa *mel filter bank* o espectro Mel é computado passando o sinal transformado por um conjunto de filtros passa-banda, conhecido como *mel filter bank*, visando imitar a resposta em frequência do sistema auditivo humano. A escala Mel é aproximadamente linear até frequências de 1 kHz, e logarítmica para frequências maiores (DAVE, 2013). A aproximação da escala Mel da frequência física pode ser expressa na Equação 2, onde f denota a frequência física em Hz, e f_{MEL} denota a frequência percebida (RAO; MANJUNATH, 2017).

$$f_{MEL} = 2595\, log_{10}\,(1 + \frac{f}{700}) \tag{2}$$

O banco de filtros Mel pode ser implementado tanto no domínio do tempo, quanto no domínio da frequência. Para a computação do MFCC, os bancos são geralmente implementados por filtros triangulares no domínio da frequência. Cada filtro calcula a média do espectro em torno da frequência central, possuindo maiores larguras de banda conforme maior for a frequência (RAO; MANJUNATH, 2017). A Figura 7 apresenta o banco de filtros Mel.

No último passo da MFCC o espectro logarítmico da escala Mel é convertido

Figura 7 – Banco de filtros Mel



Fonte: Rao e Manjunath (2017)

de volta para o domínio do tempo (DAVE, 2013). Para isso, existem duas possíveis alternativas, a DCT (*discrete cosine transform*) e a IFFT (*inverse fast fourier transform*). A vantagem da DCT sobre a IFFT é que ela reduz o número de coeficientes gerados, devido a propriedade conhecida como compactação de energia, concentrando os valores mais significativos nos primeiros termos do vetor, e descartando os últimos, obtendo melhor eficiência computacional (SIQUEIRA, 2011). A DCT é aplicada aos coeficientes de frequência Mel, produzindo um conjunto de coeficientes cepstrais. Isso resulta em um sinal com pico de frequência correspondendo ao tom do sinal, e o número de formantes representando picos de baixa frequência. Uma vez que a maioria das informações do sinal é representada pelos primeiros coeficientes de MFCC, o sistema pode se tornar robusto, extraindo apenas aqueles coeficientes, ignorando ou truncando componentes de DCT de ordem superior (RAO; MANJUNATH, 2017). Finalmente, a MFCC é calculada a partir da Equação 3, onde $c(n)$ são os coeficientes cepstrais, s_m é a saída do passo anterior, e M é o número de coeficientes de MFCC.

$$c(n) = \sum_{m=0}^{M-1} (log_{10}\ s_m)\, cos\left(\frac{\pi n(m-0.5)}{M}\right); \qquad n = 0,1,2,...,M-1 \tag{3}$$

As características MFCCs são amplamente utilizadas em sistemas de

reconhecimento de fala, identificação de locutor, processamento de linguagem natural e sistemas de controle de voz. Devido à sua capacidade de capturar características discriminativas da fala, os MFCCs são uma escolha popular em uma variedade de aplicações relacionadas ao áudio, sendo usados como características de entrada para modelos de aprendizado de máquina, como redes neurais artificiais.

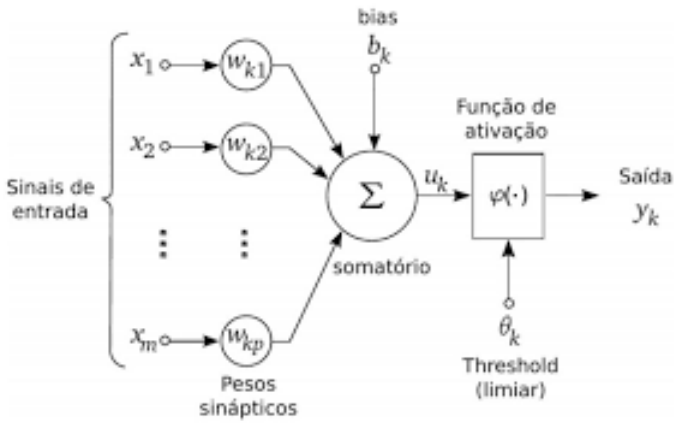
3.3 Redes neurais artificiais

Uma rede neural artificial (RNA) é uma técnica que foi inspirada na estrutura e comportamento do cérebro humano, modelada de modo a imitar a maneira como o nosso cérebro realiza suas tarefas, ou o seu "processamento" (HAYKIN, 2007).

A unidade básica de processamento de uma rede neural é chamado de neurônio, e é modelada como uma função matemática que visa emular de maneira similar o funcionamento dos neurônios do cérebro humano. Cada neurônio artificial recebe entradas, pesa-as separadamente, soma-as e passa o resultado dessa soma através de uma função não linear para gerar as saídas (SILAPARASETTY, 2020).

Um neurônio artificial é constituído pelos seguintes componentes: sinal de entrada, pesos, *bias*, somador, função de ativação e sinal de saída. O neurônio artificial mais simples é chamado de *perceptron*, que é uma estrutura que implementa um classificador binário. O algoritmo de aprendizado para um *perceptron* aprende os pesos dos sinais de entrada para traçar um limite de decisão linear (SILAPARASETTY, 2020). A Figura 8 apresenta o modelo de um neurônio artificial.

Figura 8 – Modelo de um neurônio artificial



Fonte: Haykin (2007)

O processamento de um neurônio consiste em quatro etapas: (*i*) propagação das entradas; (*ii*) cálculo da entrada líquida; (*iii*) função de ativação; (*iv*) propagação reversa.

Na primeira etapa o neurônio recebe as entradas na forma de *features* e realiza o processamento para prever a saída. A saída por sua vez é comparada aos rótulos (ou classes) para medir o erro de predição. As entradas estão representadas na Figura 8 por $(x_1, x_2, ..., x_m)$, onde x_i representa o valor da *feature* i, $i = 1, \ldots, m$ e m representa o número total de *features* de entrada (SILAPARASETTY, 2020).

Na etapa seguinte, o cálculo da entrada líquida é resultado de dois cálculos que ocorrem sucessivamente. O primeiro cálculo é o somatório da multiplicação de cada peso w_{kj} por seu valor associado x_j. Após isso, o valor do *bias* é adicionado ao produto, esse valor tem o efeito de aumentar ou diminuir a entrada líquida da função de ativação, dependendo se ele é positivo ou negativo, respectivamente (HAYKIN, 2007).

Na terceira etapa, a função de ativação adiciona ao neurônio a capacidade de aprender funções que não são uma mera combinação linear das entradas, ou seja fornece ao neurônio a capacidade de aprender funções não lineares, tentando reproduzir o limiar de disparo das sinapses em um neurônio biológico. As funções de ativação mais utilizadas são a Sigmoide, a ReLU, a Tanh e a Softmax (SILAPARASETTY, 2020). Assim, o sinal de saída de um neurônio da camada k pode ser definido pela Equação 4.

$$y_k = \phi \left(b_k + \sum_{j=1}^{m} w_{kj} x_j \right) \tag{4}$$

onde ϕ é a função de ativação, b_k é o valor do *bias*, x_j, $j = 1, \ldots, p$ é o valor de cada entrada da camada k e w_{kj} é o peso da camada k para a entrada j, e u_k é o cálculo da entrada líquida, definido pela Equação 5.

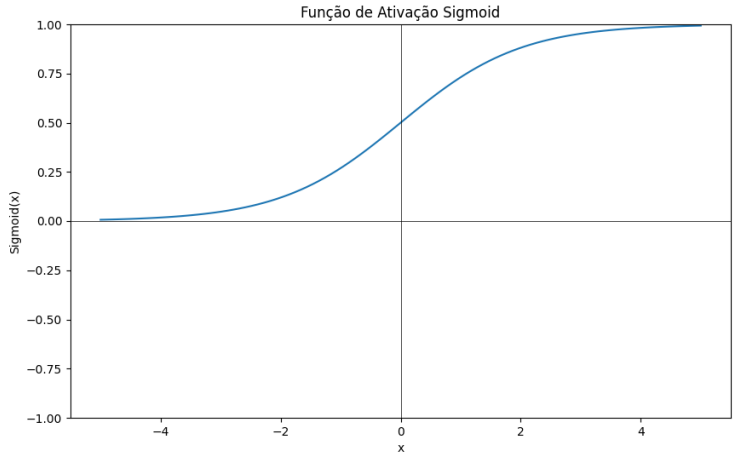
$$u_k = \sum_{j-1}^{m} w_k x_j \tag{5}$$

A função Sigmoide é uma função matemática cujo o gráfico tem o formato aproximado ao de uma função degrau, mas ao contrário desta última, a função Sigmoide é diferenciável. A função é estritamente crescente e exibe um balanceamento adequado entre um comportamento linear e não linear (HAYKIN, 2007), sendo um caso especial da função logística, resultando em valores contidos em um intervalo contínuo entre 0 e 1. A Equação 6 apresenta a função Sigmoide.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{6}$$

Essa função de ativação é útil quando o foco é o mapeamento de probabilidade, ao invés de valores precisos dos parâmetros de entrada, o que a torna útil para problemas de classificação binária, onde a saída representa a probabilidade de pertencer a uma classe (SILAPARASETTY, 2020). A Figura 9 apresenta o gráfico da função Sigmoide.

Figura 9 – Função Sigmoide



Fonte: Autor (2024)

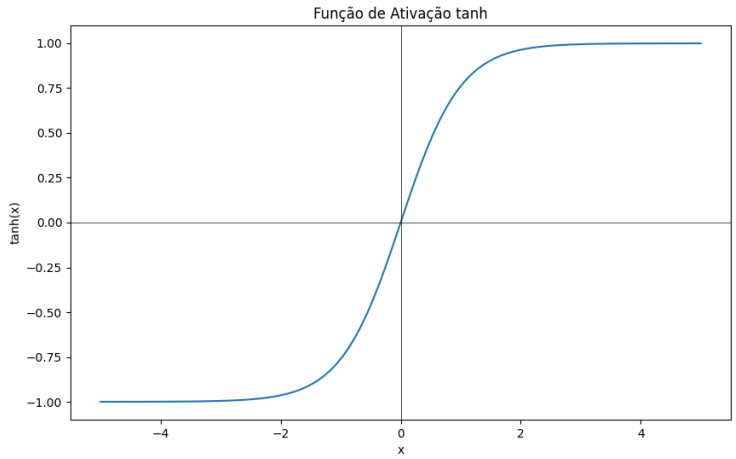
A função de ativação tangente hiperbólica, também conhecida como Tanh, tem o formato de “S", sendo contínua e diferenciável, assim como a função sigmoide. No entanto, o seu valor de saída varia de –1 a 1 e não de 0 a 1, como no caso da função sigmoide. A Equação 7 define a função Tanh.

$$tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{7}$$

O formato da função Tanh tende a normalizar cada saída da camada em torno de 0, o que frequentemente facilita e acelera a convergência durante o treinamento, especialmente quando as características possuem escalas diferentes (GÉRON, 2019). A Figura 10 apresenta o gráfico da função Tanh.

A função ReLU (do inglês “*Rectified Linear Unit*") permite eliminar valores negativos em uma rede neural artificial, como uma função linear por partes. A função tem como saída a própria entrada caso o valor seja positivo; caso contrário a saída será 0. A Equação 8 apresenta a função ReLU.

Figura 10 – Função Tanh



Fonte: Autor (2024)

$$\text{ReLU}(x) = \begin{cases} x & \text{se} \quad x \geq 0 \\ 0 & \text{se} \quad x < 0 \end{cases} \tag{8}$$

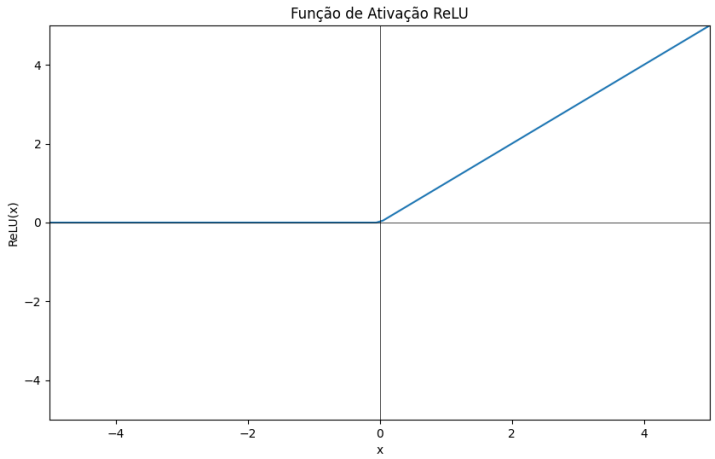
As vantagens da utilização dessa função de ativação é o rápido e efetivo treinamento de uma rede neural com *datasets* grandes e complexos, eficiente computação com apenas comparação, adição ou multiplicação, além de possuir boa escalabilidade. Entre as limitações, pode-se destacar que os valores próximos de zero podem fornecer resultados inconsistentes, o valor de saída não tem limite e pode levar a problemas computacionais com a passagem de grandes valores, além de quando a taxa de aprendizagem é muito alta, os neurônios ReLU podem se tornar inativos e "morrer" (SILAPARASETTY, 2020). A Figura 11 apresenta o gráfico da função ReLU.

Para lidar com esse problema, foram propostas variantes da função ReLU, como a Leaky ReLU, que introduz um fator de vazamento na função. Enquanto a função ReLU retorna 0 para todos os valores negativos, a Leaky ReLU retorna um valor pequeno proporcional ao valor de entrada negativo. A equação 9 define a função Leaky ReLU.

$$\text{Leaky ReLU}(x) = \begin{cases} x & \text{se} \quad x > 0 \\ \alpha x & \text{se} \quad x \leq 0 \end{cases} \tag{9}$$

O parâmetro α controla a inclinação da parte negativa da função, determinando o quanto a função "vaza", e é geralmente definido em torno de 0,01. A Leaky ReLU tem a vantagem de não zerar completamente os gradientes para valores negativos durante o

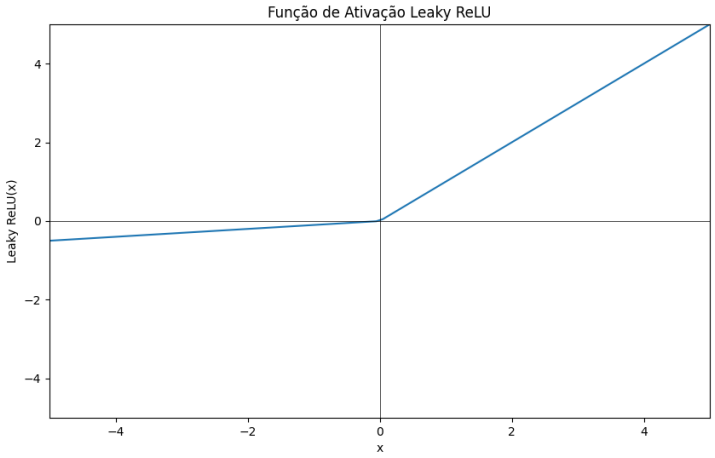
Figura 11 – Função ReLU



Fonte: Autor (2024)

treinamento, o que pode ajudar a mitigar o problema de neurônios inativos e melhorar a convergência em redes profundas (GÉRON, 2019). A Figura 12 ilustra a função Leaky ReLU.

Figura 12 – Função Leaky ReLU



Fonte: Autor (2024)

A função Softmax é uma generalização da função logística que produz a probabilidade de o resultado pertencer a um determinado conjunto de classes. A função converte um vetor k-dimensional de valores reais arbitrários para um vetor k-dimensional de valores reais no intervalo entre 0 e 1, que somam 1. É semelhante à lógica de categorização no final de uma rede neural. A Equação 10 apresenta a função Softmax.

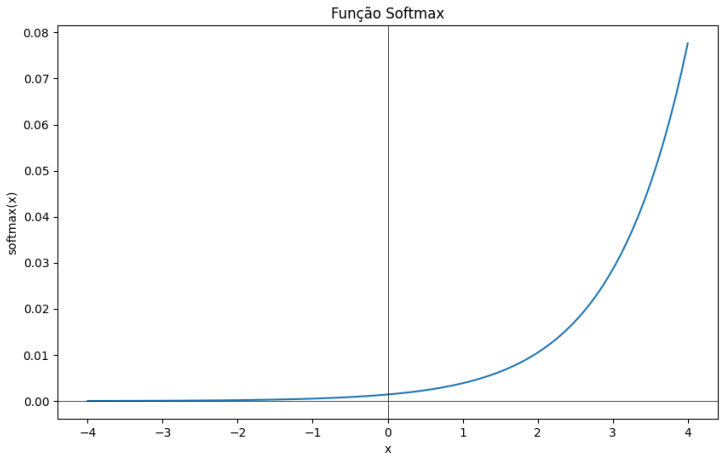
$$\sigma(x)_k = \frac{e^{x_k}}{\sum_{i=1}^{n} e^{x_i}} \tag{10}$$

Onde:

- n é o número de classes
- x é o vetor de entrada da função
- x_k i-ésimo elemento do vetor da entrada x
- $\sigma(x)_k$ é a probabilidade estimada de que a instância x pertença à classe k

Uma das principais vantagens no uso dessa função é a sua capacidade de treinamento e previsão rápidos, além de enfatizar os valores significativamente menores em relação ao máximo. No entanto, uma desvantagem importante é sua inabilidade em suportar rejeição nula, o que implica a necessidade de treinar o algoritmo com uma classe nula específica caso a rejeição nula seja necessária (SILAPARASETTY, 2020). A Figura 13 apresenta o gráfico da função Softmax.

Figura 13 – Função Softmax



Fonte: Autor (2024)

Um dos principais problemas relacionados ao treinamento de redes neurais é o sobreajuste ou *overfitting*, que é uma superadaptação da rede neural sobre o conjunto de dados, o que ocasiona uma memorização dos dados, ao invés de uma generalização, prejudicando a performance de previsão da rede quando submetida à dados novos, pois a rede só saberá reconhecer os padrões memorizados (HAYKIN, 2007). Para contornar esse problema, as técnicas mais utilizadas são o *Dropout* e a normalização em lote (*Batch Normalization*).

A função *Dropout* consiste em eliminar, ou “desligar", aleatoriamente alguns neurônios durante a etapa de aprendizagem. Esse processo tem como principal vantagem reduzir a contribuição de neurônios na propagação do sinal para a camada seguinte em uma determinada iteração. Ao desativar os neurônios, a rede é forçada a aprender padrões mais robustos e a não depender excessivamente de neurônios específicos. Isso ajuda a prevenir o sobreajuste, onde os neurônios se ajustam demasiadamente aos padrões específicos dos dados de treino (HAYKIN, 2007).

Por outro lado, a técnica de normalização em lote adiciona uma operação ao modelo antes da função de ativação de cada camada. Essa operação centraliza e normaliza as entradas em torno de zero, além de escalonar e deslocar o resultado usando dois novos parâmetros por camada (um para escalonamento, outro para deslocamento). Essa abordagem permite que o modelo aprenda a escala e a média ideais das entradas para cada camada. Estudos demonstraram que essa técnica resultou em melhorias consideráveis em redes neurais profundas, tornando-as menos sensíveis às inicializações de pesos (GÉRON, 2019).

Durante o treinamento de redes neurais, uma variedade de métricas pode ser calculada para avaliar o desempenho do modelo em relação aos dados de treinamento e validação. Essas métricas fornecem informações importantes sobre a capacidade do modelo de aprender a partir dos dados e generalizar para novos exemplos. Algumas das métricas mais comuns incluem a acurácia, que mede a proporção de previsões corretas em relação ao total de previsões, a perda (*loss*), que representa a medida de quão bem o modelo está performando durante o treinamento, a precisão, que mede a proporção de previsões positivas corretas em relação ao total de previsões positivas e o *F1-score*, que é uma média ponderada da precisão e a revocação (*recall*) e é útil para avaliar o equilíbrio entre precisão e sensibilidade do modelo Géron (2019).

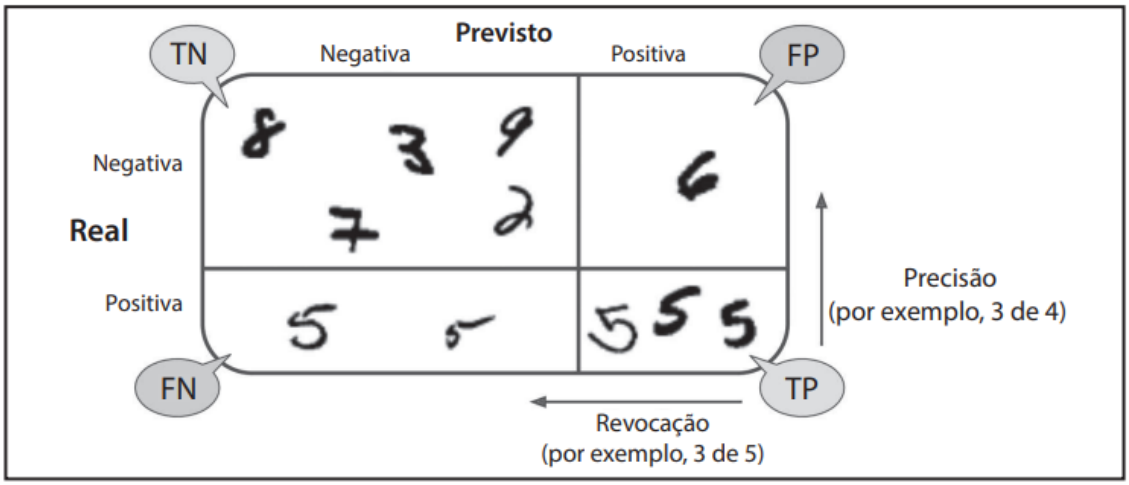
Uma maneira simples de observar essas métricas é a partir de uma matriz de confusão. É uma ferramenta útil para entender o desempenho de modelos, especialmente em problemas de classificação binária, mas também pode ser estendida para problemas de classificação multiclasse. Uma matriz de confusão organiza as previsões do modelo em relação aos valores em quatro categorias diferentes:

- Verdadeiros Positivos (*True Positives*, TP): São os casos em que o modelo previu corretamente a classe positiva.
- Verdadeiros negativos (*True Negatives*, TN): São os casos em que o modelo previu corretamente a classe negativa.

- Falsos positivos (*False Positives*, FP): São os casos em que o modelo previu erroneamente a classe positiva quando a verdadeira classe era negativa.
- Falsos negativos (*False Negatives*, FN): São os casos em que o modelo previu erroneamente a classe negativa quando a verdadeira classe era positiva.

A matriz de confusão fornece uma visão mais detalhada do desempenho do modelo do que apenas a acurácia, pois permite a identificação de onde o modelo está acertando e onde está errando. A Figura 14 apresenta um exemplo de matriz de confusão aplicada a classificação de imagens com o dígito 5.

Figura 14 – Matriz de confusão para classificador binário



Fonte: Géron (2019)

A acurácia é uma métrica que nos diz o quanto o modelo foi capaz de realizar a predição corretamente, levando em consideração tanto os casos positivos, quanto os casos negativos. A acurácia é uma medida simples e direta que indica a proporção de exemplos classificados corretamente pelo modelo em relação ao total de exemplos. Silaparasetty (2020). A equação 11 apresenta o cálculo da acurácia de um modelo preditivo.

$$\text{acurácia} = \frac{TP+TN}{TP+FP+TN+FN} \tag{11}$$

Uma métrica mais concisa e utilizada é a acurácia das previsões positivas, que é chamada de precisão. A precisão mede a proporção de exemplos classificados como positivos que realmente são positivos. É calculada como o número de verdadeiros positivos dividido pelo número de verdadeiros positivos mais falsos positivos Géron (2019). A equação 12 apresenta o cálculo da precisão de um modelo preditivo.

$$\text{precisão} = \frac{TP}{TP+FP} \tag{12}$$

A precisão é geralmente utilizada em conjunto com outra métrica chamada revocação (ou *recall*), também conhecida como sensibilidade ou taxa de verdadeiros positivos. Esta é a taxa de instâncias positivas que são corretamente detectadas pelo classificador Géron (2019). A equação 13 apresenta o cálculo da revocação de um modelo preditivo.

$$\text{revocação} = \frac{TP}{TP+FN} \tag{13}$$

O *F1-score* representa a média harmônica entre a precisão e a revocação. Enquanto a média aritmética trata todos os valores de forma igual, a média harmônica atribui mais peso aos valores mais baixos, tornando-se útil quando há desequilíbrio entre as classes. Portanto, o classificador só alcançará um *F1-score* alto se tanto a revocação quanto a precisão forem altas (GÉRON, 2019). A equação 14 mostra como calcular o *F1-score* de um modelo preditivo.

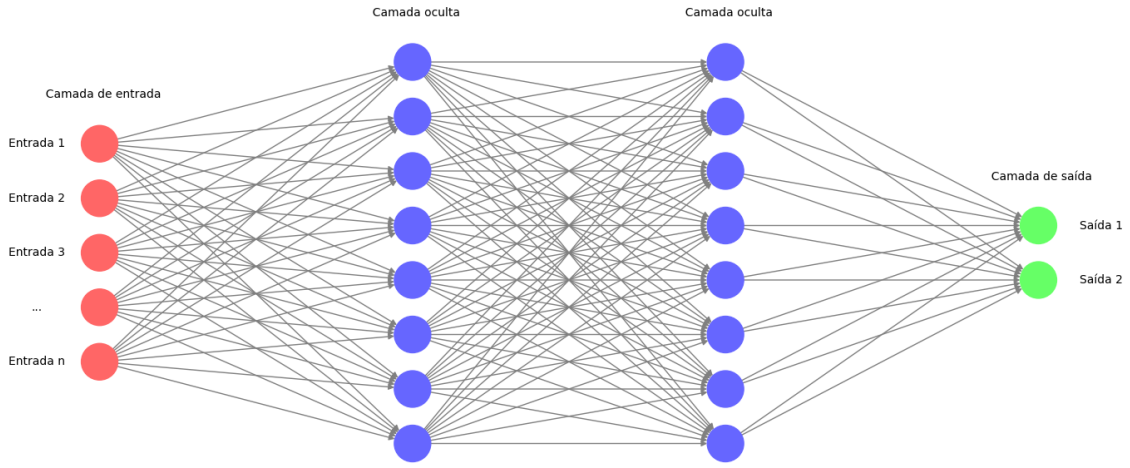
$$\textit{F1-score} = 2 * \frac{\text{precisão} * \text{revocação}}{\text{precisão} + \text{revocação}} = \frac{VP}{VP + \frac{FN+FP}{2}} \tag{14}$$

Entre os tipos mais comuns de redes neurais, destacam-se as redes neurais *feedforward*, que são o tipo mais básico de rede neural, com camadas de neurônios conectadas de forma sequencial sem formação de ciclos. Essas redes incluem as redes MLP (*Multilayer Perceptrons*), que consistem em uma ou mais camadas ocultas de neurônios, e são amplamente utilizadas para tarefas de classificação e regressão em diversos campos. Além disso, existem as redes neurais convolucionais (CNN), que foram desenvolvidas especificamente para processamento de imagens e são compostas por camadas convolucionais capazes de extrair características hierárquicas das imagens de entrada. Outro tipo importante é a rede neural recorrente (RNN), que possui conexões retroalimentadas entre os neurônios, permitindo que elas processem sequências de dados, tornando-as adequadas para tarefas como previsão de séries temporais e processamento de linguagem natural.

3.3.1 *Multilayer Perceptron* (MLP)

Uma rede neural MLP gerlamente é composta de uma camada de entrada (*passthrough*), uma ou mais camadas chamadas de camadas ocultas (*hidden layers*) e uma camada final chamada de camada de saída, onde toda camada, exceto a camada de saída, inclui um neurônio de viés e está totalmente conectada à próxima camada. Quando uma RNA possui duas ou mais camadas ocultas, é chamada de rede neural profunda (DNN, do inglês *Deep Neural Network*). (GÉRON, 2019). A Figura 15 apresenta uma típica arquitetura de uma rede MLP.

Figura 15 – Exemplo de rede MLP



Fonte: Autor (2024)

Os perceptrons de múltiplas camadas têm sido aplicados com sucesso na resolução de uma variedade de problemas complexos por meio de treinamento supervisionado, utilizando o algoritmo conhecido como retropropagação de erro (*error back-propagation*) (HAYKIN, 2007). O algoritmo alimenta cada instância de treinamento para a rede e calcula a saída de cada neurônio em cada camada consecutiva, em seguida, mede o erro de saída da rede, que corresponde à diferença entre a saída desejada e a saída real da rede. O algoritmo, então, determina a contribuição de cada neurônio para o erro em cada neurônio de saída na última camada oculta. Em seguida, avalia a contribuição desses erros provenientes de cada neurônio na camada oculta anterior, e assim por diante, retrocedendo até a camada de entrada (GÉRON, 2019). Durante a retropropagação os pesos sinápticos de todas as camadas são ajustados de acordo com uma regra de correção de erro, fazendo com que a resposta real da rede se mova para mais perto da resposta desejada (HAYKIN,

2007).

Em relação às funções de ativação em arquiteturas de redes MLP, duas opções populares são a Tangente Hiperbólica (Tanh) e a Unidade Linear Retificada (ReLU). A Tanh é preferida devido à sua saída no intervalo de -1 a 1 no início do treinamento, o que ajuda a normalizar as saídas da camada, muitas vezes acelerando a convergência. Por outro lado, a ReLU é valorizada por sua computação rápida (GÉRON, 2019). Para tarefas de classificação em que as classes são mutuamente exclusivas, a função Softmax é frequentemente aplicada na camada de saída. No entanto, em situações em que as classes não são mutuamente exclusivas ou quando há apenas duas classes, a função logística, como a sigmoide, é uma escolha comum (GÉRON, 2019).

Em relação ao número de camadas ocultas e neurônios, a seleção desses parâmetros desempenha um papel crucial na arquitetura da rede neural. O dimensionamento desses elementos influencia diretamente na capacidade da rede de aprender representações complexas dos dados. No entanto, tomar essa decisão não é trivial e depende da complexidade do problema em questão. Geralmente, é aconselhável começar com redes mais simples e aumentar a complexidade conforme necessário para obter melhores resultados no treinamento (GÉRON, 2019).

Quanto ao número de camadas ocultas, é comum iniciar com uma quantidade moderada e aumentá-las conforme necessário. A inclusão de camadas ocultas pode facilitar a rede a aprender representações mais complexas e abstratas dos dados. Contudo, é crucial evitar um número excessivo de camadas, pois isso pode resultar em sobreajuste, especialmente quando o conjunto de dados é insuficiente para justificar a complexidade adicional (GÉRON, 2019).

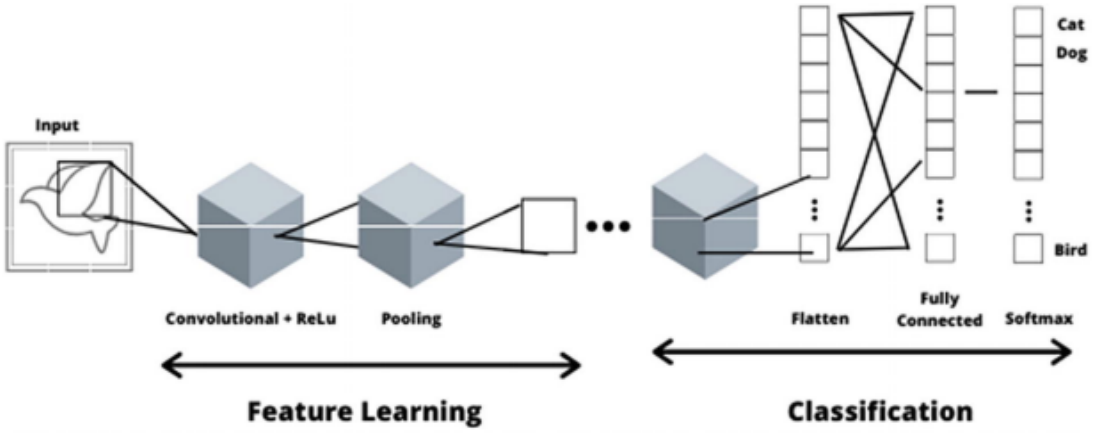
Quanto ao número de neurônios em cada camada, a estratégia pode variar de acordo com a complexidade do problema. Geralmente, começar com uma quantidade moderada de neurônios e ajustar conforme necessário é uma abordagem sensata. No que diz respeito às camadas ocultas, é comum dimensioná-las em formato de funil, onde cada camada subsequente possui progressivamente menos neurônios. Essa configuração permite que as camadas iniciais capturem características mais gerais, enquanto as camadas finais se especializam em detalhes mais específicos dos dados, facilitando uma generalização eficaz da rede (GÉRON, 2019).

3.3.2 *Convolutional Neural Network* (CNN)

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. Elas fornecem serviços de pesquisa de imagens, carros autônomos, sistemas automáticos de classificação de vídeo e muito mais. Além disso, a CNN não está restrita à percepção visual, sendo também bem-sucedida em outras tarefas, como reconhecimento de voz ou processamento de linguagem natural (PLN) (GÉRON, 2019).

Uma arquitetura típica de uma rede CNN inclui uma camada de entrada, camadas convolucionais, camadas de *pooling* (ou agrupamento) e camadas totalmente conectadas (SILAPARASETTY, 2020). O componente central de uma CNN é a camada convolucional, na qual os neurônios na primeira camada não estão conectados a todos os dados de entrada (como nas redes MLP), mas sim a uma porção dos dados. Em seguida, cada neurônio na segunda camada convolucional é conectado apenas a neurônios em uma região localizada dentro de um pequeno retângulo na camada anterior. Essa estrutura permite que a rede capture características de baixo nível na primeira camada oculta e as combine em características de nível superior na camada seguinte, e assim por diante (GÉRON, 2019). A Figura 16 ilustra uma representação típica de uma rede CNN.

Figura 16 – Arquitetura CNN padrão

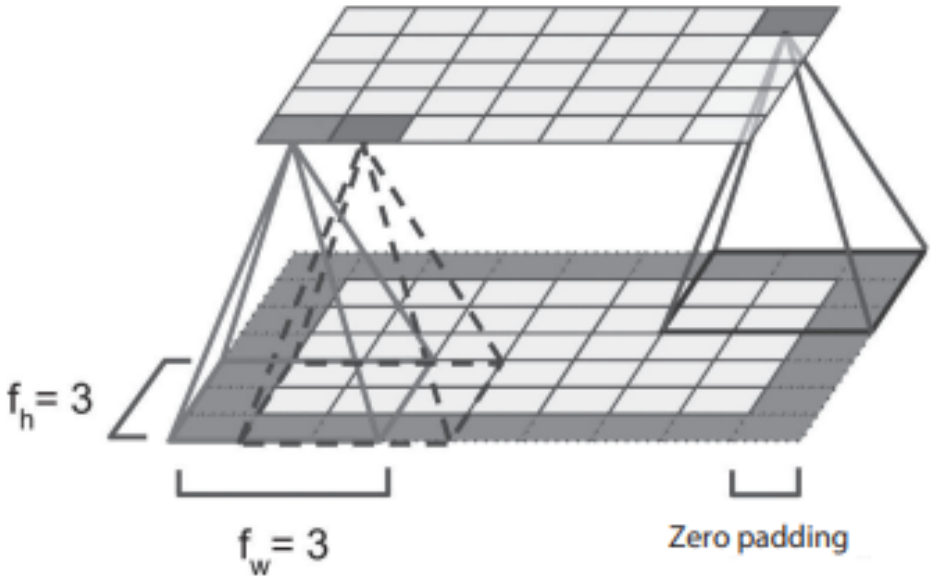


Fonte: Silaparasetty (2020)

Um neurônio localizado na linha i, coluna j de uma determinada camada está conectado às saídas dos neurônios na camada anterior, localizada nas linhas i a $i + f_h - 1$ e colunas j a $j + f_w - 1$, onde f_h e f_w representam a altura e a largura do campo receptivo, respectivamente. É comum adicionar zeros ao redor das entradas para que uma camada

tenha a mesma altura e largura da camada anterior, técnica conhecida como *zero padding*. A Figura 17 ilustra essa conexão entre as camadas convolucionais.

Figura 17 – Conexões das camadas convolucionais



Fonte: Géron (2019)

O componente responsável pela operação de convolução na primeira parte de uma camada convolucional é conhecido como *kernel* ou filtro. O objetivo dessa operação é extrair características de baixo nível, como bordas, cores, orientação de gradiente, etc., dos dados de entrada. O filtro se desloca para a direita com um determinado passo até percorrer toda a largura. Em seguida, desliza ou convolui para baixo até o início da matriz com o mesmo passo e repete esse processo até que toda a matriz tenha sido percorrida (SILAPARASETTY, 2020).

Nas camadas convolucionais seguintes, os mapas de ativação da camada anterior servem como entrada. Portanto, cada camada subsequente essencialmente descreve as áreas na matriz original onde características específicas de baixo nível são identificadas. Ao aplicar um conjunto adicional de filtros na segunda camada convolucional, são geradas ativações que representam características de nível mais elevado. À medida que progredimos na rede e atravessamos mais camadas convolucionais, os mapas de ativação resultantes capturam características progressivamente mais complexas (SILAPARASETTY, 2020).

A camada de *pooling* tem como finalidade realizar uma subamostragem dos dados de entrada, resultando na redução do tamanho espacial das características convolucionais.

Esse processo é fundamental para diminuir o custo computacional, a utilização de memória e o número de parâmetros, o que, por sua vez, torna o treinamento do modelo mais eficiente (SILAPARASETTY, 2020). Similarmente às camadas convolucionais, cada neurônio na camada de *pooling* está conectado às saídas de um número limitado de neurônios na camada anterior, os quais estão situados dentro de um pequeno campo receptivo retangular (GÉRON, 2019).

Um neurônio de *pooling* desempenha um papel fundamental na redução da dimensionalidade dos dados e na extração de características importantes deles. Ao contrário dos neurônios convolucionais, os neurônios de *pooling* não possuem pesos, em vez disso, eles utilizam funções de agregação, como máximo, mínimo ou média, para combinar as informações das regiões de entrada cobertas por um determinado kernel (GÉRON, 2019).

Entre os métodos de *pooling*, o *max pooling* é amplamente utilizado devido à sua eficácia na preservação das características mais importantes dos dados enquanto reduz sua dimensão. Esta técnica seleciona o valor máximo da região coberta pelo kernel, o que não apenas reduz a dimensionalidade dos dados, mas também atua como um supressor de ruído. Em comparação, o *average pooling* calcula a média dos valores na região do kernel. Embora o *average pooling* possa ser útil em certos contextos, o *max pooling* tende a funcionar melhor, pois preserva características-chave ao selecionar os valores mais altos. Por outro lado, o *min pooling* seleciona o valor mínimo da região coberta pelo kernel, o que pode ser útil em certas aplicações, mas é menos comum em comparação com o *max pooling* e o *average pooling* (SILAPARASETTY, 2020).

Na arquitetura de uma rede neural convolucional, a etapa de classificação é geralmente realizada por camadas totalmente conectadas, como ilustrado na Figura 16. No entanto, uma particularidade das camadas convolucionais e de *pooling* é que elas operam com dados em formato de matriz bidimensional (2D). Portanto, para integrar camadas totalmente conectadas à rede, é necessário converter os dados para um formato unidimensional, e é essa a função da camada *Flatten*. Essa camada realiza o "achatamento" dos dados, transformando a matriz bidimensional em um vetor unidimensional. Essa transformação é essencial para que a rede neural possa processar os dados de maneira eficaz e produzir resultados precisos (SILAPARASETTY, 2020).

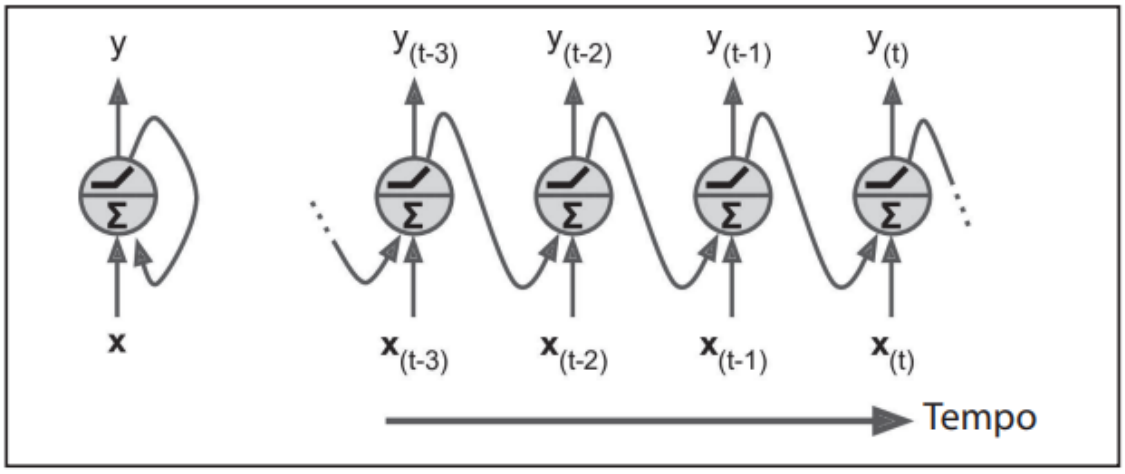
É importante destacar que a função da camada *Flatten* é reorganizar os dados de entrada sem alterar seu conteúdo informativo, apenas ajustando sua representação para torná-la compatível com as camadas subsequentes da rede neural. Em outras palavras, a

camada *Flatten* não modifica os valores dos dados, apenas os reorganiza espacialmente para facilitar o processamento pelas camadas subsequentes da rede.

3.3.3 *Long Short-Term Memory* (LSTM)

Os modelos de redes mencionados anteriormente (MLP e CNN) são tipos de redes neurais *feedforward*, onde as ativações fluem somente em uma direção, da camada de entrada até a camada de saída. Por outro lado, em uma rede neural recorrente, essa dinâmica é diferente, pois também inclui conexões que retroalimentam informações para camadas anteriores (GÉRON, 2019). A forma mais básica de uma RNN consiste em um único neurônio que recebe entradas, gera uma saída e envia essa saída de volta para si mesmo, conforme ilustrado na Figura 18 (esquerda). Em cada intervalo de tempo t (também chamado de *frame*), esse neurônio recorrente recebe tanto as entradas $x(t)$ quanto a sua própria saída do intervalo de tempo anterior.

Figura 18 – Neurônio recorrente



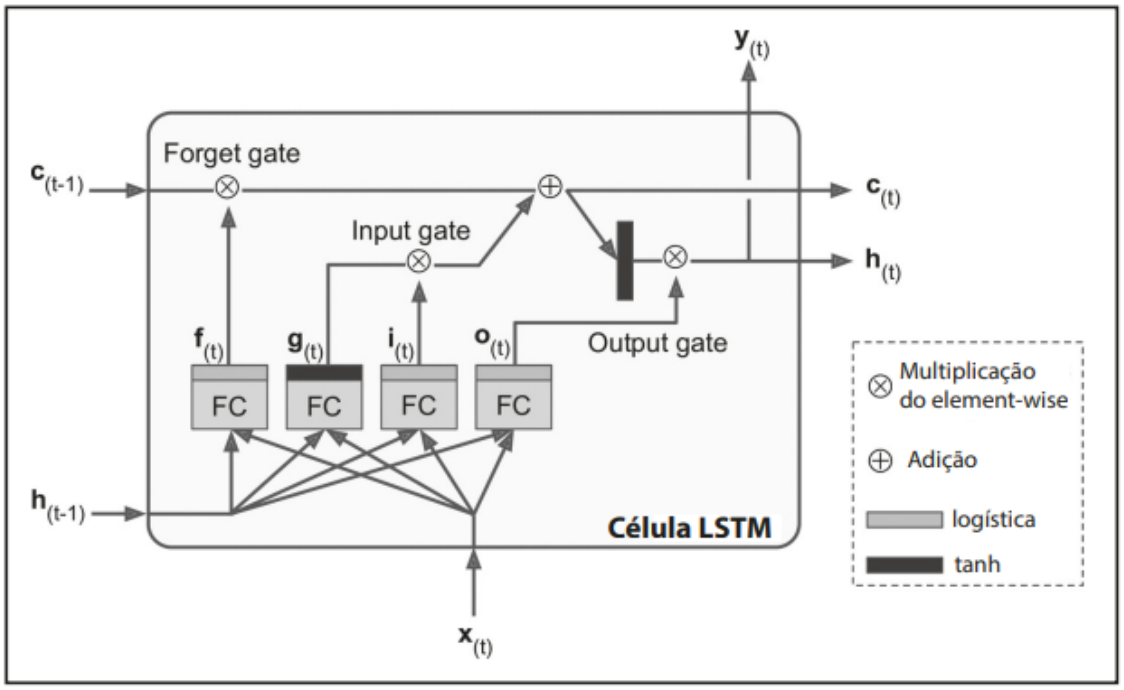
Fonte: Géron (2019)

A natureza recorrente de um neurônio em um intervalo de tempo específico implica que sua saída seja uma função de todas as entradas recebidas nos intervalos de tempo anteriores, conferindo-lhe uma espécie de "memória". Uma componente fundamental de uma rede neural que mantém e utiliza informações ao longo do tempo é chamada de célula de memória. Uma das células de memória mais populares e eficazes é a Memória de Longo Curto Prazo (LSTM - *Long Short-Term Memory*) (GÉRON, 2019).

Uma célula LSTM possui a capacidade de capturar dependências de longo prazo,

mantendo um gradiente de erro constante, o que permite que redes recorrentes aprendam continuamente ao longo de vários passos, capturando relações causais complexas. Esse gradiente de erro é retropropagado ao longo do tempo e das camadas, contribuindo para a aprendizagem eficaz de sequências de dados (SILAPARASETTY, 2020). A Figura 19 apresenta uma ilustração da célula LSTM.

Figura 19 – Célula LSTM



Fonte: Géron (2019)

Conforme ilustrado na figura, a LSTM emprega uma estrutura conhecida como portão (*gate*) para controlar o fluxo de dados dentro e fora de uma célula. Esses portões funcionam de maneira análoga às portas condicionais (AND, OR, XOR, etc.), decidindo com base em um conjunto de condições quais informações permitirão passar pelo portão. Essas condições são determinadas pela própria LSTM e não precisam ser programadas explicitamente (SILAPARASETTY, 2020).

Primeiro, o vetor de entrada atual $x_{(t)}$ e o estado de curto prazo anterior $h_{(t-1)}$ são fornecidos para quatro camadas diferentes totalmente conectadas. Todos servem a um propósito diferente.

A camada principal é aquela que produz $g_{(t)}$ tem como função analisar as entradas atuais $x_{(t)}$ e o estado anterior $h_{(t-1)}$. Na célula LSTM sua saída é parcialmente armazenada no estado de longo prazo.

As três outras camadas funcionam como controladores de portão. Elas empregam

a função de ativação logística, gerando saídas no intervalo de 0 a 1. Essas saídas alimentam operações de multiplicação elemento por elemento, onde valores de 0 fecham o portão e valores de 1 o abrem.

O portão de esquecimento (*forget gate*), controlado por $f_{(t)}$, determina quais partes do estado de longo prazo devem ser esquecidas. O portão de entrada (*input gate*), controlado por $i_{(t)}$ decide quais partes de $g_{(t)}$ devem ser adicionadas ao estado de longo prazo. Por fim, o portão de saída (*output gate*), controlado por $o_{(t)}$, regula quais partes do estado de longo prazo devem ser lidas e exibidas nesse intervalo de tempo, influenciando tanto $h_{(t)}$ quanto $y_{(t)}$ (GÉRON, 2019).

Em resumo, uma célula LSTM desempenha várias funções cruciais no processamento de sequências de dados. Primeiramente, ela é capaz de reconhecer entradas importantes por meio do mecanismo de "portão de entrada"(*input gate*), permitindo que informações relevantes sejam incorporadas ao estado de memória de longo prazo. Em seguida, o mecanismo de "portão de esquecimento"(*forget gate*) permite que a célula decida quais informações devem ser retidas no estado de memória e quais devem ser descartadas, garantindo a retenção apenas das informações essenciais ao longo do tempo. Por fim, a célula LSTM é capaz de acessar e utilizar as informações armazenadas no estado de memória quando necessário, facilitando a extração de padrões relevantes em dados sequenciais. Essas características explicam por que as LSTMs têm sido tão eficazes na captura de padrões de longo prazo em uma variedade de aplicações, incluindo análise de séries temporais, processamento de texto, reconhecimento de fala e muito mais (GÉRON, 2019).

4 TRABALHOS RELACIONADOS

Após o levantamento de referencial teórico, feito por meio da revisão de escopo da literatura, é necessário uma discussão sobre o estado da arte e uma comparação com a proposta de pesquisa apresentada. Para a realização desta comparação, foram selecionados os trabalhos encontrados na literatura e que estão relacionados com a proposta em desenvolvimento. Os trabalhos foram selecionados com base nas três abordagens que constituem o domínio do problema e a intenção de desenvolvimento desta dissertação: transporte de animais, análise de áudio e análise de estresse.

Analisando a literatura com base no transporte de animais, Strappini et al. (2013) investigaram os principais motivos de lesões e hematomas em bovinos, relacionando com as condições de transporte, embarque e desembarque, a partir de câmeras de espectros visível e infravermelho instaladas no caminhão. Com uma abordagem mais subjetiva, Hoffman e Lühl (2012), Minka e Ayo (2007) utilizaram formulários para a coleta de dados e análise das principais causas de estresse e lesões de bovinos durante o manejo e transporte.

Levando em consideração a perspectiva da análise de áudio, na literatura é possível encontrar diferentes abordagens. Meen et al. (2015) investigaram a viabilidade de avaliação do bem-estar de bovinos através da análise dos sons emitidos, identificando diferentes tipos de comportamentos relacionados com a ruminação, alimentação, interação social, interação sexual, estresse e remanescente. Nesta linha de pensamento, Jahns (2008), Deshmukh et al. (2012) abordaram a análise de som para identificar as condições dos animais, com base nas formas de comunicação realizadas pelos animais. Outra abordagem para a análise de som foi realizada nos trabalhos de Chung et al. (2013), Röttgen et al. (2020), Lee et al. (2014), Yeon et al. (2006), onde o objetivo foi monitorar e identificar a ocorrência do cio em bovinos. Também foram encontrados na literatura trabalhos que utilizam a análise de áudio para o monitoramento de comportamento ingestivo de bovinos (CHELOTTI et al., 2016; CLAPHAM et al., 2011). Já nos trabalhos de (TORRE et al., 2015; YAJUVENDRA et al., 2013) o objetivo foi investigar a estrutura acústica e as características das vocalizações de bovinos através de análises estatísticas nos domínios da amplitude e frequência dos sinais.

Quanto ao estresse animal, Moura et al. (2008), Manteuffel e Schön (2002) realizaram um estudo sobre a identificação de estresse em porcos através da análise de vocalizações. Nessa mesma linha, Lee et al. (2015) analisou e desenvolveu um sistema de

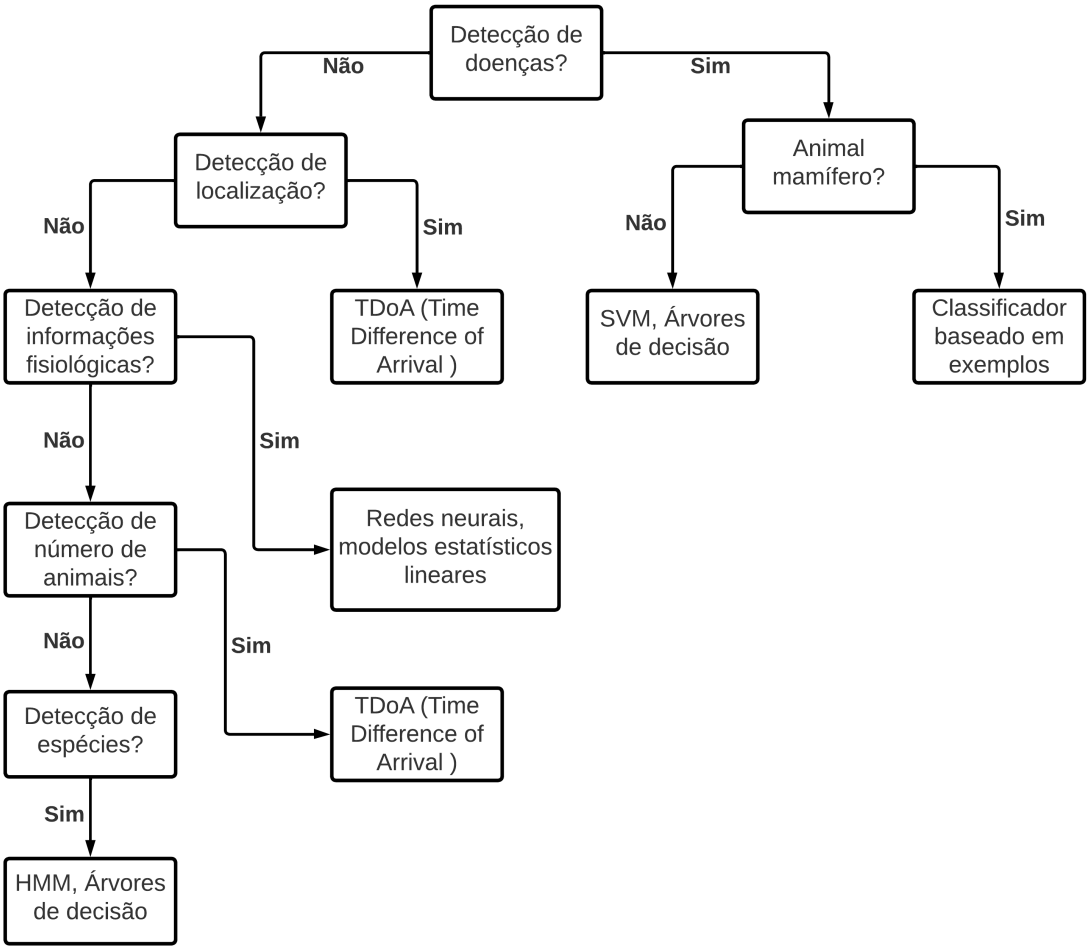
alerta de estresse com aves. Para a análise de estresse com bovinos, Ikeda e Ishii (2008), Gavojdian et al. (2023) buscaram identificar as mudanças nas vocalizações de vacas durante condições psicologicamente estressantes. No que diz respeito à aplicação de redes neurais na análise de vocalizações de animais, o estudo conduzido por Vidana-Vila et al. (2023) buscou desenvolver um detector automático de vocalizações de bovinos por meio de redes neurais convolucionais (CNN - *Convolutional Neural Network*). Já nas pesquisas de Jung et al. (2021), Pandeya et al. (2022) e Sattar (2022) as técnicas de MFCC (*mel frequency cepstrum coefficients*) foram empregadas em conjunto com redes neurais convolucionais para a classificação de diferentes comportamentos de bovinos com base em suas vocalizações. Da mesma forma, Şaşmaz e Tek (2018) realizaram a classificação de vocalizações de diferentes animais utilizando as técnicas de MFCC e redes neurais convolucionais. No âmbito de estudos que empregaram redes neurais recorrentes (RNN - *Recurrent Neural Network*), merecem menção os trabalhos de Huang et al. (2021), Shorten (2023) e Duan et al. (2021). O primeiro concentrou-se na detecção do comportamento alimentar de aves, o segundo teve como objetivo a identificação dos padrões de comportamento de vacas, enquanto o terceiro analisou o comportamento ruminante de ovelhas.

Com o intuito de explorar a literatura através de uma revisão bibliográfica e examinar as principais técnicas de análise acústica aplicadas ao monitoramento de animais, McLoughlin, Stewart e McElligott (2019) compilaram diferentes métodos automatizados com potencial para inferir sobre o estado de bem-estar animal. No contexto específico de bovinos, os autores identificaram que as análises estatísticas nos domínios da amplitude e frequência são amplamente empregadas no estudo das vocalizações, assim como é disseminada a utilização de *deep learning*, com o predomínio no uso de redes neurais. A Figura 20 apresenta, de forma esquemática, por meio de uma árvore de decisão, a síntese realizada pelos autores. Essas abordagens abrem perspectivas promissoras para a compreensão e monitoramento do bem-estar animal, evidenciando o avanço significativo proporcionado pela aplicação de métodos automatizados na análise acústica.

Na Figura 20, pode-se observar que, de acordo com McLoughlin, Stewart e McElligott (2019), os principais métodos automatizados empregados em estudos bioacústicos para análise de informações fisiológicas são as redes neurais e os modelos estatísticos. O estudo do estresse animal se enquadra nessa categoria, o que reforça a escolha adequada dos métodos utilizados para desenvolver a presente pesquisa.

Para descrever de maneira mais detalhada os trabalhos encontrados na literatura e

Figura 20 – Principais métodos utilizados em estudos bioacústicos.



Fonte: Adaptado de McLoughlin, Stewart e McElligott (2019)

proporcionar uma visão mais abrangente e aprofundada sobre as pesquisas que serviram como referência do presente trabalho, a seguir serão apresentados os trabalhos que apresentam maior afinidade com o escopo e os objetivos da pesquisa desenvolvida. Os critérios adotados para a seleção desses estudos consideraram os dados utilizados e suas metodologias de coleta, as técnicas analíticas e computacionais empregadas, e os resultados alcançados.

O trabalho de Chung et al. (2013) teve como objetivo desenvolver um sistema capaz de detectar a ocorrência de cio de vacas nativas (*Bos Taurus coreanea*) através da análise dos sons emitidos. Na análise de áudio foi utilizado o algoritmo MFCC. Para o treinamento e classificação dos sons foram utilizados os algoritmos CFS (*Correlation-based Feature Selection*) para a definição do melhor conjunto de características do som para o reconhecimento e classificação de padrões, e SVDD

(*Support vector data description*) para a classificação dos sons, uma vez em que esse algoritmo é eficiente em problemas de classificação de uma classe. Para a realização da validação do sistema foram utilizadas 100 vocalizações de animais no cio e 180 vocalizações normais. Foram analisadas dois conjuntos de propriedades dos sons, no qual o primeiro é constituído por todas as propriedades dos sinais (360) e alcançou uma taxa de detecção de cio de 89,4%. O segundo conjunto, formado pelas propriedades definidas pelo algoritmo CFS, que reduziu as propriedades do som para 62, alcançado uma taxa de detecção de cio de 83,6%. Embora tenha apresentado uma taxa de detecção menor, com a utilização do algoritmo CFS se obteve uma redução no consumo de memória, e uma melhora no tempo de execução em cerca de 5 vezes.

O objetivo do trabalho de Meen et al. (2015) foi estudar o bem-estar animal a partir da identificação do comportamento dos animais através da análise conjunta de som e imagem. Foram utilizadas 4 câmeras e 4 microfones para captar os sons e imagens de 141 animais da raça *Holstein Friesian*. As imagens e sons coletados foram analisados por especialistas em comportamento animal, que realizaram a detecção manual do comportamento dos animais, e após isso, a detecção do comportamento foi realizada pelos próprios autores. Os sons que eram visivelmente atribuíveis a um animal foram classificados em seis distintos grupos de comportamento: deitado e ruminação, alimentação, interação social, interação sexual, estresse e comportamento remanescente. No total, 836 sons foram relacionados com algum comportamento, sendo 228 estresse, 252 interação social, 167 alimentação, 27 deitado e ruminação, 4 interação sexual e 158 remanescente. A métrica utilizada no estudo foi a frequência máxima média e realizada a análise de variância (teste de Fisher, ou LSD) entre as frequências dos sons para a classificação dos comportamentos dos animais. Foram utilizados os *softwares* Ultra Vox 3.0 e SPSS Statistics 21.0 para a determinação da frequência máxima dos sons e análises estatísticas, respectivamente. Os resultados obtidos apresentaram uma menor frequência máxima média para comportamentos que indicam um melhor nível de bem-estar, como o comportamento deitado e ruminação. Por outro lado, os comportamentos classificados como interação sexual e estresse apresentaram a maior frequência máxima média em relação aos demais comportamentos.

O objetivo do trabalho de Moura et al. (2008) foi desenvolver um software para o monitoramento de estresse de leitões com base no sons emitidos pelos animais. Para a validação do software, foram realizados dois testes. O primeiro teste, de controle, consistiu em reproduzir uma condição estressante para os animais, onde os leitões foram

pegos do curral e ligeiramente agitados. Foi considerada uma vocalização normal, ou não estressante, quando os leitões estavam em grupos. A vocalização registrada nos testes foram divididos em três tipos: alerta, apreensão e contenção. O segundo teste foi realizado com 10 leitões onde cada um foi pego do curral e cuidadosamente colocado em uma câmara. Os sons emitidos pelos animais foram captados por microfones elétricos posicionados no centro do curral. Para cada tipo de vocalização, foi realizada uma análise a fim de comparar os sons, usando os valores médios e seus respectivos intervalos de confiança, utilizando testes de análise de comparações múltiplas de Tukey. Os testes identificaram 12 sons de condição de alerta, 7 em condição de apreensão e 10 em condição de contenção. As vocalizações do estado de contenção apresentaram maior amplitude e menor frequência sonora, enquanto o estado de alerta apresentou menor amplitude e maior frequência sonora. O teste de comparações múltiplas de Tukey apresentou diferença significativa entre as médias das amplitudes dos três tipos de vocalização. Para a comparação entre os testes que reproduziram condições estressantes nos leitões, foi observado um menor pico de frequência nos sons emitidos.

O trabalho de Deshmukh et al. (2012) teve como objetivo a utilização de técnicas de análise de padrões de vocalização para classificar e identificar o estado animal. O estudo foi realizado em duas etapas, na primeira etapa foram colocados 60 animais (30 búfalos e 30 vacas), sendo 10 bezerros, 10 novilhos e 10 adultos de cada raça, em recintos individuais com microfones posicionados no teto. Os dados coletados foram manipulados para que ao menos 50 vocalizações de cada animal fossem captadas, em três diferentes condições: isolamento normal (quando o animal estava isolado no recinto), estado de calor (induzido por substâncias químicas), e ordenha (com duração entre 1 e 2 horas). No total foram coletadas 5873 amostras de som, cada uma com duração entre 2 e 8 segundos. Os dados coletados foram analisados a partir das suas características espectro-temporais para cada tipo de condição. Na condição de ordenha, foi observado que o espectro estava concentrado em regiões de baixa frequência e se manteve estável durante toda a vocalização, com a quantidade de aperiodicidade[1] mínima. Para as condições de estado de calor e isolamento normal a aperiodicidade foi maior. Para a condição de isolamento normal, a maior proporção do espectro está concentrada em regiões de alta frequência, com maior aperiodicidade e menor estrutura harmônica se comparado às outras duas condições. Na análise do espectro, para diferenciar as vocalizações dos três tipos de condições foram utilizados os seguintes parâmetros: harmonia, proporção periódica,

[1]Grau de irregularidade do sinal sonoro (estrutura similar ao ruído).

passo médio, declive espectral, estabilidade espectral e proporção estável. Para a tarefa de classificação das amostras foi utilizado o algoritmo MFCC. Na segunda etapa do estudo foi realizado com 10 animais de cada raça divididos entre bezerros, novilhos e adultos, com a finalidade de identificar os animais a partir de suas vocalizações, como também para identificar o estado de cada animal. Foram utilizados 85% dos dados para o treinamento e 15% para o teste do algoritmo. Os resultados apontam para uma acurácia entre 62-79% na identificação dos animais (bezerro, novilho ou adulto) e uma acurácia entre 71-92% para a identificação das condições (isolamento normal, estado de calor ou ordenha).

Lee et al. (2014) desenvolveram um algoritmo para determinar o melhor conjunto de formantes do espectro para a utilização na detecção do cio em vacas (*Bos Taurus coreanea*) através da vocalização dos animais. No total, foram registrados 163 sons de cio e 140 sons normais para a utilização na detecção do cio dos animais. O algoritmo desenvolvido tem como entrada o conjunto de todos os formantes da amostra de som (F1-F19) e sua saída é um subconjunto otimizado de formantes para ser utilizado no etapa de identificação de sons de cio. Para a detecção do cio, foi utilizado o algoritmo baseado em *machine learning AdaBoost.M1*. Para validar o algoritmo desenvolvido, foram avaliados três conjuntos de formantes de espectro na identificação do cio. O primeiro conjunto foi composto pelos principais formantes utilizados na literatura para a classificação de sons (F1, F2, F3 e F4). O segundo conjunto foi composto por todos os formantes do espectro (F1-F19). O terceiro conjunto foi composto pelos formantes definidos pelo algoritmo desenvolvido pelos autores (F1, F2, F4, F7, F14 e F19). Como resultados, o conjunto composto pelos formantes F1, F2, F3 e F4 apresentou uma taxa de detecção de cio de 65,3%, taxa de falso positivo de 20% e taxa de falso negativo de 14,7%. O segundo conjunto, formado por todos os formantes do espectro (F1-F19) alcançou uma taxa de detecção de cio de 88%, taxa de falso positivo de 5,7% e taxa de falso negativo de 6,3%. O terceiro conjunto, composto pelos formantes definidos pelo algoritmo (F1, F2, F4, F7, F14 e F19) alcançou uma taxa de detecção de cio de 92,5%, taxa de falso positivo de 5% e taxa de falso negativo de 2,5%.

O objetivo do trabalho de Jahns (2008) foi desenvolver um software para a identificação e tradução das vocalizações realizadas por vacas. O sistema desenvolvido funciona da seguinte maneira: as vocalizações realizadas pelos animais são captadas, amplificadas e digitalizadas. A partir do processamento digital dos sinais, um decodificador de sons produz uma sequência de sinais que representam as vocalizações dos animais. Por fim, um interpretador transcreve os sinais processados em um texto

reconhecível por humanos, e que representa o estado emocional, ou condição do animal. Para a validação do software, foram coletadas 688 vocalizações de animais, onde 70% foi utilizado para o treinamento do sistema e 30% (210) foi usado para testar a solução. Para o reconhecimento dos sons foram utilizados os algoritmos MFCC e HMM (*Hidden Markov Models*). Na etapa de treino do sistema, cada um dos padrões foi construído a partir de vocalizações de mesmo significado, o que resulta em padrões não sensíveis à variações nas vocalizações individuais dos animais. Na etapa de reconhecimento, as vocalizações desconhecidas são pré-processadas do mesmo modo que são processadas as vocalizações da etapa de treino, e após isso, são comparadas com os padrões de referência armazenados. O sistema determina a qual dos padrões de referência armazenados a vocalização desconhecida melhor se assemelha e atribui a ela o estado emocional, ou condição do animal. Nos testes, foram avaliados sete estados emocionais e condições dos animais para a realização do reconhecimento de som (vacas chamando bezerros, bezerros chamando vacas, tossindo, cio, ordenha, inalação ruidosa e fome). Os resultados apresentaram, respectivamente, 94%, 100%, 93%, 88%, 74%, 91% e 100% de taxa de reconhecimento de padrão.

Manteuffel e Schön (2002) buscaram desenvolver um sistema para monitorar e registrar a quantidade de vocalizações de estresse de porcos para ser aplicado em ambientes de criação, transporte e abate. Foram aplicadas a 30 porcos em situações de estresse para a coleta de vocalizações. Nos leitões, foi realizada a imobilização por contenção no tórax, mantendo-os acima do chão. Nos porcos em fase de crescimento foi realizada a imobilização forçando-os na região do dorso. Nos porcos adultos a imobilização foi feita com laços no focinho. Grunhidos dos animais, bem como outros tipos de ruídos que ocorrem nas instalações dos porcos, em condições normais, foram coletados como sons não estressantes. Para a etapa de reconhecimento os sons coletados foi utilizado o LPC (*Linear Prediction Coding*) em conjunto de redes neurais. Foram utilizados 12 coeficientes de LPC, que equivalem às 6 primeiras frequências de ressonâncias, calculados em janelas de tempo de 46,44 ms de duração. A rede neural possui 193 neurônios em 4 camadas, onde apenas as entradas classificadas como sons de estresse são encaminhadas para a saída. Para a validação do sistema, foi analisada a ocorrência de estresse em porcos em dois diferentes regimes de alimentação. O primeiro regime foi realizado com uma relação de porcos/local de alimentação de 6:1 com 24 porcos de engorda. O segundo regime foi realizado com uma relação de porcos/local de alimentação de 1:1 com 11 porcos. Foram registrados os sons, a partir de microfones

posicionados à 2 metros do chão, no centro do curral. Em paralelo, um sistema de vídeo registrou o comportamento e a vocalização dos animais. Como resultados, o sistema desenvolvido detectou uma maior ocorrência de vocalizações de estresse no primeiro regime de alimentação, onde havia competição dos animais pelo alimento. Para o segundo regime, o sistema detectou poucas ocorrências de vocalizações de estresse, ocasionadas por brigas casuais dos animais.

Lee et al. (2015) procuraram implementar um sistema de monitoramento automático para a detecção e alerta de estresse de aves em instalações avícolas comerciais. O experimento foi conduzido com 120 galinhas divididas em 4 grupos com 15 repetições, onde cada repetição possuía dois animais por cela. Cada grupo foi exposto ao estresse físico a partir da mudança de temperatura na cela (10ºC$\pm$ 2, 21ºC$\pm$ 2 e 34ºC$\pm$ 2), além disso um grupo foi exposto ao estresse mental (medo), atingido a partir de pancadas na cela com uma vara, em temperatura ambiente. Os sons emitidos pelos animais durante os experimentos foram registrados por uma filmadora digital posicionada no interior das celas. Os sons foram classificados a partir de rotulação manual, baseada na análise acústica combinada com a análise espectral visual de toda gravação. Foram rotulados todos os sons que puderam ser identificados como vocalização, com base na observação visual do espectrograma e confirmação auditiva feita pelo operador. Para a classificação e reconhecimento de som, as características analisadas dos sons foram definidas a partir das principais características utilizadas na literatura. No domínio do tempo as características analisadas foram a *Root mean square* (RMS), a potência, a energia, o extremo absoluto, a intensidade, a proporção de ruído harmônico, o *jitter*, o *shimmer* e a frequência fundamental. No domínio da frequência, as características analisadas foram as formantes do espectro (F1-F9) e a densidade espectral (PSD) do PSD1 ao PSD39. Para a definição do melhor subconjunto de características para o reconhecimento de padrão, foi utilizado o algoritmo CFS, que considera a utilidade individual de cada característica para a predição de classes, em conjunto com o nível de inter-correlação entre elas. Para a tarefa de detecção de estresse foi utilizado o algoritmo SVM (*Support Vector Machine*) hierárquico multi-classe. Para a validação do sistema foram utilizadas 407 amostras de som de experimentos de temperatura induzida, sendo 149 à 10ºC$\pm$ 2, e 258 à 34ºC$\pm$ 2. Também foram utilizadas 114 amostras de som de experimentos de medo induzido, além de 136 amostras de som de condições normais. O subconjunto de características definido pelo algoritmo CFS foi os formantes F1 e F3, RMS, *pitch* médio, *pitch* máximo, *shimmer*, *jitter* e PSD38, reduzindo o número de características analisadas para o reconhecimento

de padrão de 54 para 8. Os resultados apresentaram uma taxa de detecção de estresse de 86,6%, taxa de falso positivo de 9,6% e taxa de falso negativo de 3,8%.

O objetivo do trabalho de Ikeda e Ishii (2008) foi analisar e identificar a mudança na vocalização de vacas (*Japanese Black*) sob duas condições psicologicamente estressantes, sendo a primeira a fome e a segunda a separação de seu bezerro durante o desmame. Os experimentos foram realizados com os mesmos animais sob as duas condições. A condição de fome foi induzida retardando o horário de alimentação habitual dos animais. A condição de separação foi analisada durante a fase de desmama dos bezerros, onde mãe e filhote são alojados em ambientes separados. No total, foram registrados 51 vocalizações de fome e 283 vocalizações de separação, dos quais 28 sons de fome e 181 sons de separação foram excluídos devido à sobreposição por outros sons ou saturação dos sinais. Na análise de som, foi considerado a densidade espectral (*power espectrum*), utilizando a transformada rápida de Fourier (FFT) e o algoritmo LPC. As seis mais baixas frequências ressonantes foram escolhidas como variáveis para a análise discriminante linear, devido que nessas frequências se concentraram a maior densidade do espectro. Os resultados obtidos na condição de fome apresentaram maiores frequências médias e coeficientes de variação, para todas as frequências ressonantes analisadas, se comparado aos resultados obtidos para a condição de separação. O estudo também identificou que para as ambas condições as frequências ressonantes eram sobretons harmônicos, nas quais frequências ressonantes de ordem maiores eram múltiplos da frequência ressonante de primeira ordem (frequência fundamental).

Yeon et al. (2006) analisaram a vocalização de vacas (*Bos taurus coreana*) para identificar as diferenças entre os estados de cio e alimentação antecipada. Para a realização do experimento foram usados 26 animais, nos quais 8 estavam no cio. Os sons foram coletados com um filmadora localizada 1 metro do curral. Para o estado de alimentação antecipada, foram coletados os sons no período da manhã, quando os animais não estavam esperando receber alimentação. A coleta de sons no estado de cio foi realizada com animais nos quais especialistas em comportamento animal identificaram o cio. As análises estatísticas foram realizadas utilizando o *software* SPSS 9.0. Foram feitas análises multivariadas da variância para comparar as propriedades acústicas das vocalizações de estado de cio e alimentação antecipada, usando o procedimento de modelo linear geral (GLM). As análises estatísticas foram realizadas com um nível de confiança de 95% utilizando os métodos de Tukey-Kramer e testes de comparações múltiplas de Bonferroni. Os parâmetros analisados no domínio da frequência foram a

duração da vocalização, intensidade, tom e os quatro primeiros formantes do espectro (F1-F4). Para a validação do estudo foram analisadas 441 vocalizações dos 26 animais utilizados no experimento, sendo 146 vocalizações de cio e 295 vocalizações de alimentação antecipada. Os resultados apresentaram duração e intensidade de 1,85±0,79 s e 71,4±6,2 dB, e 1,86±0,44 s e 68,8±4,2 dB para as vocalizações de alimentação antecipada e cio, respectivamente. Na análise do espectrograma, a condição de cio não apresentou uma tonalidade clara na vocalização, por esse motivo esse parâmetro foi desconsiderado na análise. Para as formantes do espectro, o primeiro formante apresentou maior frequência na condição de cio, com uma frequência de 832±150 Hz em comparação com a frequência de 784±127 Hz da condição de alimentação antecipada. Para os outros formantes do espectro analisadas (F2, F3 e F4), a condição de alimentação antecipada apresentou maior média nas frequências se comparado à condição de cio. Para a classificação das condições dos animais, os melhores resultados por análise discriminante foram alcançado com os parâmetros de duração, intensidade e formantes onde a classificação incorreta foi de 13,8%, com 86,2% dos animais sendo atribuídos ao grupo correto.

O estudo conduzido por Torre et al. (2015) teve como objetivo a investigação e caracterização das vocalizações entre vacas e bezerros durante a comunicação mãe-filho. A pesquisa envolveu a análise de 344 vocalizações, das quais 205 eram de vacas e 139 de bezerros, provenientes de 31 animais (17 vacas e 14 bezerros). A coleta de dados foi realizada utilizando microfones direcionais, posicionados a uma distância de 8 a 30 metros dos animais, em um ambiente livre de manejos ou interações estressantes. As vocalizações foram registradas com uma taxa de amostragem de 44,1 kHz, no formato WAV (*Waveform Audio Format*), com resolução de amplitude de 16 bits. A análise acústica compreendeu a aplicação de técnicas estatísticas (ANOVA, ANCOVA, MANOVA) sobre os parâmetros de F0 (frequência fundamental), *jitter*, *shimmer*, frequências F1-F8, amplitude e duração das vocalizações. Os resultados destacaram diferenças significativas entre as vocalizações de vacas e bezerros, sobretudo na análise de frequência, onde para as vacas foram obtidas as médias para F0 de 81,1±0,9 e para os formantes F1-F8 de 228,3±1,8 até 3181±2,6 Hz, enquanto para os bezerros foram obtidas as médias para F0 de 142,8±1,8, e para os formantes F1-F8 de 391,7±5,3 até 5813±68,7 Hz.

O estudo realizado por Gavojdian et al. (2023) investigou as vocalizações de baixas frequências (LFC - *Low Frequency Calls*) e altas frequências (HFC - High

Frequency Calls) em vacas da raça *Romanian Holstein*. As vocalizações LFC são consideradas indicadores de baixo estresse e emoções positivas, ao passo que as HFC são associadas a estados afetivos negativos. A coleta de dados foi conduzida em um grupo homogêneo de 20 animais em confinamento, utilizando microfones direcionais a uma distância de 6 metros dos animais. As vocalizações registradas foram armazenadas no formato WAV, com uma taxa de amostragem de 44,1 kHz e resolução de amplitude de 16 bits. No total, foram coletadas 1144 vocalizações, sendo 952 HFC e 192 LFC. Para a análise acústica, foram adotadas duas abordagens. A primeira utilizou uma arquitetura de rede neural que combina uma CNN 2D e uma RNN com a unidade GRU (*Gated Recurrent Unit*), conforme proposto por Ye e Yang (2021). A segunda abordagem envolveu análise estatística de parâmetros no domínio da frequência, como F0 e os formantes F1-F8. Ambas abordagens foram avaliadas através de validação cruzada (*fold cross-validation*), com k = 5. Na arquitetura da rede neural, a CNN 2D foi empregada na extração das principais características dos espectrogramas de áudio, enquanto a RNN foi utilizada para o aprendizado das características extraídas relacionadas às vocalizações. Na tarefa de distinguir entre LFC e HFC, os resultados revelaram que o modelo estatístico obteve uma acurácia de 87,2$\pm$4,1%, enquanto o modelo baseado em redes neurais atingiu uma acurácia de 89,4$\pm$3,8%. Na classificação individual dos animais, o modelo estatístico alcançou uma precisão de 68,9$\pm$5,1%, ao passo que a rede neural apresentou uma acurácia de 72,5$\pm$4,7%.

Jung et al. (2021) desenvolveram um classificador para quatro tipos de vocalizações de bovinos da raça *Bos taurus coreanae*, sendo elas: cio, alimentação antecipada, tosse e normal. A coleta de dados foi realizada por sensores de monitoramento, os quais registraram os sons por meio de um microcontrolador Raspberry Pi 3+ e microfones posicionados a aproximadamente 3 metros acima dos celeiros onde os animais estavam confinados. O microcontrolador desempenhou a função de filtrar os sons coletados, armazenando apenas aqueles com amplitude superior a 60 dB, após passarem por uma etapa de filtragem de ruídos. Posteriormente, os sons foram rotulados manualmente com base em análises de vídeo dos celeiros, onde as vocalizações do tipo normal foram atribuídas quando não foi possível rotular como cio, alimentação antecipada ou tosse. No total, foram coletadas 897 vocalizações, distribuídas entre 207 vocalizações de cio, 178 de alimentação antecipada, 56 de tosse e 456 vocalizações normais. O algoritmo MFCC foi utilizado para a extração de características dos sons, servindo como entrada para o modelo de classificação. O modelo de classificação comportamental de voz

dos bovinos empregou uma arquitetura de rede neural convolucional 2D, sendo 80% dos dados destinados ao treinamento e 20% para teste. Os resultados revelaram que o modelo alcançou acurácias de 74% na classificação de vocalizações de alimentação antecipada, 76% para vocalizações normais, 86% para vocalizações de tosse e 95% para vocalizações de cio.

O estudo conduzido por Pandeya et al. (2022) descreve o desenvolvimento de uma ferramenta semiautomática de rotulação de eventos de vocalizações em bovinos. A abordagem proposta utiliza dados de áudio ou vídeo como entrada e, por meio da análise espectral de áudio, sugere automaticamente possíveis momentos de vocalizações aos usuários. Inicialmente, foram coletados dados reais em estábulos e na internet, além de dados sintéticos gerados automaticamente, abrangendo quatro categorias de eventos: vocalizações de vacas, bezerros, touros e vocalizações mistas, caracterizadas por sobreposição de vocalizações. O conjunto de dados combinado, composto por 47.400 vocalizações, distribuiu-se entre 2.520 vocalizações de touros, 2.864 de bezerros, 2.916 de vacas e 2.968 vocalizações mistas. Para a tarefa de classificação, foi empregada uma rede neural FRCNN (*Faster Region-based Convolutional Neural Network*). A arquitetura da rede foi treinada especificamente para detectar eventos sonoros no eixo temporal, mantendo o eixo de frequência constante. A implementação do *software* foi realizada em linguagem *Python*, utilizando as bibliotecas *librosa*, *pyaudio* e *matplotlib* para extrair o espectrograma Mel das entradas de áudio, padronizado em 32 x 320, servindo como entrada para a rede neural. Os resultados obtidos demonstram uma precisão significativa na classificação de eventos de vocalizações, atingindo 89,4% para vocalizações de touros, 82,7% para vocalizações de bezerros, 85,3% para vocalizações de vacas e 65,8% para vocalizações mistas.

A Figura 21 apresenta uma síntese dos trabalhos correlatos, onde se ressalta o foco de aplicação, a proveniência dos dados utilizados, o tipo de análise realizada e os principais resultados obtidos dos trabalhos. Dentre as principais características dos trabalhos é possível destacar o foco, o qual está compreendido em identificação de cio, análise de bem-estar, análise de comportamento e classificação de vocalizações. Quanto às técnicas utilizadas no desenvolvimento dos trabalhos, é possível destacar a utilização do algoritmo MFCC para a análise e processamento de áudio, algoritmos de aprendizado de máquina para a classificação dos sons, e análises de variância para comparação das propriedades e características acústicas das vocalizações.

Em comparação entre a presente pesquisa e os trabalhos correlatos apresentados,

Figura 21 – Análise comparativa de trabalhos correlatos

Aplicação	Proveniência dos dados	Tipo de análise	Principais resultados	Trabalhos
Identificação de cio	Coleta de vocalizações de condição normal e de cio	Análise de som (MFCC, CFS e SVDD)	Taxa de detecção de cio de 83,6%	(CHUNG et al., 2013)
	Coleta de vocalizações de condição normal e de cio	Análise de som (AdaBoost.M1)	Taxa de detecção de cio de 92,5%	(LEE et al., 2014)
	Coleta de vocalizações de condição normal e de cio	Análise de variância (Tukey-Kramer e testes de comparações múltiplas de Bonferroni)	Menor frequência média em condição de cio	(YEON et al., 2006)
Análise de bem-estar	Coleta de vocalizações em condições normais e estressantes	Análise de variância (testes de comparações múltiplas de Tukey)	Menor pico de frequência em condições de estresse	(MOURA et al., 2008)
	Coleta de vocalizações em condições normais e estressantes	Análise de som (LPC e redes neurais)	Maior incidência de estresse em regime de disputa por alimento	(MANTEUFFEL; SCHÖN, 2002)
	Coleta de vocalizações de condições normais e estressantes	Análise de som (CFS e SVM)	Taxa de detecção de estresse de 86,6%	(LEE et al., 2015)
	Coleta de vocalizações em condições estressantes	Análise de som (LPC)	Maiores frequências médias em condições de estresse	(IKEDA; ISHII, 2008)
Análise de comportamento	Coleta de vocalizações de animais	Análise de som (MFCC e HMM)	Taxa de identificação de estado entre 74-100%	(JAHNS, 2008)
	Coleta de vocalizações em diferentes estados	Análise de som (MFCC)	Taxa de identificação de estado entre 71-92%	(DESHMUKH et al., 2012)
	Coleta de áudio e vídeo de animais em campo	Análise de variância (teste de Fischer)	Maior frequência máxima média em situações de estresse	(MEEN et al., 2015)
	Coleta de vocalizações de animais em confinamento	Análise de som (MFCC) e redes neurais (CNN)	Taxa de classificação de estado entre 74-95%	(JUNG et al., 2021)
Classificação de vocalizações	Coleta de vocalizações em ambiente livre	Análise de variância (ANOVA, ANCOVA)	Diferenças significativas entre vocalizações de vacas e bezerros	(TORRE et al., 2015)
	Coleta de vocalizações de animais em confinamento	Redes neurais (RNN e CNN) e análise estatística	Taxa de classificação de vocalizações entre 68-87%	(GAVOJDIAN et al., 2023)
	Vocalizações reais e sintéticas	Análise de som (espectro Mel) e redes neurais (FRCNN)	Taxa de classificação de vocalizações entre 65-89%	(PANDEYA et al., 2022)

Fonte: Autor (2024)

é possível destacar algumas similaridades: a utilização de técnicas de processamento de áudio, uso de redes neurais, análise estatística nos domínios da frequência e amplitude, classificação e identificação de estresse. Contudo, a pesquisa desenvolvida se destaca principalmente no objetivo de realizar um estudo comparativo entre diferentes arquiteturas de redes neurais para um modelo de classificação capaz identificar situações de estresse, além de conduzir uma análise estatística sobre parâmetros nos domínios da frequência e amplitude para avaliar os principais fatores relacionados ao estado de estresse animal.

A principal contribuição desta pesquisa para o estado da arte está na oportunidade

de aprimorar a compreensão dos aspectos que interferem no nível de bem-estar dos animais, bem como também contribuir no avanço do conhecimento sobre o desempenho de diferentes arquiteturas de redes neurais na tarefa de classificação de estresse animal. Embora as variáveis e técnicas avaliadas neste trabalho já tenham sido investigadas de forma independente, ou em revisões bibliográficas, até o momento não havia sido realizado um estudo prático que contemplasse todas essas variáveis e técnicas de forma simultânea.

5 PROJETO E DESENVOLVIMENTO

5.1 Equipamento de coleta de dados de transporte

O equipamento de coleta de dados de transporte bovino construído não chegou a ser embarcado para coletar dados sensoriais, de imagem e som dentro dos caminhões de transporte. Contudo, ele foi projetado e construído, sendo também um resultados deste trabalho de dissertação. As características do produto são apresentadas nas seções seguintes.

5.1.1 Requisitos e projeto lógico do produto

O desenvolvimento deste trabalho, conforme apresentado na Seção 1.2, teve como finalidade a construção de um equipamento para a identificação e alerta de estresse animal durante o transporte de bovinos entre fazendas e frigoríficos. Resumidamente, o sistema coleta dados ambientais, imagens e sons emitidos pelos animais durante as etapas de embarque, deslocamento e desembarque. Os dados são armazenados e os sons obtidos seriam usados na construção do sistema de inferência para a identificação de estresse animal e alerta ao condutor do caminhão por meio de um aplicativo para dispositivos móveis. A Tabela 5 apresenta os requisitos do sistema.

Como maneira de exemplificar a proposta, a Figura 22 apresenta uma visão pictórica dos elementos pertencentes ao produto desenvolvido. O produto tem um módulo acoplado na carroceria do caminhão, composto por câmeras, microfones, sensores e um componente de armazenamento e processamento, alimentado a partir da bateria do próprio caminhão. O módulo de alerta de estresse animal também é apresentado, fazendo uso de *Bluetooth* para a transferência dos dados entre o sistema acoplado na carroceria do caminhão e o dispositivo móvel do motorista.

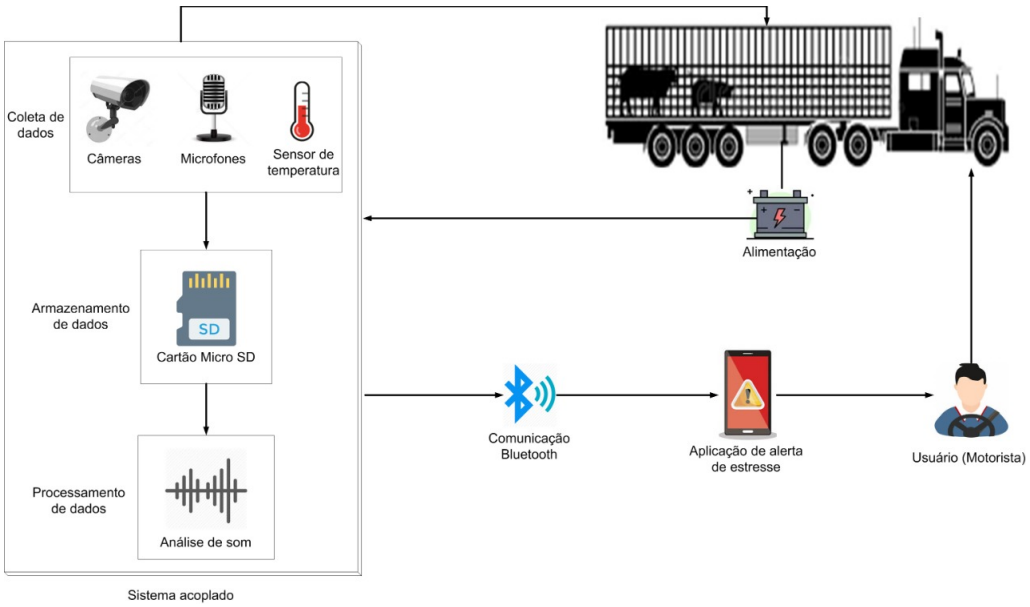
O produto é constituído por módulos de *hardware* e *software*, responsáveis pela coleta, pelo armazenamento, pelo processamento e pela apresentação da informação obtida a partir dos dados. A Figura 23 apresenta a estrutura em módulos do produto, representando os cinco módulos existentes na solução: (*1*) coleta e armazenamento, (*2*) processamento de áudio, (*3*) identificação de estresse animal, (*4*) comunicação e (*5*) interface de usuário.

Tabela 5 – Requisitos do Sistema

Requisito	**Descrição**
Coletar imagens e sons	O sistema deve coletar imagens e sons dos animais em transporte por meio de câmeras e microfones posicionados no interior da carroceria do caminhão.
Coletar variáveis de ambiente	O sistema deve coletar variáveis ambientais como temperatura, umidade e vibração durante o transporte dos animais por meio de sensores posicionados no interior da carroceria do caminhão.
Armazenar dados	O sistema deve registrar, em uma unidade de armazenamento, os dados coletados pelas câmeras, microfones e sensores durante o transporte dos animais.
Processamento de áudio	O sistema deve realizar a aplicação de filtros e processamento digital dos sons coletados durante o transporte.
Identificar estresse animal	O sistema deve ser capaz de identificar a ocorrência de estresse animal durante o transporte por meio da análise dos sons emitidos.
Alertar motorista	O sistema deve enviar alertas para um aplicativo voltado para dispositivos móveis que seja capaz de receber as informações processadas e enviar um sinal de alerta ao condutor do caminhão sempre que for identificada ocorrência de estresse animal.

1. **Módulo de coleta e armazenamento:**
    - Entrada: Sons, imagens e variáveis de ambiente.
    - Algoritmo: Realiza a coleta dos dados e o armazenamento estruturado dos dados em tempo real.
    - Saída: Dados armazenados e estruturados.
2. **Módulo de processamento de áudio:**
    - Entrada: Dados de som.
    - Algoritmo: Realiza a aplicação de filtros e processamento digital dos sons.
    - Saída: Dados de áudio filtrados e processados.
3. **Módulo de identificação de estresse animal**
    - Entrada: Dados de áudio filtrados e processados e variáveis de ambiente.
    - Algoritmo: Realiza a análise e classificação do som, identificando situações

Figura 22 – Visão geral do produto



Fonte: Autor (2024)

relativas ao estresse animal.

- Saída: Informação sobre o estresse animal.

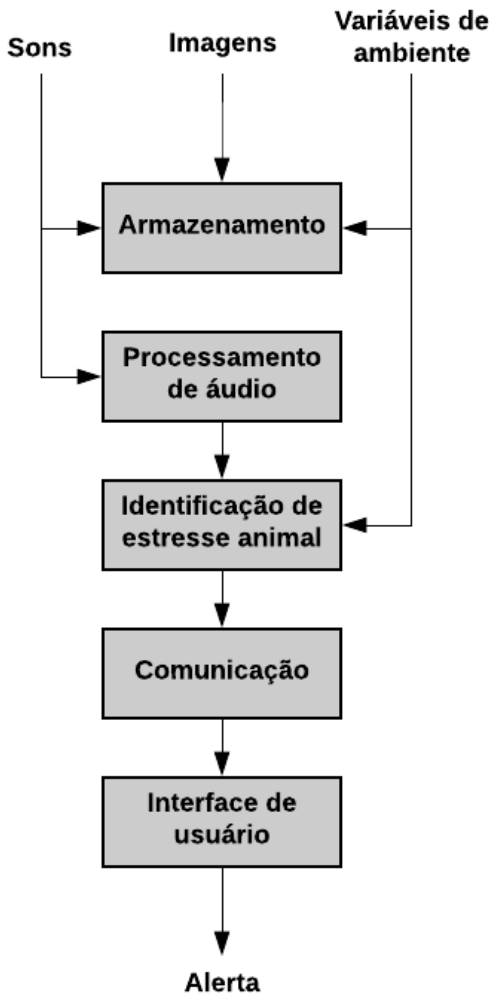
4. **Módulo de comunicação**

   - Entrada: Dados sobre estresse animal.
   - Algoritmo: Estabelece a conexão *Bluetooth* entre o dispositivo de processamento e o dispositivo móvel do motorista e realiza a transmissão dos dados de estresse animal.
   - Saída: Informação de estresse animal transmitido ao aplicativo móvel.

5. **Módulo de interface de usuário**

   - Entrada: Informação sobre estresse animal.
   - Algoritmo: Aplicativo móvel recebe dados sobre o estresse animal e apresenta alertas ao motorista sempre que for identificado o estresse durante o transporte.
   - Saída: Alerta de estresse animal.

Figura 23 – Estrutura em módulos do produto



Fonte: Autor (2024)

5.1.2 Levantamento e análise dos componentes do sistema

Para realizar a análise dos componentes necessários para a construção do módulo de coleta de dados foi realizada uma vistoria em um caminhão de transporte de gado, como ilustrado na Figura 24. Durante essa inspeção, foram feitas medições de comprimento e largura dos caminhões, bem como identificados os locais adequados para instalação dos equipamentos.

A partir da vistoria do caminhão, da definição dos requisitos do sistema e da elaboração do modelo inicial, foram elencados e analisados os componentes necessários para o desenvolvimento do produto. O levantamento de equipamentos foi realizado

Figura 24 – Caminhão de transporte bovino

Fonte: Autor (2024)

seguindo alguns critérios como capacidade de processamento e/ou armazenamento, velocidade, preço, usabilidade, interface de entrada e saída, interface de comunicação, alimentação e funcionalidades. Para a escolha dos equipamentos foi realizada uma busca em sites especializados em comércio online.

Com base nos requisitos definidos, foi identificada a necessidade de câmeras para a captura de imagens do embarque e desembarque dos animais, como também para identificar situações de brigas, tombos e estresse durante o transporte. Os microfones das câmeras são usados para a coleta de sons dos animais. Os principais critérios para a escolha das câmeras foram a qualidade da imagem gerada, modo de armazenamento (interno ou externo), interface de saída, tamanho, existência de microfone embutido e interface de comunicação sem fio.

Devido à possibilidade de comunicação via rede *wifi* e o protocolo IP, além de contar com microfone integrado, as câmeras definidas para a execução do trabalho foram as câmeras IP. Como características, possuem resolução de até 720p, com alcance de imagem de até 15 metros, visão noturna com alcance de até 10 metros, sensor de movimento, conectividade via cabo Rj45 e *wifi* , *slot* para armazenamento externo via cartão de memória. As câmeras possuem a alimentação de 5V/2A DC, com temperatura de operação entre 0ºC e 55ºC e umidade de operação entre 20% UR e 85% UR. As câmeras também dispõem de microfones omnidirecionais embutidos, de resolução de 32 bits e frequência de amostragem de 44,1 kHZ. A Figura 25 apresenta as câmeras definidas para a execução do trabalho.

Os sensores são os elementos responsáveis pela coleta dos valores das variáveis de ambiente, tais como temperatura, umidade e vibração. Para os dados de temperatura e umidade foram priorizados os sensores que realizam simultaneamente a coleta das

Figura 25 – Câmera IP

Fonte: Autor (2024)

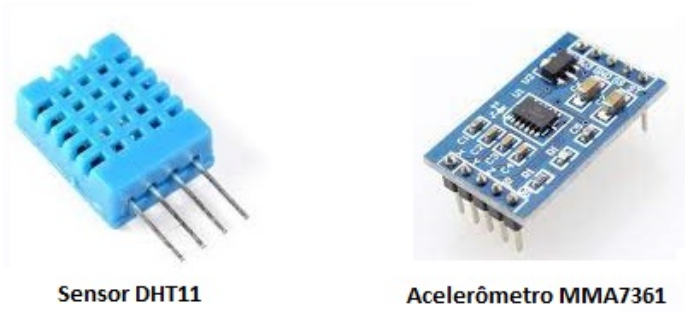
duas variáveis, além de características como a faixa de medição, precisão, alimentação, interface de saída e resolução. Para a escolha do sensor para a coleta da vibração, foram considerados os critérios de sensibilidade, interface de comunicação, alimentação, faixa de medição e tamanho.

O sensor escolhido para a coleta das variáveis de temperatura e umidade foi o sensor DHT11. Ele permite a realização de leituras de temperaturas na faixa entre 0ºC e 50ºC, com precisão de medição de $\pm$2ºC. Para leituras de umidade, o sensor realiza medição faixa entre 20% e 90% de umidade relativa (UR), com precisão de $\pm$% UR. O sensor possui resolução de 16 bits, com alimentação de 3,5V à 5,5V DC, corrente de 200μA a 500mA e tempo de resposta de 2 segundos.

Para a coleta de vibração, o sensor escolhido foi o acelerômetro triaxial MMA7361. Este sensor é capaz de medir a inclinação e aceleração tridimensional de um objeto, gerando uma variação de tensão, para cada eixo, correspondente à aceleração detectada. Como principais características, o MMA7361 possui faixa de operação entre -40ºC e 80ºC, com alimentação de 3,5V e 5V, corrente de 400μA a 600μA, tempo de resposta de 2ms, alta sensibilidade, baixo consumo de energia, condicionamento de sinal com filtro passa-baixa, compensação de temperatura, autoteste, detecção de queda livre e modo *sleep* para economia de energia. A Figura 26 apresenta os sensores definidos para a realização de coleta das variáveis de ambiente.

O processador é o elemento central da arquitetura do produto, sendo responsável pela execução dos algoritmos de identificação de estresse animal. O principal limitante na utilização de um processador nesta pesquisa é o fato da necessidade dele estar embarcado na carroceria do caminhão, o que restringiu o número de opções viáveis para o desenvolvimento do produto. Devido a este fator, foi definido a utilização de um

Figura 26 – Sensores para coleta de variáveis de ambiente



Fonte: Autor (2024)

minicomputador Raspberry, onde foram analisadas características como a capacidade de processamento e armazenamento, velocidade, interfaces de entrada e saída, interfaces de comunicação e preço.

O modelo definido para a realização do trabalho foi o Raspberry Pi 3 B+. Como características, ele possui um processador Broadcom *quad-core* com *clock* de 1,2GHz e arquitetura baseada em 64 bits, 1GB de memória RAM, 40 pinos de entrada e saída, 4 portas USB 2.0, suporte à conexão *wi-fi* (5GHz) e *Bluetooth* 4.2 BLE, conector *ethernet*, além de interface para câmera, *display* e *slot* para cartão micro SD. A Figura 27 apresenta a unidade de processamento escolhida.

Figura 27 – Raspberry Pi 3 B+

Fonte: Autor (2024)

Como parte dos objetivos e requisitos da pesquisa, o armazenamento de dados é uma funcionalidade que deve ser atendida pelo sistema. Para o armazenamento de dados, atualmente existem como opções os discos rígidos e as memórias *flash*. Considerando as

condições das estradas rurais, que geralmente estão em condições precárias, a utilização de uma memória mecânica, como os discos rígidos, se torna inviável devido à vibração presente na estrutura do caminhão durante o percurso. Sendo assim, levando em consideração o fator de vibração, além da integração com o elemento de processamento definido, optou-se pela utilização de uma memória *flash*, o cartão Micro SD. A capacidade atual de armazenamento dessa mídia permite que sejam armazenadas várias horas de gravação em cada cartão, permitindo o registro de toda a viagem de transporte.

Como elemento responsável por possibilitar o funcionamento de todo o sistema, a alimentação tem papel imprescindível no desenvolvimento da solução. Por se tratar de um sistema embarcado em um caminhão em deslocamento, sem o acesso à rede elétrica, as alternativas possíveis para a alimentação se tornam escassas. Como decisão de projeto, foi definido que o sistema utilizaria a bateria do caminhão em conjunto de um inversor de 12v para a sua alimentação. Para a definição do inversor as principais características analisadas foram a compatibilidade com o caminhão utilizado no transporte dos animais, potência e o preço.

O aplicativo de comunicação foi desenvolvido em outra parte deste projeto e opera por meio de *bluetooth*, visto que não há sinal de *wifi* ou de telefonia celular em diversas partes das propriedades rurais e estradas no Rio Grande do Sul.

A construção do equipamento de coleta foi elaborada com base em um sistema elétrico, composto por um quadro elétrico contendo dois disjuntores de 30A, quatro tomadas duplas, eletrodutos e um inversor conversor. Embora o equipamento tenha capacidade para suportar até oito dispositivos eletrônicos, inicialmente foi planejado o uso de apenas quatro câmeras junto com o Raspberry Pi. A escolha pelo quadro elétrico foi motivada por questões de segurança em casos de sobrecarga de energia, além de proporcionar facilidade e agilidade na desativação do sistema quando necessário. A Figura 28 apresenta o módulo de coleta de dados de transporte.

O módulo construído foi submetido a testes, nos quais quatro câmeras IP foram ligadas simultaneamente em todas as tomadas, visando avaliar a integridade e o correto funcionamento do equipamento. Além disso, os disjuntores foram testados para garantir o adequado desempenho do quadro de distribuição elétrica. É importante destacar que foram adquiridas quantidades adicionais de eletrodutos e cabos elétricos para permitir a extensão dos cabos durante a instalação nos caminhões de transporte.

Figura 28 – Equipamentos de coleta de dados de transporte

Fonte: Autor (2024)

5.2 Coleta e organização dos dados de vocalização bovina

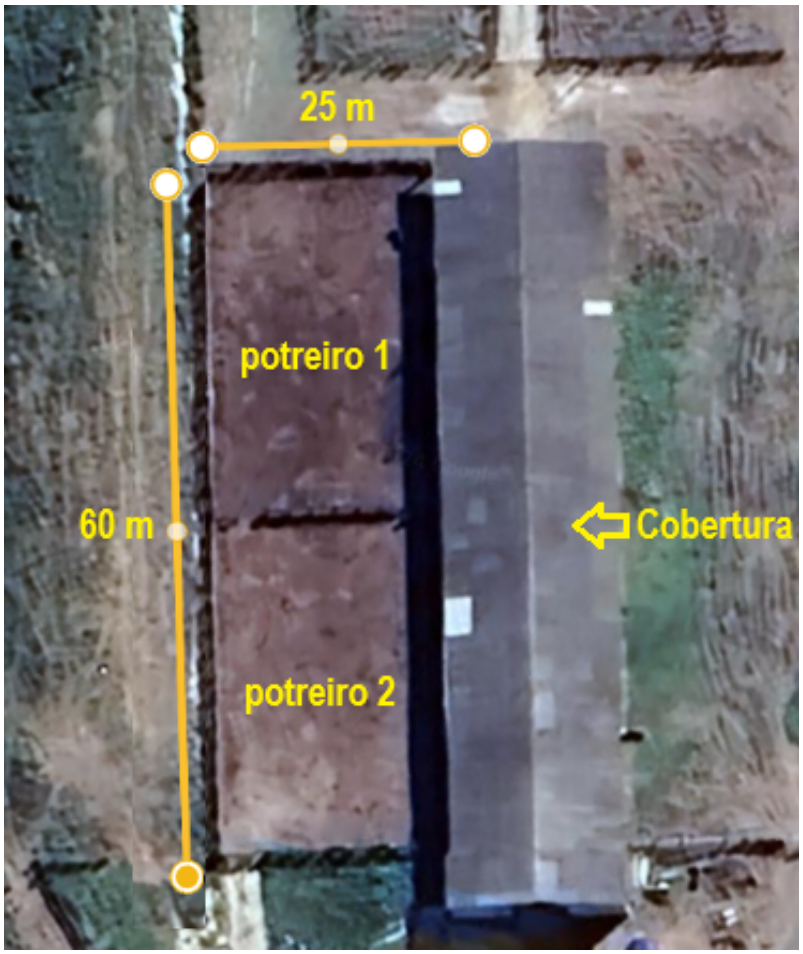
5.2.1 Organização do experimento de coleta

Com a impossibilidade de embarcar em caminhões a estrutura definida para a coleta de dados durante o transporte, a captação das vocalizações dos animais foi conduzida em dois contextos distintos: confinamento e manejo. Cada contexto representa um momento em que diferentes níveis de estresse animal podem ser observados. As coletas foram realizadas com 48 animais na raça Brangus (*Bos taurus indicus*).

No ambiente de confinamento, situado na zona rural da cidade de Bagé-RS (31°18'56"S 53°59'54"W), os animais estavam divididos em dois espaços amplos, de aproximadamente 1500m^2 no total (30m x 25m cada potreiro), onde podiam conviver, interagir e se alimentar livremente. Neste período, as interações com humanos eram minimizadas, ocorrendo somente quando era necessário repor a alimentação dos animais, garantindo assim a manutenção de um ambiente controlado. A Figura 29 apresenta uma imagem aérea do local de confinamento, onde observa-se a cobertura, de

aproximadamente 975m^2 (65m x 15m), que cobre os cochos de alimentação e à esquerda os dois potreiros onde estavam os animais.

Figura 29 – Local de confinamento dos animais



Fonte: Autor (2024)

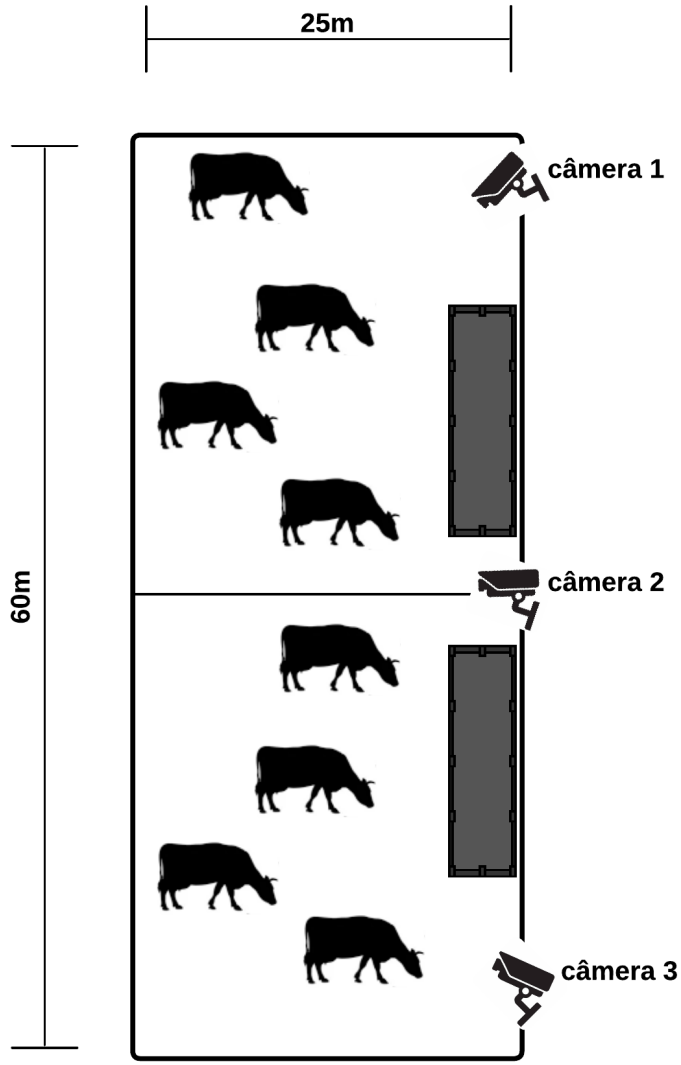
Para a captação dos sons no ambiente de confinamento, foram instaladas três câmeras em pontos estratégicos: uma na extremidade esquerda, outra central e a terceira na extremidade direita. As câmeras foram posicionadas próximas aos cochos, locais nos quais os animais se concentravam na maior parte de seu tempo durante o confinamento. A Figura 30 apresenta o esquema de posicionamento das câmeras.

As imagens foram registradas ao longo de um período de 14 dias, totalizando aproximadamente 1000 horas de filmagens. Vale destacar que, no ambiente de confinamento, as vocalizações registradas foram categorizadas como sendo de natureza controlada, refletindo o estado sereno dos animais e a ausência de interações estressantes. Este protocolo de coleta proporcionou uma fundamentação sólida para análises subsequentes das vocalizações em diferentes contextos. A Figura 31 apresenta os animais

em confinamento.

Outro momento de interação dos animais com os humanos ocorria no momento em que se fazia necessária o manejo dos animais. Nesse cenário, observou-se um manejo mais intensivo dos animais, caracterizado por interações enérgicas que envolviam gritos, gestos e cutucões. Foram acompanhadas duas sessões de manejo dos animais, com duração aproximada de 1 hora cada.

Figura 30 – Esquema de posicionamento das câmeras



Fonte: Autor (2024)

Durante os manejos, os animais foram conduzidos para um curral estreito que os levava até a balança. Nesse percurso, a interação entre humanos e animais tornava-se mais explícita, culminando na contenção dos animais dentro da gaiola de pesagem até a realização da leitura do peso. Após esse procedimento, o animal era libertado da gaiola, abrindo espaço para a entrada do próximo animal a ser pesado. A Figura 32 apresenta o

local de manejo dos animais.

Figura 31 – Animais em confinamento

Fonte: Autor (2024)

Durante os manejos foi evidente a agitação e um aumento nos níveis de estresse dos animais, atestado por especialistas em comportamento animal, acompanhado por uma maior frequência de produção de vocalizações em comparação com os períodos de confinamento. O monitoramento e registro deste processo foi realizado por meio de câmeras digitais. Esse enfoque visual proporciona uma análise mais detalhada das interações e comportamentos durante as pesagens.

Dessa forma, após a captura das filmagens em ambientes de estresse e não estresse, as gravações foram submetidas a uma etapa de análise com a finalidade de identificar momentos em que ocorreram vocalizações.

Figura 32 – Manejo dos animais

Fonte: Autor (2024)

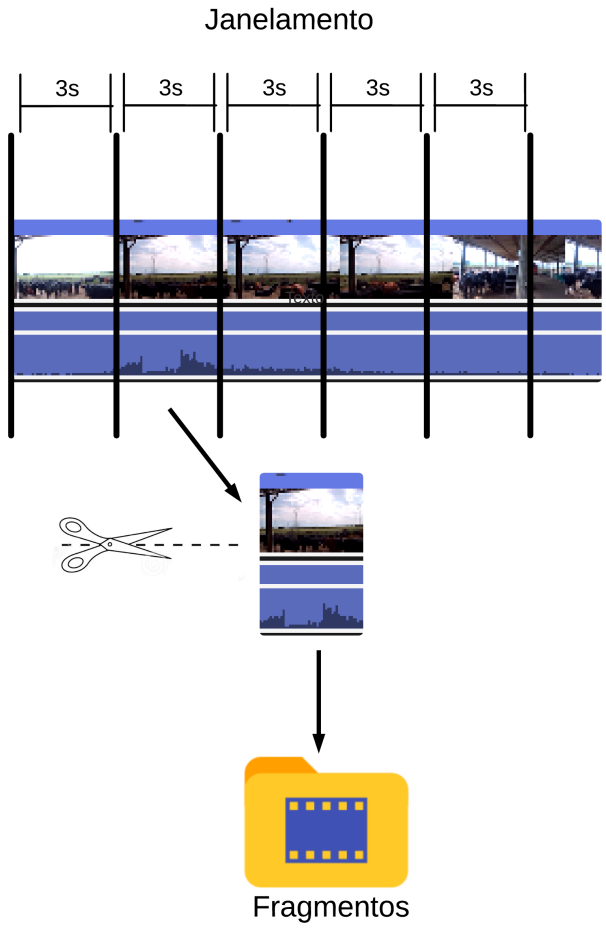
5.2.2 Preparação da base de dados

A construção da base de dados sobre o estresse animal originou-se das filmagens capturadas durante a fase de coleta de dados. Nesse estágio, foi necessária a análise dos registros e a categorização das vocalizações com base nos níveis de estresse manifestados pelos animais. Com o intuito de automatizar esse processo de preparação da base de vocalizações, foi desenvolvido um *software* em linguagem Python para detecção de picos de amplitude sonora nas filmagens.

Como as vocalizações de bovinos são tipicamente de natureza harmônica e geralmente variam entre 1,3s e 2,1s de duração (TORRE et al., 2015), o funcionamento do *software* envolve a criação de janelas de 3 segundos, nas quais são verificados picos de amplitude sonora. Esses picos indicam a presença de sons discrepantes, e caso identificados, os trechos correspondentes a essas janelas de tempo são individualmente salvos para análises manuais subsequentes. A Figura 33 demonstra o funcionamento do *software* desenvolvido.

Ao final da execução do programa, obtêm-se fragmentos dos dados originais nos

Figura 33 – Esquema de funcionamento do *software* de análise de vídeo



Fonte: Autor (2024)

quais ocorreram picos de amplitude sonora. Esses fragmentos foram submetidos a uma análise manual minuciosa, um a um, com o propósito de verificar a presença ou ausência de vocalizações de animais.

A análise manual foi empregada com a utilização do *software* Movavi Video Editor. Na análise, foram descartados os fragmentos que não se enquadravam como vocalizações ou cujos sons não possuíam qualidade satisfatória devido a baixa amplitude ou sobreposição de sons.

Após a verificação manual e individual dos fragmentos, foi realizada a extração dos sons, no formato wav (*Waveform Audio Format*), dos arquivos de vídeo restantes. Sobre os arquivos de áudio, foi submetido um filtro digital com o intuito de eliminar possíveis ruídos presentes nos sons, visando assegurar vocalizações isoladas para análises que estejam melhor alinhadas com as características acústicas das vocalizações emitidas

pelos animais.

Assim, das aproximadamente 1000 horas de filmagens durante o confinamento, foi possível extrair 357 vocalizações individuais dos animais. Dado o contexto de controle, esses sons foram categorizados como vocalizações normais.

No que se refere à coleta de dados durante os manejos, a análise dos registros de vídeo foi realizada manualmente utilizando o *software* Movavi Video Editor. Durante esse processo, foram identificadas um total de 186 vocalizações individuais emitidas pelos animais. Para manter a consistência dos dados, as vocalizações cuja duração eram inferiores a 3 segundos foram ajustadas para atender a esse padrão. Esse procedimento foi adotado para garantir que todos os arquivos de vocalizações tivessem o mesmo comprimento padrão de 3 segundos, o que também foi aplicado às vocalizações capturadas em ambiente de confinamento.

Assim como nas vocalizações em confinamento, as vocalizações em manejo também foram submetidas a uma etapa de filtragem de ruídos, visando garantir a qualidade das amostras de áudio. Dada a natureza desse contexto, caracterizado por interações intensas entre animais e humanos, esses sons foram rotulados como vocalizações de estresse.

Após a extração das vocalizações, elaborou-se duas pastas, uma com as vocalizações de confinamento e a outra com as vocalizações de manejo, bem como uma planilha contendo colunas que representam o nome do arquivo de áudio, o diretório em que está armazenado e o rótulo atribuído a cada um dos arquivos (estresse ou normal). Essa planilha foi empregada como entrada no processo de extração das *features* MFCC e treinamento das redes neurais, conforme será detalhado na próxima seção. Este método organizado e estruturado proporcionou uma base sólida para a implementação eficiente das arquiteturas de redes neurais, contribuindo para uma análise precisa das vocalizações registradas durante contextos positivos e negativos.

5.3 Classificação de sons

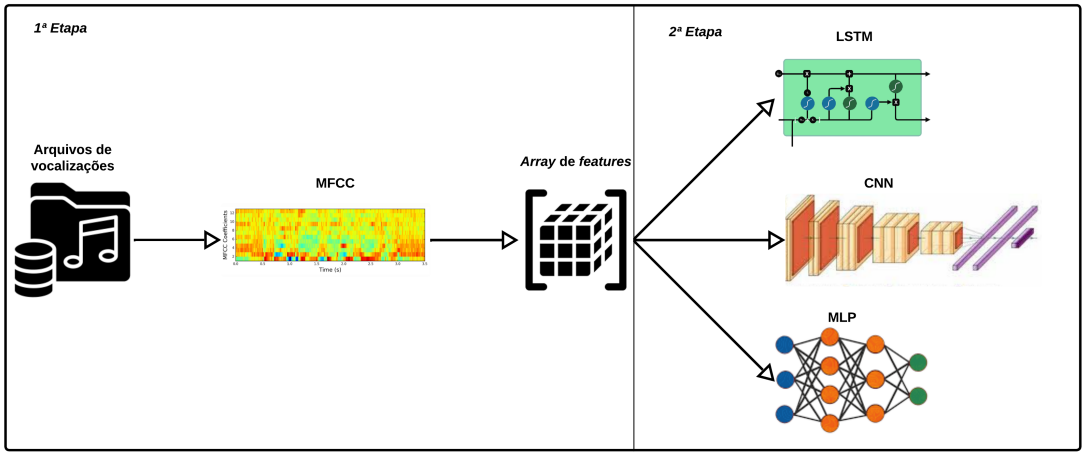
A tarefa de classificação e identificação de som animal é parte central do trabalho em desenvolvimento. Jahns (2008) defende que as principais tarefas envolvidas no reconhecimento de som são: (*i*) construir uma base de dados confiável contendo um repertório de vocalizações rotuladas de animais; (*ii*) 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.

Seguindo essa definição, para o desenvolvimento do módulo de identificação de estresse animal foram implementados *softwares* desenvolvidos na linguagem Python com a utilização das bibliotecas especializadas, detalhadas na Seção 2.4. Com base na fundamentação teórica, as técnicas escolhidas para a análise e classificação de som foram o MFCC (*Mel-frequency cepstral coefficients*) e redes neurais. A escolha dessas técnicas foi feita com base na revisão da literatura, onde foi evidenciada a viabilidade de suas utilizações para a identificação de sons animais nos trabalhos de Chung et al. (2013), Deshmukh et al. (2012), Jahns (2008), Manteuffel e Schön (2002), Jung et al. (2021), Pandeya et al. (2022).

A construção desse módulo foi composta por duas etapas, sendo a primeira etapa responsável pela extração das *features* de MFCC dos arquivos de som. A segunda etapa consiste na criação dos modelos de identificação e classificação de som, a partir do treinamento de diferentes redes neurais. A Figura 34 apresenta a lógica de funcionamento da etapa de classificação de sons.

Figura 34 – Etapas do módulo de classificação de sons



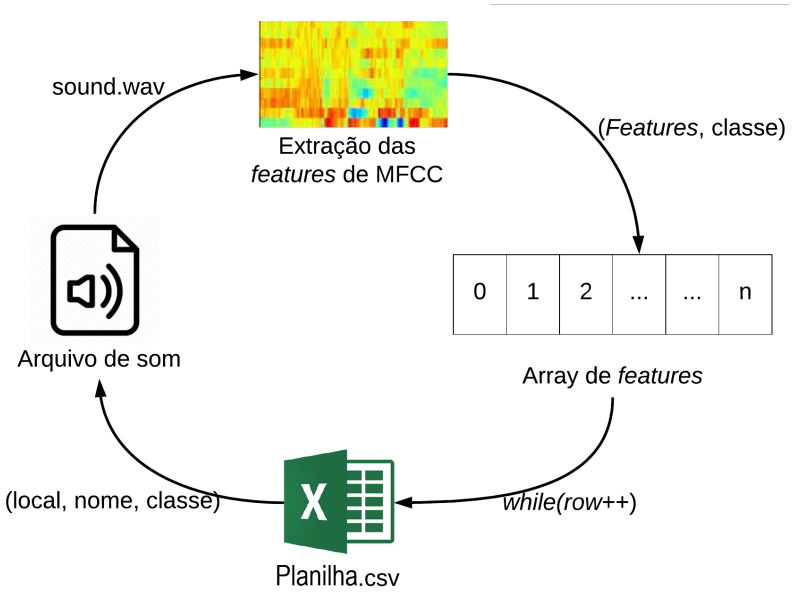
Fonte: Autor (2024)

5.3.1 Extração das características MFCC

Na primeira etapa, um ciclo é iniciado a partir da leitura da primeira linha da planilha, onde são identificados o nome do arquivo de áudio, o diretório onde se encontra

e a classe à qual pertence (normal ou estresse). Em seguida, a vocalização é submetida ao algoritmo MFCC para a extração de suas características. As características e a classe do som são armazenadas em um *array* dinâmico, que é preenchido à medida que as características de MFCC são extraídas. Posteriormente, o ciclo é retomado, com a leitura da próxima linha da planilha. O ciclo é encerrado ao final da planilha, quando o *array* de *features* estiver completo com todas as características de MFCC dos arquivos de som. A Figura 35 demonstra a lógica de funcionamento do extrator de características MFCC.

Figura 35 – Lógica de funcionamento da extração de características MFCC das vocalizações



Fonte: Autor (2024)

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 a partir 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, e o espaçamento entre amostras foi estabelecido em 512, indicando o número de amostras que a janela de análise se desloca para calcular os coeficientes MFCC entre quadros adjacentes. 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 (SATTAR, 2022; ZHENG; ZHANG; SONG, 2001; DUAN et al., 2021; TIWARI, 2010; JUNG et al., 2021).

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, onde os modelos podem acabar favorecendo a classe majoritária e não aprendendo adequadamente a distinguir a classe minoritária.

Ao término do processo de extração das características MFCC, para cada arquivo de áudio, foi obtida uma matriz de dimensões 13 x 259. Cada linha dessa matriz reflete um dos 13 coeficientes MFCC, enquanto cada coluna representa um dos 259 *frames* temporais. Os valores presentes em cada posição na matriz denotam os coeficientes MFCC em momentos específicos ao longo da vocalização.

Sobre os dados extraídos, foi aplicada a normalização conhecida como *z-score*. Essa técnica de normalização é amplamente 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 15 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.

$$z\text{-}score = \frac{x - \mu}{\sigma} \tag{15}$$

Dessa forma, os dados normalizados apresentam uma média e um desvio padrão próximos a um. Essa normalização é fundamental para evitar que características com escalas diferentes influenciem desproporcionalmente 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.

Na segunda etapa, com o objetivo de realizar um estudo comparativo entre diversos modelos de redes neurais voltados à classificação de vocalizações bovinas, optou-se por três categorias de redes neurais artificiais: *Multilayer Perceptron* (MLP), *Convolutional Neural Network* (CNN) e *Long Short-Term Memory* (LSTM). O modelo MLP, junto com a algoritmo de aprendizado *Backpropagation*, é o modelo mais clássico de redes neurais artificiais; CNN são uma família de abordagens mais modernas,

com extremo sucesso em reconhecimento de imagens e com alguma literatura em reconhecimento de sons; LSTM é uma rede neural recorrente que poderia ter influência positiva no fato de que sons são séries temporais, aplicações para as quais esse tipo de rede é usada.

Para cada categoria de rede neural, foram desenvolvidas três arquiteturas distintas, 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.

5.3.2 implementação das arquiteturas MLP

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

1. Arquitetura Básica:
    - Número de camadas: 3 (1 camada de entrada, 1 camada oculta, 1 camada de saída)
    - Número de neurônios: 64, 32, 1
    - Funções de ativação: leaky ReLU, leaky ReLU, sigmoid
2. Arquitetura Intermediária
    - Número de camadas: 4 (1 camada de entrada, 2 camadas ocultas, 1 camada de saída)
    - Número de neurônios: 128, 64, 32, 1
    - Funções de ativação: leaky ReLU, leaky ReLU, leaky ReLU, sigmoid
3. Arquitetura Robusta
    - Número de camadas: 5 (1 camada de entrada, 3 camadas ocultas, 1 camada de saída)
    - Número de neurônios: 256, 128, 64, 32, 1
    - Funções de ativação: leaky ReLU, leaky ReLU, leaky ReLU, leaky ReLU, sigmoid

A primeira configuração, denominada Arquitetura Básica, consistiu em três camadas, sendo uma de entrada, uma oculta e uma de saída. O número de neurônios em cada camada foi definido como 64, 32 e 1, respectivamente. As funções de ativação utilizadas foram leaky ReLU para as camadas ocultas e sigmoid para a camada de saída. Esta configuração básica visa fornecer uma abordagem inicial para a classificação de vocalizações de estresse em bovinos, capturando as características essenciais das vocalizações e fornecer uma compreensão inicial do problema.

Já a segunda configuração, chamada de Arquitetura Intermediária, contou com quatro camadas, sendo uma de entrada, duas ocultas e uma de saída. O número de neurônios foi ajustado para 128, 64, 32 e 1, respectivamente, para cada camada. Assim como na configuração anterior, as funções de ativação utilizadas foram leaky ReLU para as camadas ocultas e sigmoid para a camada de saída. Esta configuração busca melhorar a capacidade de representação das características das vocalizações adicionando uma camada oculta adicional e mais neurônios. A Arquitetura Intermediária pretende capturar relações mais complexas nos dados de entrada, aumentando a capacidade do modelo de aprender padrões específicos.

Por fim, a terceira configuração, denominada Arquitetura Robusta, foi a mais complexa, composta por cinco camadas, incluindo uma de entrada, três ocultas e uma de saída. O número de neurônios aumentou progressivamente, sendo definido como 256, 128, 64, 32 e 1 para cada camada, respectivamente. As funções de ativação também foram mantidas consistentes com as configurações anteriores, utilizando leaky ReLU para as camadas ocultas e sigmoid para a camada de saída. Esta configuração representa uma abordagem mais sofisticada, com múltiplas camadas ocultas, aumentando ainda mais a capacidade do modelo de aprender as características essenciais das vocalizações. Com um número maior de neurônios e camadas, a Arquitetura Robusta visa alcançar um desempenho ainda melhor na tarefa de classificação, buscando uma melhor generalização dos padrões presentes nos dados.

Destaca-se que as escolhas das funções de ativação foram avaliadas por meio de testes preliminares. Além das funções Tanh e ReLU, diferentes combinações de funções entre camadas também foram consideradas. No entanto, a função de ativação Leaky ReLU em todas as camadas foi selecionada devido aos seus melhores resultados médios.

5.3.3 implementação das arquiteturas CNN

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 *flattten*, 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 *flattten*, 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 *flattten*, 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*.

Para a definição das funções de ativação para as arquiteturas foram analisadas testes preliminares as funções Tanh e ReLU. Contudo, as funções Leaky ReLU apresentaram melhores desempenhos médios.

5.3.4 implementação das arquiteturas LSTM

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

1. Arquitetura Básica:
    - Número de camadas: 2 (1 camada LSTM, 1 camada de saída)
    - Camada LSTM: 64 unidades, ativação tanh
    - camada de saída: 1 neurônio, ativação sigmoid
2. Arquitetura Intermediária

- Número de camadas: 3 (2 camadas LSTM, 1 camada de saída)
- Camadas LSTM: 2 camadas, 128 e 64 unidades, ativação tanh
- Camada de saída: 1 neurônio, ativação sigmoid

3. Arquitetura Robusta

- Número de camadas: 5 (3 camadas LSTM, 1 camada totalmente conectada, 1 camada de saída)
- Camadas LSTM: 3 camadas, 256, 128 e 64 unidades, ativação tanh
- Camada totalmente conectada: 128 neurônios, ativação Leaky ReLU
- Camada de saída: 1 neurônio, ativação sigmoid

A Arquitetura Básica consiste em uma única camada LSTM seguida por uma camada de saída. Com 64 unidades na camada LSTM e uma função de ativação tanh. Essa arquitetura busca capturar padrões temporais básicos nas vocalizações para realizar a classificação.

A segunda arquitetura, denominada Intermediária, é composta por duas camadas LSTM, cada uma com 128 e 64 unidades, respectivamente, seguidas por uma camada de saída. Com funções de ativação tanh em ambas as camadas LSTM. Essa arquitetura visa aprofundar a capacidade de captura de padrões temporais mais complexos nas vocalizações.

Já a Arquitetura Robusta apresenta uma complexidade ainda maior, com três camadas LSTM, cada uma com 256, 128 e 64 unidades, respectivamente, seguidas por uma camada totalmente conectada com 128 neurônios e uma camada de saída. Com funções de ativação tanh em todas as camadas LSTM, e com função de ativação Leaky ReLU para a camada totalmente conectada, essa arquitetura busca capturar os padrões temporais mais complexos e abstratos presentes nas vocalizações, visando alcançar uma classificação ainda mais precisa.

Para as redes LSTM, os testes preliminares indicaram que as funções de ativação Tanh obtiveram melhores desempenhos médios em comparação com as funções ReLU e Leaky ReLU. Além disso, foram exploradas várias combinações de funções de ativação entre as camadas, mas não foram observadas melhorias significativas.

5.3.5 Configurações de treinamento

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

O otimizador Adam, abreviação de "*Adaptive Moment Estimation*", é uma técnica de otimização popularmente utilizada em redes neurais profundas para ajustar os pesos durante o treinamento. Ele combina conceitos do método de otimização de *momentum* e da taxa de aprendizagem adaptativa. Uma das vantagens do Adam é sua capacidade de lidar bem com problemas de otimização não estacionários e de alta dimensionalidade, comuns em redes neurais profundas. Além disso, o algoritmo é relativamente simples de implementar e geralmente requer poucos ajustes de hiperparâmetros (GÉRON, 2019).

A inicialização Glorot, também conhecida como inicialização Xavier, é um método popular para inicializar os pesos de uma rede neural de forma eficaz. A ideia por trás da inicialização Glorot é ajustar a escala de inicialização dos pesos de forma a manter a variância dos sinais de entrada e saída aproximadamente constante em todas as camadas da rede. Isso ajuda a evitar que os sinais se tornem muito pequenos ou muito grandes à medida que passam pela rede durante o treinamento (GÉRON, 2019).

A determinação da taxa de aprendizagem adequada pode ser um desafio durante o treinamento de redes neurais. Uma taxa muito alta pode resultar em divergência durante o treinamento, enquanto uma taxa muito baixa pode levar a uma convergência lenta e ineficiente. Uma estratégia para lidar com esse dilema é iniciar o treinamento com uma

taxa de aprendizagem alta e, em seguida, reduzi-la à medida que o progresso se estabiliza. Essa abordagem, conhecida como Agendamento de Desempenho, envolve monitorar o erro de validação a cada N épocas e diminuir a taxa de aprendizado por um fator λ quando o erro não diminuir mais (GÉRON, 2019). Essa técnica permite ajustar dinamicamente a taxa de aprendizado conforme necessário durante o treinamento, otimizando assim o processo de aprendizado da rede neural.

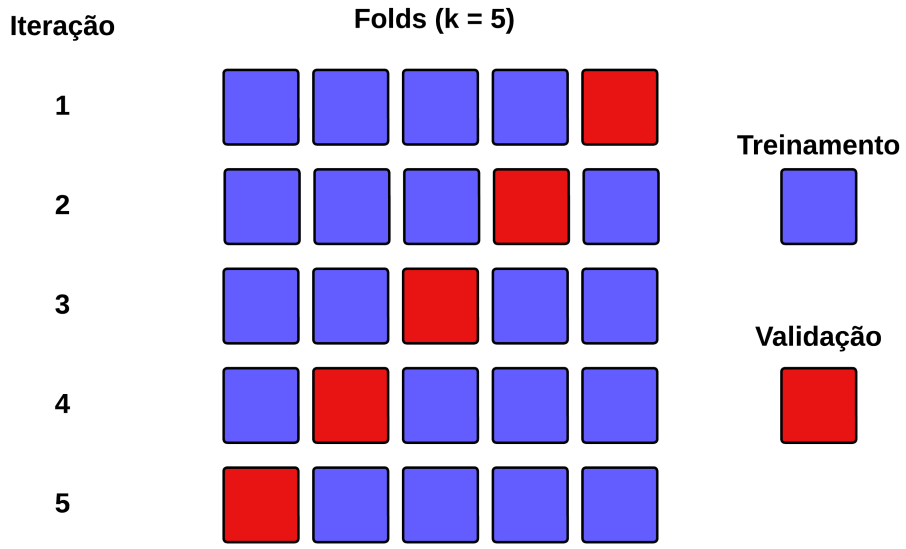
Em todas as arquiteturas implementadas, foram empregadas técnicas de regularização, como *dropout* e *batch normalization*, com o objetivo de mitigar o sobreajuste dos modelos. A taxa de aprendizagem foi fixada em 0,01 e ajustada dinamicamente durante o treinamento, sendo reduzida a taxa pela metade (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 (GÉRON, 2019). 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 (SILAPARASETTY, 2020). A Figura 36 demonstra o funcionamento da validação cruzada.

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.

Sendo assim, após a conclusão dos treinamentos das redes neurais, os resultados de cada execução da validação cruzada, como a acurácia, precisão, revocação e *F1-score*, foram registrados para análises posteriores por meio de técnicas estatísticas de variância. No total, para cada arquitetura implementada, obteve-se 25 registros de cada uma das métricas analisadas. Os resultados serão apresentados e discutidos na seção 6.2.

Figura 36 – Funcionamento da técnica de validação cruzada



Fonte: Autor (2024)

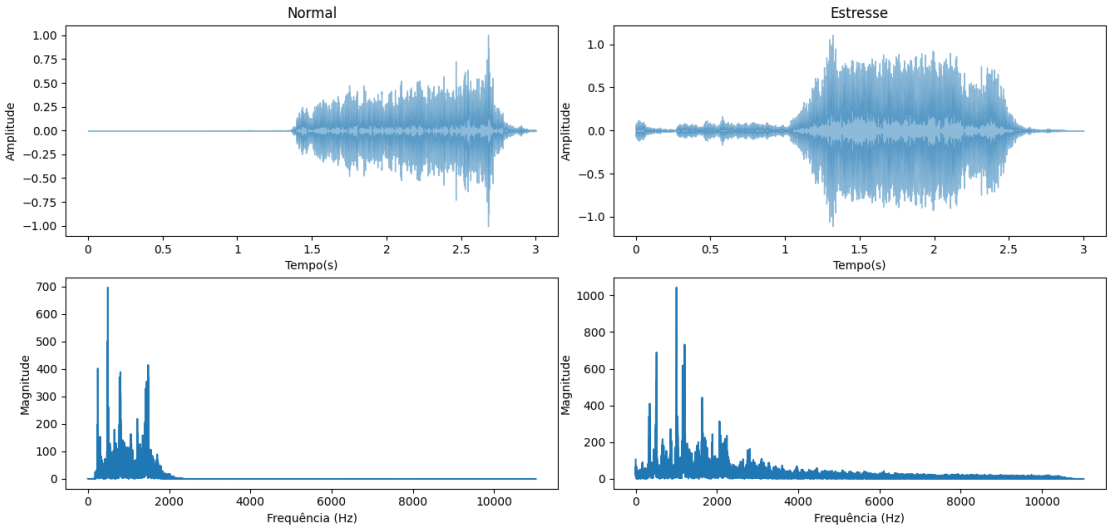
5.4 Análise acústica

Para compreender e analisar as características das vocalizações coletadas, foi necessário realizar análises nos domínios da amplitude e frequência. Para esse propósito, empregou-se um *software*, utilizando as bibliotecas descritas na Seção 2.4 com a finalidade de extrair e visualizar os sinais de áudio nos domínios do tempo e frequência, bem como a sua representação na forma de onda. A Figura 37 apresenta duas vocalizações durante os períodos de confinamento e manejo, categorizadas como normais e sob estresse, nos domínios do tempo e da frequência.

No âmbito da frequência, foram examinadas a densidade espectral de energia (*Power Spectral Density* - PSD) e o espectrograma das vocalizações, buscando compreender a distribuição das principais frequências e identificar eventuais disparidades entre elas. A Figura 38 exibe gráficos comparativos da análise no domínio da frequência.

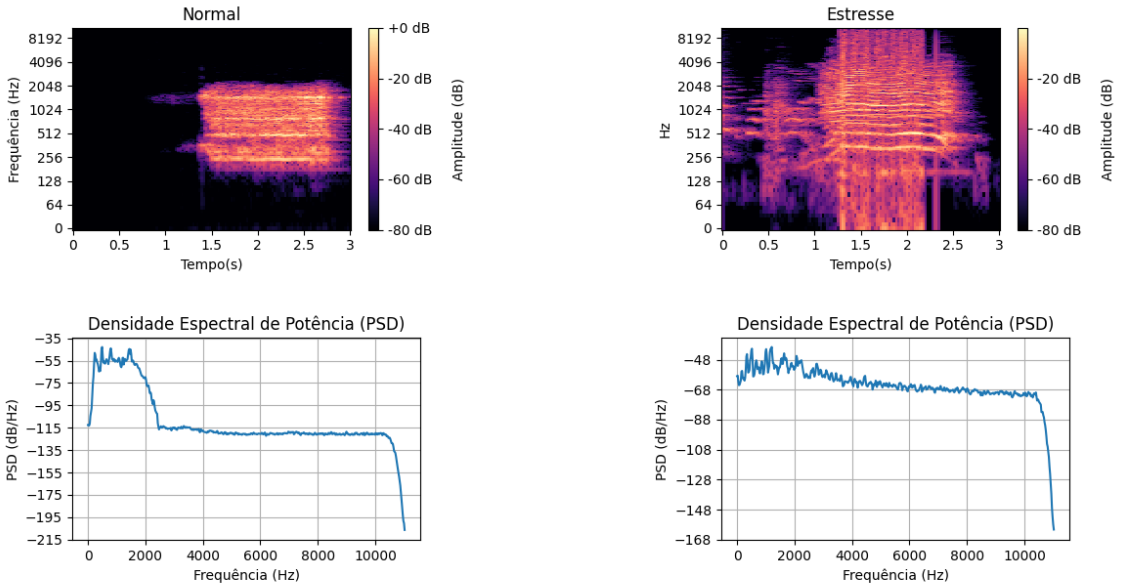
A análise visual revelou diferenças entre as vocalizações durante períodos de confinamento e manejo. Em termos de frequência, as representações visuais demonstraram que as vocalizações durante momentos de estresse apresentavam frequências mais elevadas, evidenciadas claramente nos espectrogramas e nos gráficos de PSD. Essa análise visual foi importante para destacar indícios relevantes sobre mudança de produção vocal em diferentes contextos, incentivando investigações mais aprofundadas sobre as características acústicas das vocalizações.

Figura 37 – Comparação nos domínios do tempo e frequência de vocalizações normais e de estresse



Fonte: Autor (2024)

Figura 38 – Comparação do espectrograma e PSD de vocalizações normais e de estresse



Fonte: Autor (2024)

Assim, além da análise visual das vocalizações, realizou-se um estudo abordando as principais características acústicas examinadas em trabalhos correlatos sobre as vocalizações de bovinos, incluindo a frequência fundamental (F0), os formantes F1 - F4, *jitter*, *shimmer*, harmonia e intensidade das vocalizações (TORRE et al., 2015; YEON et al., 2006; MOURA et al., 2008; IKEDA; ISHII, 2008; MEEN et al., 2015; LEE et al.,

2015; LEE et al., 2014).

Esta análise torna-se importante para a comparação entre vocalizações provenientes de situações de confinamento e de manejo, visando identificar os principais fatores que possam caracterizar os estados de tranquilidade ou estresse dos animais. Além disso, a análise desempenha um papel significativo como um ponto de comparação entre os resultados deste estudo e aqueles encontrados na literatura. Para cada uma das características examinadas, foram calculadas a média e o desvio padrão para todos os valores associados a cada tipo de vocalização.

O *software* empregado na extração dos parâmetros percorreu as duas pastas de arquivos construídas na etapa 5.2.2, sendo a primeira destinada a vocalizações normais e a segunda a vocalizações associadas ao estado de estresse. Em cada iteração sobre os arquivos, foram extraídos os parâmetros de F0, formantes de F1 a F4, o *jitter*, o *shimmer*, a harmonia, a intensidade e a classificação do tipo de vocalização (normal ou estresse).

Posteriormente, os dados foram agregados em uma planilha no formato CSV, onde cada linha representa uma vocalização, enquanto as colunas abrangem as características acústicas. No total a planilha resultante ficou com 543 linhas, sendo 357 representando vocalizações normais e 186 representando vocalizações de estresse. Essa planilha serviu como base para calcular as médias gerais e os desvios-padrão de cada um dos parâmetros, separados de acordo com a categoria de vocalização.

As médias obtidas para os parâmetros selecionados serão discutidas mediante uma análise estatística, conforme detalhado na Seção 6.1. Nesta seção, cada parâmetro será analisado individualmente, a fim de verificar a existência de diferenças estatisticamente significativas que possibilitem discernir entre vocalizações normais e aquelas associadas ao estado de estresse. Adicionalmente, será conduzida uma análise comparativa com os resultados encontrados na literatura, no que se refere às vocalizações animais e sua vinculação à identificação de situações de estresse.

6 RESULTADOS E DISCUSSÃO

6.1 Análise acústica

Para analisar os resultados obtidos na etapa de análise acústica, foi empregada uma análise estatística por testes t de Student sobre as características acústicas das vocalizações coletadas. Essa análise permite avaliar se existem diferenças significativas entre as médias de grupos independentes. O objetivo principal da análise estatística é determinar se a variabilidade observada nas médias entre os grupos é maior do que a variabilidade esperada devido ao acaso.

A análise estatística foi realizada com base nos parâmetros acústicos de frequência fundamental (F0), formantes do espectro F1-F4, *jitter*, *shimmer*, harmonia e intensidade extraídos das vocalizações de dois grupos independentes:

- Confinamento: Coletas realizadas em local onde os animais tinham liberdade dentro de um espaço amplo, alimentação e poucas interações com humanos. Essas vocalizações são consideradas como normais.
- Manejo: Coletas realizadas durante o manejo dos animais, onde haviam intensas interações com humanos, com gritos, cutucões e barulho. Essas vocalizações são consideradas de estresse.

O teste de hipóteses foi realizado para cada parâmetro definido, comparando-se as médias obtidas para cada grupo independente. Sendo assim, as hipóteses levantadas são:

- Hipótese Nula (H0): Não há diferença significativa entre as médias dos parâmetros acústicos para vocalizações de confinamento e de manejo. Qualquer variação observada é atribuída ao acaso.
- Hipótese Alternativa (HA): Existe uma diferença significativa entre as médias dos parâmetros acústicos para vocalizações de confinamento e de manejo. A variação observada não é devida ao acaso, indicando uma relação verdadeira entre situações de estresse e mudanças nos parâmetros de vocalizações.

O teste t de Student foi empregado para avaliar se as médias observadas dos grupos de confinamento e manejo apresentaram diferenças estatisticamente significativas ou são simplesmente variações aleatórias nos dados, com um nível de confiança de 5%. A Equação 16 ilustra o cálculo do teste t de Student.

$$t = \frac{m_1 - m_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}} \tag{16}$$

Onde:

- m_1 e m_2 são as médias de cada grupo
- s_1 e s_2 são os desvios padrão para cada grupo
- n_1 e n_2 são a quantidade de amostras de cada grupo

A Tabela 6 apresenta os resultados obtidos para o teste de variância para cada um das características acústicas analisadas, onde é possível observar que apenas para F0 max e para a harmonia não houve diferenças significativas entre as vocalizações de confinamento e manejo.

Tabela 6 – Análise de variância dos parâmetros vocais por testes de t de Student

Parâmetro	**Confinamento**	**Manejo**	***P-value***
Média F0 (Hz)	172,87 ± 25,27	276,23 ± 47,38	< 0,001***
Min F0 (Hz)	81,74 ± 14,57	132,54 ± 29,24	< 0,001***
Max F0 (Hz)	436,17 ± 57,25	458,70 ± 51,46	0,1133
Média F1 (Hz)	848,34 ± 51,94	940,47 ± 54,83	< 0.001***
Média F2 (Hz)	1445,04 ± 61,42	1582,79 ± 58,74	< 0.001***
Média F3 (Hz)	2029,96 ± 89,06	2421,29 ± 103,72	< 0.001***
Média F4 (Hz)	3435,90 ± 93,31	3802,71 ± 92,55	0,004**
Jitter (%)	1,61 ± 0,63	3,85 ± 0,81	< 0.001***
Shimmer (%)	14,93 ± 1,7	19,08 ± 1,8	< 0.001***
Harmonia (dB)	7,12 ± 0,17	7,66 ± 0,18	0,198
Intensidade (dB)	48,74 ± 3,35	57,59 ± 4,37	< 0.001***

Os resultados obtidos destacam que, para a maioria dos parâmetros analisados, as vocalizações de animais durante o manejo apresentam médias estatisticamente distintas em comparação com as vocalizações durante o confinamento. Essa disparidade sugere que o contexto do manejo pode influenciar significativamente as características acústicas das vocalizações dos animais, indicando a existência de padrões distintos entre os dois cenários.

Para a frequência fundamental (F0), tanto a média quanto o mínimo mostram diferenças significativas entre os grupos de confinamento e manejo ($p < 0,001$). A média é maior no grupo de manejo. No entanto, a frequência fundamental máxima (Max F0) não apresenta diferença significativa entre os grupos ($p = 0,1133$). Os resultados obtidos indicam uma variação considerável na modulação vocal entre os contextos de confinamento e manejo.

As médias dos primeiros quatro formantes (F1, F2, F3, F4) exibem diferenças significativas entre os grupos ($p < 0,001$), sendo maiores no grupo de manejo. As médias dos formantes no grupo de manejo foram consistentemente superiores em comparação com o grupo de confinamento. Esses resultados sugerem que as características espectrais da vocalização bovina são sensíveis às condições de manejo, possivelmente refletindo a influência do ambiente na produção vocal dos animais.

Para os parâmetros de variabilidade vocal, *jitter* e *shimmer*, há diferenças significativas entre os grupos ($p < 0,001$), ambos sendo maiores no grupo de manejo. Em relação ao parâmetro de *jitter*, constatou-se que os animais em manejo apresentaram vocalizações com índices mais elevados ($3,85 \pm 0,81\%$) em comparação com o grupo de confinamento ($1,61 \pm 0,63\%$). Essa diferença na variabilidade temporal das vocalizações pode indicar uma resposta vocal mais instável ou agitada nos bovinos submetidos ao manejo. Analisando o parâmetro *shimmer*, observou-se que o grupo de manejo apresentou um índice de variação maior ($19,08 \pm 1,8\%$) em comparação com o grupo de confinamento ($14,93 \pm 1,7\%$). Isso sugere uma maior variação na amplitude das vocalizações nos bovinos sob condições de manejo, indicando possíveis alterações na regularidade e estabilidade vocal.

A medida de harmonia em decibéis não mostrou diferença significativa entre os grupos ($p = 0,198$), com médias de $7,12 \pm 0,17$ (confinamento) e $7,66 \pm 0,18$ (manejo). Isso sugere que, ao contrário de outros parâmetros, a harmonia vocal não foi afetada de maneira significativa pelas condições de manejo.

No que se refere à intensidade vocal, constatou-se que o grupo de manejo exibiu uma intensidade vocal mais elevada ($57,59 \pm 4,37$ dB) em comparação com o grupo de confinamento ($48,74 \pm 3,35$ dB), apresentando diferença significativa entre os grupos ($p < 0,001$). Essa diferença sugere uma vocalização mais intensa em situações de manejo, indicando uma resposta vocal mais vigorosa e enérgica nessas condições.

Esses resultados indicam que as condições de manejo têm um impacto significativo nas características vocais, com diferenças significativas em parâmetros como frequência fundamental, formantes, variabilidade vocal e intensidade entre os grupos de confinamento e manejo.

Em comparação aos trabalhos correlatos, é possível observar semelhanças nos resultados. Na frequência fundamental (F0), os valores médios para as frequências máximas, mínimas e médias foram, respectivamente, $436,17 \pm 57,25$ (confinamento) e $458,70 \pm 51,46$ (manejo), $81,74 \pm 14,57$ (confinamento) e $132,54 \pm 29,24$ (manejo), e

172,87 ± 25,27 (confinamento) e 276,23 ± 47,38 Hz (manejo).

Na literatura, em análise com bovinos, Yeon et al. (2006) detectaram frequência média de F0 no alcance entre 214 ± 66,40 (alimentação antecipada) e 221,96 ± 61,20 Hz (cio). Yajuvendra et al. (2013) no seu estudo de caracterização das vocalizações observaram valores médios para F0 de 191,57 ± 2,40 Hz. Similarmente, Torre et al. (2015) ao caracterizarem vocalizações de vacas de altas e baixas frequências (HFC e LFC) e de bezerros observaram valores médios de F0 de 152,8 ± 3,1 (HFC), 81,1 ± 0,9 e 142,8 ± 1,8 Hz (bezerros) respectivamente. O estudo de Green (2020) observou valores médios para F0 de 115,00 ± 7,84 (parto) e 177,00 ± 15,13 Hz (separação). Em outro estudo, Green (2020) observou valores médios de F0 de 183,5 ± 9,90 (contexto positivo) e 286,0 ± 40,19 Hz (contexto negativo).

Na análise de F0 máximo, na literatura encontra-se resultados de 352,12 ± 4,59 Hz no trabalho de (YAJUVENDRA et al., 2013). No estudo de Torre et al. (2015), observa-se F0 máximo de 198,7 ± 3,6 (HFC), 84,7 ± 1,0 (LFC) e 153,3 ± 2,1 Hz (bezerros). Já Green (2020) identificou valores de F0 máximo de 167,00 ± 9,52 (parto) e 218,00 ± 13,32 Hz (separação), além de valores de 473,1 ± 76,67 (contexto positivo) e 559,3 ± 65,54 Hz (contexto negativo).

Avaliando os resultados para F0 mínimo, Yajuvendra et al. (2013) verificaram valores de 78,79 ± 1,69 Hz, enquanto Torre et al. (2015) encontraram F0 mínimos de 91,0 ± 2,8 (HFC), 74,8 ± 1,0 (LFC) e 121,0 ± 1,6 Hz (bezerros). Green (2020) analisando diferentes contextos observou F0 mínimos de 75,80 ± 2,38 (parto) e 70,60 ± 2,42 (separação), 66,17 ± 2,11 (contexto positivo) e 76,32 ± 3,57 Hz (contexto negativo).

Na análise dos quatro primeiros formantes (F1, F2, F3, F4) das vocalizações, as médias obtidas foram 848,34 ± 71,94 (F1 confinamento), 940,47 ± 94,83 (F1 manejo), 1445,04 ± 81,42 (F2 confinamento), 1582,79 ± 78,74 (F2 manejo), 2029,96 ± (F3 confinamento), 2421,29 ± 143,72 (F3 manejo), e 3435,90 ± 123,31 (F4 confinamento) e 3802,71 ± 162 55 Hz (F4 manejo). Na literatura é possível observar resultados semelhantes para as frequências médias dos formantes.

Para o primeiro formante (F1), Yeon et al. (2006) encontraram os valores médios de 784,05 ± 127,45 (alimentação antecipada) e 832,25 ± 150,77 Hz (cio). Yajuvendra et al. (2013) detectaram F1 médio entre 689,12 ± 17,53 e 1015,88 ± 17,53 Hz. Torre et al. (2015) observaram F1 médio de 228,3 ± 1,8 (HFC) e 391,7 ± 5,3 Hz (bezerros).

No segundo formante (F2), Yeon et al. (2006) observaram valores médios de 1710,40 ± 166,97 (alimentação antecipada) e 1570,87 ± 127,83 Hz (cio). Yajuvendra

et al. (2013) encontraram os valores entre 1675,97 ± 29,06 e 1942,70 ± 9,19 Hz. Já Torre et al. (2015) verificaram valores de 634,3 ± 6,6 (LFC), 644,6 ± 3,7 (HFC) e 1162,1 ± 16 Hz (bezerros).

Avaliando o terceiro formante (F3), Yeon et al. (2006) verificaram valores médios de 2675,46 ± 188,88 (alimentação antecipada) e 2418,96 ± 183,76 Hz (cio). Yajuvendra et al. (2013) constataram F3 médios entre 3079,93 ± 35,66 e 3412,47 ± 11,27 Hz. No estudo de Torre et al. (2015) observaram valores para F3 de 1064 ± 11,7 (LFC), 1073 ± 2,8 (HFC) e 1939 ± 24,6 Hz (bezerros).

Para o quarto formante (F4), Yeon et al. (2006) identificaram valores médios de 3818,71 ± 201,40 (alimentação antecipada) e 3559,43 ± 410,34 Hz (cio). Yajuvendra et al. (2013) verificaram valores entre 4939,19 ± 26,56 e 5296,28 ± 8,39 Hz. Torre et al. (2015) constataram frequências médias para F4 de 1513 ± 16,1 (LFC), 1478 ± 2,5 (HFC) e 2722 ± 34,2 Hz (bezerros).

Comparando com a literatura os valores médios obtidos para o *jitter*, de 1,61 ± 0,63 em confinamento e 3,85 ± 0,81 em manejo, e o *shimmer*, de 14,93% ± 1,7 em confinamento e 19,08 ± 1,8% em manejo, também é possível encontrar semelhanças.

Em relação ao *jitter*, na literatura é possível observar os valores entre 1,09 ± 0,03 e 2,77 ± 0,10% (YAJUVENDRA et al., 2013). No trabalho de Torre et al. (2015) identificaram *jitter* de 2,0 ± 0,18 (LFC), 4,0 ± 0,41 (HFC) e 1,0 ± 0,01% (bezerros). Já Green (2020) constatou em contextos positivos valores entre 1,0 ± 0,2 e 3,0 ± 0,6%, e em contextos negativos entre 3,0 ± 0,6 e 5,0 ± 0,8%.

Analisando o *shimmer* em relação aos trabalhos correlatos, Yajuvendra et al. (2013) identificou valores entre 4,85 ± 0,1 e 8,97 ± 0,31%. Torre et al. (2015) observaram valores para o *shimmer* de 17,0 ± 1,1 (LFC), 17,0 ± 1,0 (HFC) e 15,0 ± 1,0% (bezerros). Já no estudo de Green (2020) é possível observar valores em contextos positivos entre 9,0 ± 1,0 e 17,0 ± 1,6%, e para contextos negativos entre 14,0 ± 2,0 e 18,0 ± 3,0%.

Na análise comparativa entre os resultados obtidos para harmonia na presente pesquisa, de 7,12 ± 0,17 (confinamento) e 7,66 ± 0,18 dB (manejo), e os trabalhos correlatos, o estudo de Yajuvendra et al. (2013) detectaram valores entre 5,86 ± 0,68 e 12,71 ± 0,21 dB, enquanto na pesquisa de Green (2020) foram observadas harmonias em contextos positivos entre 7,87 ± 0,79 e 12,53 ± 1,02 dB, e para contextos negativos harmonias entre 7,72 ± 0,41 e 11,97 ± 0,49 dB.

Para a intensidade da vocalização, a presente pesquisa identificou os valores de 48,74 ± 3,35 dB para os animais em confinamento e 57,59 ± 4,37 dB para os animais em

manejo. Na literatura, Yajuvendra et al. (2013) observaram intensidades entre 60,15 $\pm$ 0,8 e 87,72 $\pm$ 0,16 dB, enquanto na pesquisa de Yeon et al. (2006) verificaram a intensidade de 71,4 $\pm$ 6,2 (alimentação antecipada) e 68,8 $\pm$ 4,2 dB (cio).

A tabela 7 sintetiza os principais resultados obtidos na literatura sobre as diferenças nos parâmetros acústicos das vocalizações de bovinos em condições diversas, proporcionando uma análise mais detalhada das similaridades e diferenças entre os resultados encontrados.

Tabela 7 – Comparação entre os resultados obtidos e a literatura

	Presente trabalho	Yajuvendra et al. (2013)	Torre et al. (2015)	Yeon et al. (2006)	Green (2020)
Média F0 (Hz)	172 e 276	191	81 e 152	214 e 221	183 e 286
Min F0 (Hz)	81 e 132	78	74 e 121	NA	66 e 76
Max F0 (Hz)	436 e 458	352	84 e 198	NA	473 e 559
Média F1 (Hz)	848 e 940	689 e 1015	228 e 391	784 e 832	NA
Média F2 (Hz)	1445 e 1582	1675 e 1942	634 e 1162	1710 e 1570	NA
Média F3 (Hz)	2029 e 2421	3079 e 3412	1064 e 1939	2418 e 2675	NA
Média F4 (Hz)	3435 e 3802	4939 e 5296	1513 e 2722	3818 e 3559	NA
Jitter (%)	1,61 e 3,85	1,09 e 2,77	2 e 4	NA	1 e 5
Shimmer (%)	14 e 19	4,85 e 8,97	15 e 17	NA	9 e 18
Harmonia (dB)	7,12 e 7,66	5,86 e 12,71	NA	NA	7,72 e 11,97
Intensidade (dB)	48 e 57	60 e 87	NA	68 e 71	NA

Os resultados obtidos nesta etapa de análise acústica evidenciam que as vocalizações de bovinos em situações de estresse exibem características distintas daquelas produzidas por bovinos em condições normais, com diferenças especialmente marcantes no que diz respeito às características de frequência. Isso indica que técnicas que exploram essas características, como o MFCC, têm o potencial de serem escolhas promissoras ao desenvolver ferramentas automatizadas para classificação de vocalizações bovinas em contextos de estresse. Além disso, a utilização de redes neurais para essa tarefa pode ser pertinente, uma vez que esses modelos são capazes de aprender padrões complexos nos dados de entrada, permitindo uma classificação mais precisa e eficaz das vocalizações em categorias de estresse e não estresse. Essa abordagem poderia oferecer uma contribuição significativa para o monitoramento da saúde e bem-estar dos bovinos, possibilitando uma intervenção precoce em situações de estresse e melhorando a qualidade de vida dos animais.

6.2 Redes neurais

Para analisar os resultados obtidos no treinamento das diferentes arquiteturas de redes neurais apresentadas na seção 5.3, foram escolhidas as métricas de acurácia, precisão, revocação e *F1-score*. A análise das métricas foi realizada a partir de uma análise de variância (ANOVA) de modo a comparar os valores médios entre as arquiteturas pertencentes as mesmas redes neurais, bem como comparar o desempenho de redes neurais distintas.

Dessa forma, a Tabela 8 apresenta os valores médios obtidos para a acurácia, precisão, revocação e *F1-score* para as três arquiteturas das redes MLP, CNN e LSTM após a execução do treinamento a partir de 5 repetições de validação cruzada.

Tabela 8 – Médias para os parâmetros de acurácia, precisão, revocação e *F1-score*

Rede Neural	**Arquitetura**	**Acurácia**	**Precisão**	**Revocação**	***F1-score***
MLP	Básica	87,72%	90,53%	83,22%	86,72%
	Intermediária	89,32%	91,23%	86,13%	88,61%
	Robusta	90,62%	91,84%	88,38%	90,07%
LSTM	Básica	80,67%	78,10%	83,22%	80,58%
	Intermediária	83,21%	80,66%	85,70%	83,10%
	Robusta	85,59%	82,53%	88,92%	85,61%
CNN	Básica	96,94%	95,22%	98,60%	96,88%
	Intermediária	97,88%	96,74%	98,92%	97,82%
	Robusta	98,81%	98,39%	99,14%	98,76%

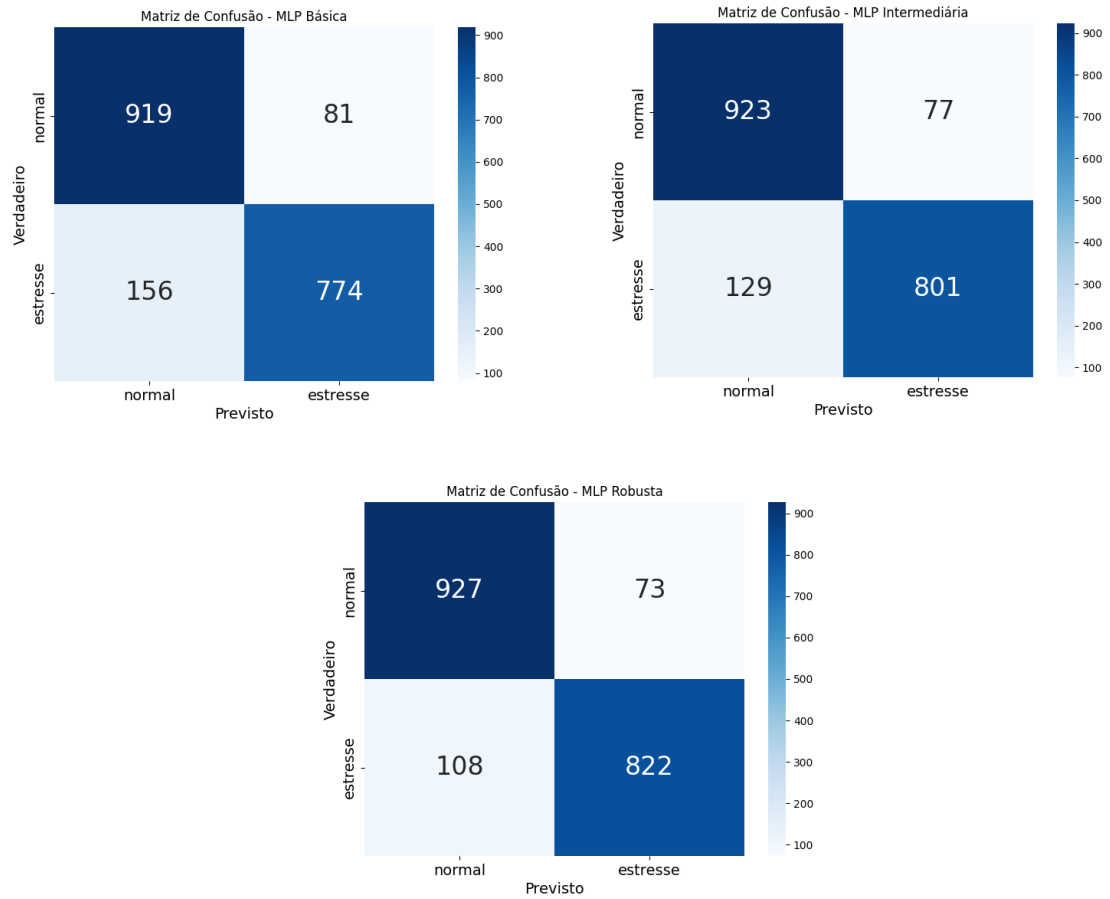
Na tabela, observa-se que as arquiteturas mais complexas apresentaram desempenhos superiores em comparação com aquelas de menor complexidade. No entanto, apesar de sua maior complexidade, evidenciada pelo aumento no número de camadas e de neurônios por camada, os ganhos em acurácia, precisão, revocação e *F1-score* foram modestos. Isso sugere que os recursos adicionais de memória e computação investidos não se traduziram em melhorias proporcionais na capacidade de classificação de vocalizações.

Analisando o desempenho geral das redes neurais, observa-se que as arquiteturas CNN apresentaram os melhores resultados, com acurácias variando entre 96,94% e 98,81%, precisões entre 95,22% e 98,39%, revocação entre 98,60% e 99,14%, e *F1-scores* entre 96,88% e 98,76%. Em seguida, as arquiteturas MLP obtiveram resultados satisfatórios, com acurácias entre 87,72% e 90,62%, precisões entre 90,53% e 91,84%, revocações entre 83,22% e 88,38%, e *F1-scores* entre 86,72% e 90,07%. Por outro lado, as arquiteturas LSTM demonstraram desempenho inferior entre todas as arquiteturas

analisadas, apresentando acurácias entre 80,67% e 85,59%, precisões entre 78,10% e 82,53%, revocações entre 83,22% e 88,92%, e *F1-scores* entre 80,58% e 85,61%.

A Figura 39 apresenta as matrizes de confusão durante o treinamento das arquiteturas MLP implementadas. Observa-se que as arquiteturas obtiveram um desempenho superior na classificação de vocalizações normais (média de 92,3%) em comparação com a classificação de vocalizações de estresse (média de 85,91%). Os resultados mais satisfatórios foram alcançados pela arquitetura robusta, que classificou corretamente 92,7% das vocalizações normais e 88,38% das vocalizações de estresse. Os resultados piores foram obtidos com a arquitetura básica, classificando corretamente 91,9% das vocalizações normais e 83,22% das vocalizações de estresse. Os resultados podem indicar que as arquiteturas MLP tiveram maiores dificuldades de capturar as nuances específicas de vocalizações de estresse, devido a uma variabilidade ou imprevisibilidade desse tipo de vocalização.

Figura 39 – Matrizes de confusão das arquiteturas MLP

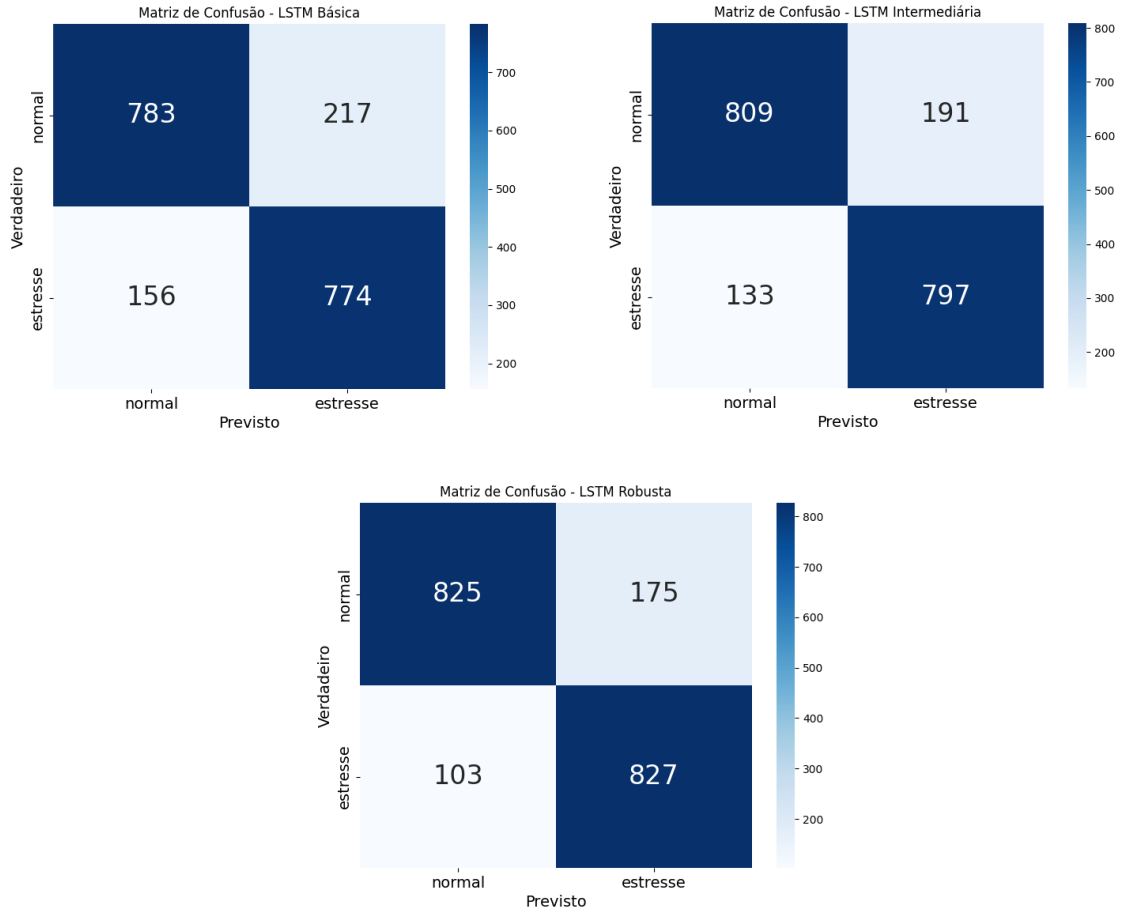


Fonte: Autor (2024)

A Figura 40 ilustra a média das matrizes de confusão durante o treinamento das

arquiteturas LSTM implementadas. Embora as arquiteturas LSTM tenham apresentado desempenho geral inferior em comparação com as arquiteturas MLP, elas se destacaram ao classificar vocalizações de estresse (média de 85,95%) em relação às vocalizações normais (média de 80,57%), uma tendência inversa à observada nas arquiteturas MLP. Destaca-se que a arquitetura robusta obteve os melhores resultados, classificando corretamente 82,5% das vocalizações normais e 88,92% das vocalizações de estresse. O pior desempenho, assim como nas arquiteturas MLP, foi observado na arquitetura básica, que classificou corretamente 76,72% das vocalizações normais e 83,22% das vocalizações de estresse.

Figura 40 – Matrizes de confusão das arquiteturas LSTM

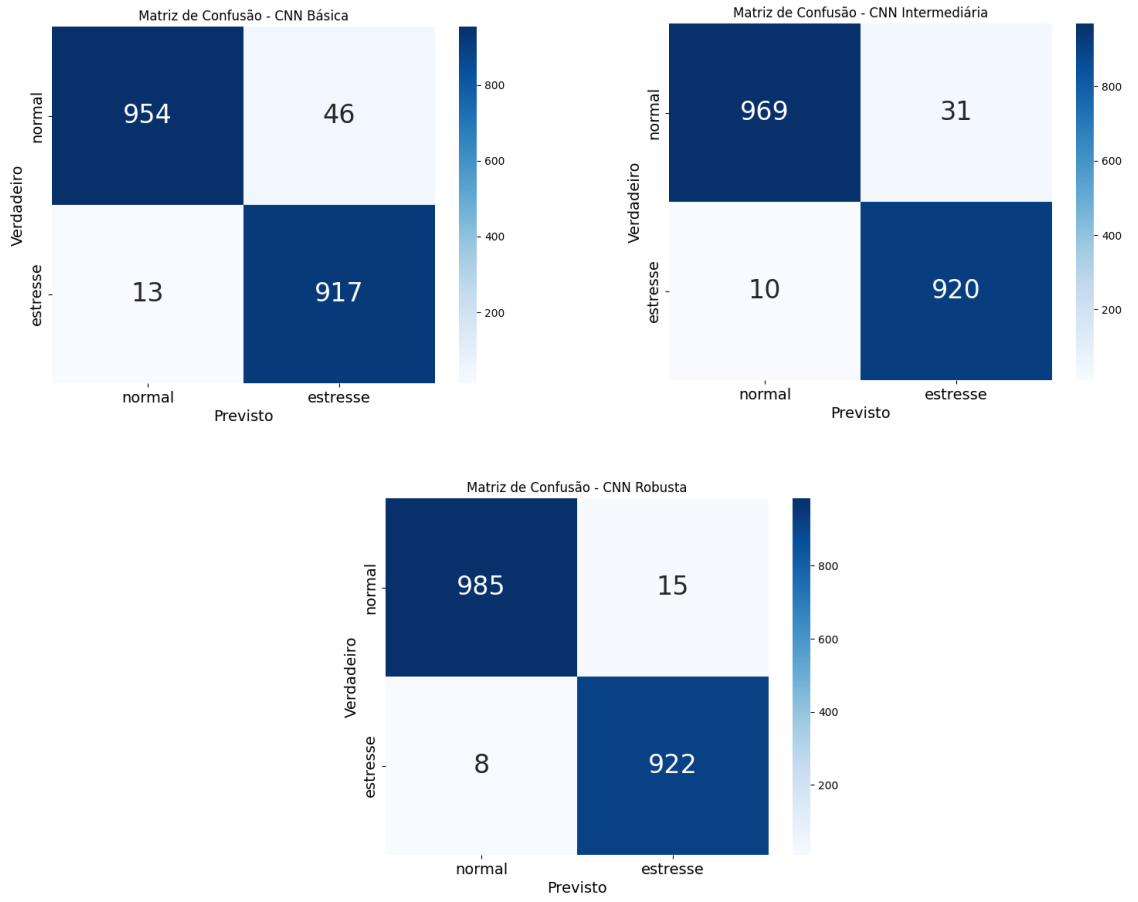


Fonte: Autor (2024)

A Figura 41 apresenta a média dos valores para as matrizes de confusão no treinamento das arquiteturas CNN implementadas. Nota-se que as arquiteturas CNN alcançaram um desempenho bastante satisfatório na classificação das vocalizações, de modo que a arquitetura CNN mais básica superou os resultados de arquiteturas robustas

do tipo MPL e LSTM. Assim como nas redes LSTM, as arquiteturas CNN mostraram melhor desempenho na classificação das vocalizações de estresse (média de 98,89%) em comparação com as vocalizações normais (média de 96,93%). Especificamente, a arquitetura robusta alcançou uma taxa de revocação de 98,5% para vocalizações normais e 99,14% para vocalizações de estresse. Apesar de apresentar um desempenho ligeiramente inferior, a arquitetura básica ainda obteve resultados impressionantes, com revocações de 95,4% das vocalizações normais e 98,6% das vocalizações de estresse. Esses resultados ressaltam a eficácia das arquiteturas CNN na tarefa de classificação de vocalizações animais, evidenciando seu potencial para aplicações práticas em estudos comportamentais e de bem-estar animal.

Figura 41 – Matrizes de confusão das arquiteturas CNN



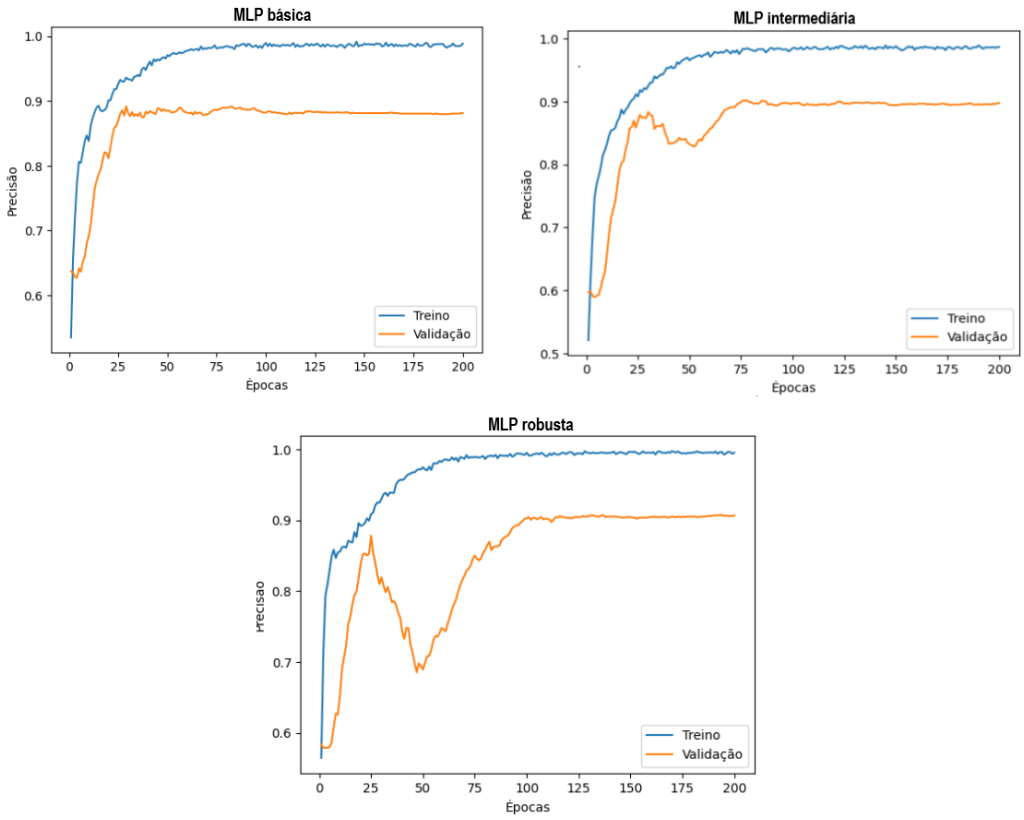
Fonte: Autor (2024)

Outra análise crucial a ser conduzida em relação ao treinamento das arquiteturas implementadas é o acompanhamento do histórico das taxas de perda e acurácia ao longo das épocas de treinamento. Esse monitoramento é essencial devido à sua capacidade

de fornecer entendimento sobre o progresso do treinamento e a estabilidade do modelo. Ao observar o comportamento das curvas de perda e acurácia durante o treinamento, é possível identificar possíveis problemas, como *overfitting* ou *underfitting*, e ajustar os hiperparâmetros do modelo conforme necessário para otimizar o desempenho. Além disso, o histórico das métricas de desempenho fornece uma visão geral do processo de aprendizado da rede neural e pode ajudar a determinar se mais épocas de treinamento são necessárias ou se o modelo já convergiu para uma solução satisfatória.

A Figura 42 exibe os gráficos de precisão ao longo do treinamento das arquiteturas MLP. Nos gráficos é perceptível um rápido aumento na precisão nas primeiras épocas para todas as arquiteturas e, a partir desse ponto, comportamentos distintos entre elas se evidenciaram. A arquitetura básica mantém uma precisão quase constante até a época 200. Por outro lado, a arquitetura intermediária sofre uma perda de precisão até a época 50, recuperando-se em seguida e alcançando um platô na época 75, mantendo-se estável até o final do treinamento. Já a arquitetura robusta apresentou uma grande queda na precisão até a época 50, seguida por uma recuperação até alcançar um platô por volta da época 100.

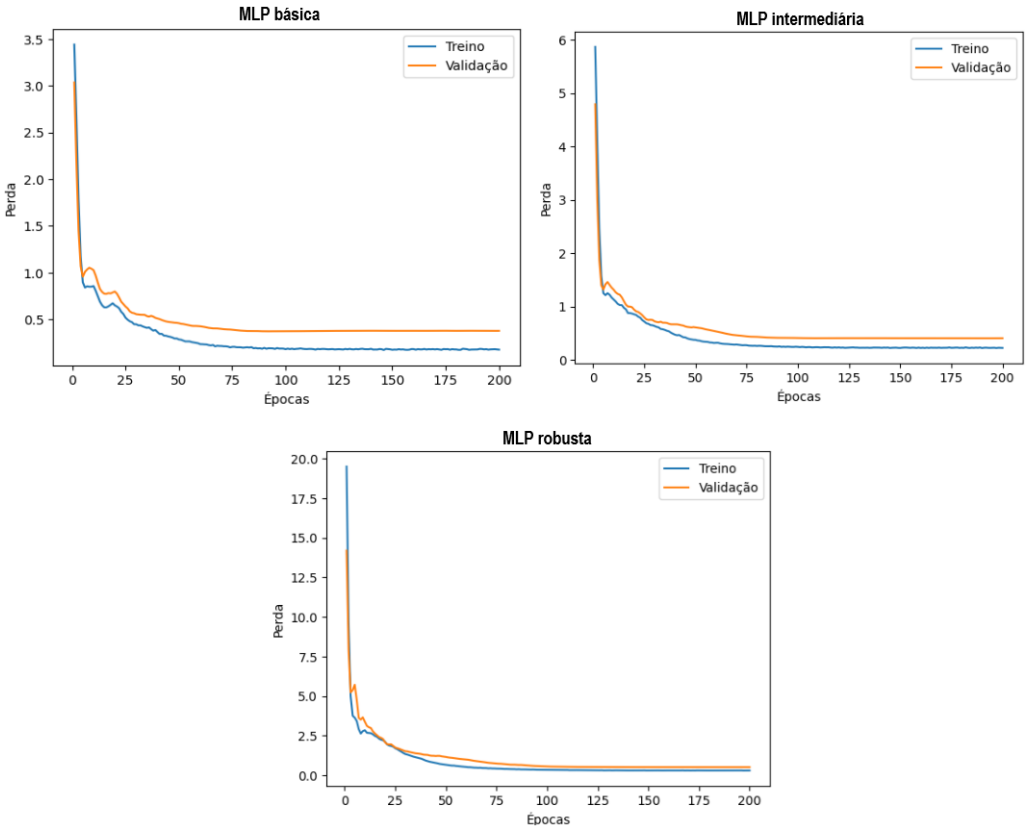
Figura 42 – Precisões nos treinamentos das arquiteturas MLP



Fonte: Autor (2024)

A Figura 43 ilustra os gráficos de perda durante os treinamentos das três arquiteturas MLP. Observa-se as três arquiteturas um padrão consistente onde as perdas na etapa de validação são comparativamente próximas às perdas na etapa de treinamento. Entretanto, à medida que a complexidade da arquitetura aumenta, a discrepância entre as perdas de validação e treinamento diminui. Esse fenômeno sugere que arquiteturas mais complexas têm uma capacidade maior de generalização, visto que conseguem manter um desempenho semelhante em ambos os conjuntos de dados.

Figura 43 – Perdas nos treinamentos das arquiteturas MLP

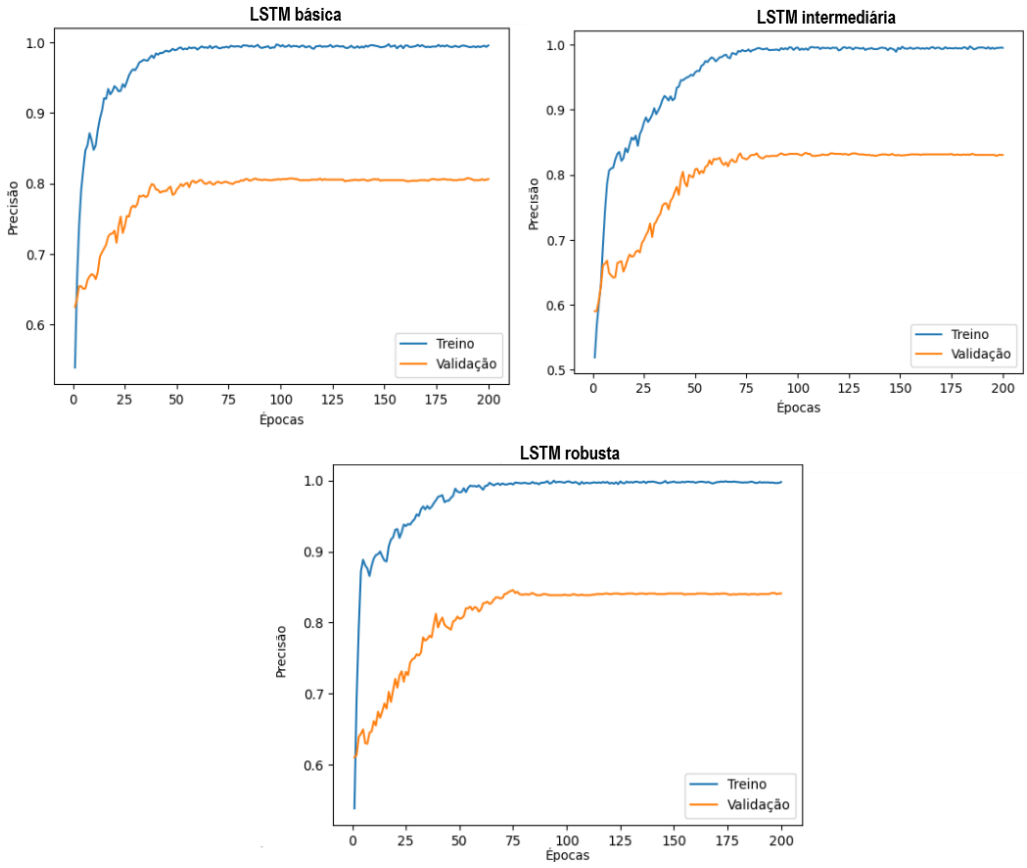


Fonte: Autor (2024)

A Figura 44 apresenta o gráficos de precisão durante o treinamento das três arquiteturas LSTM. Para essa rede, nota-se um padrão consistente em todas as arquiteturas, caracterizado por um aumento gradual da precisão ao longo das épocas até atingir o platô por volta da época 75, mantendo-se estável até o final do treinamento. Embora as arquiteturas tenham demonstrado comportamentos semelhantes, diferenças sutis foram observadas em relação aos níveis de precisão alcançados, sendo que as arquiteturas mais robustas apresentaram valores ligeiramente maiores.

A Figura 45 exibe os gráficos de perda durante o treinamento das três arquiteturas

Figura 44 – Precisões nos treinamentos das arquiteturas LSTM

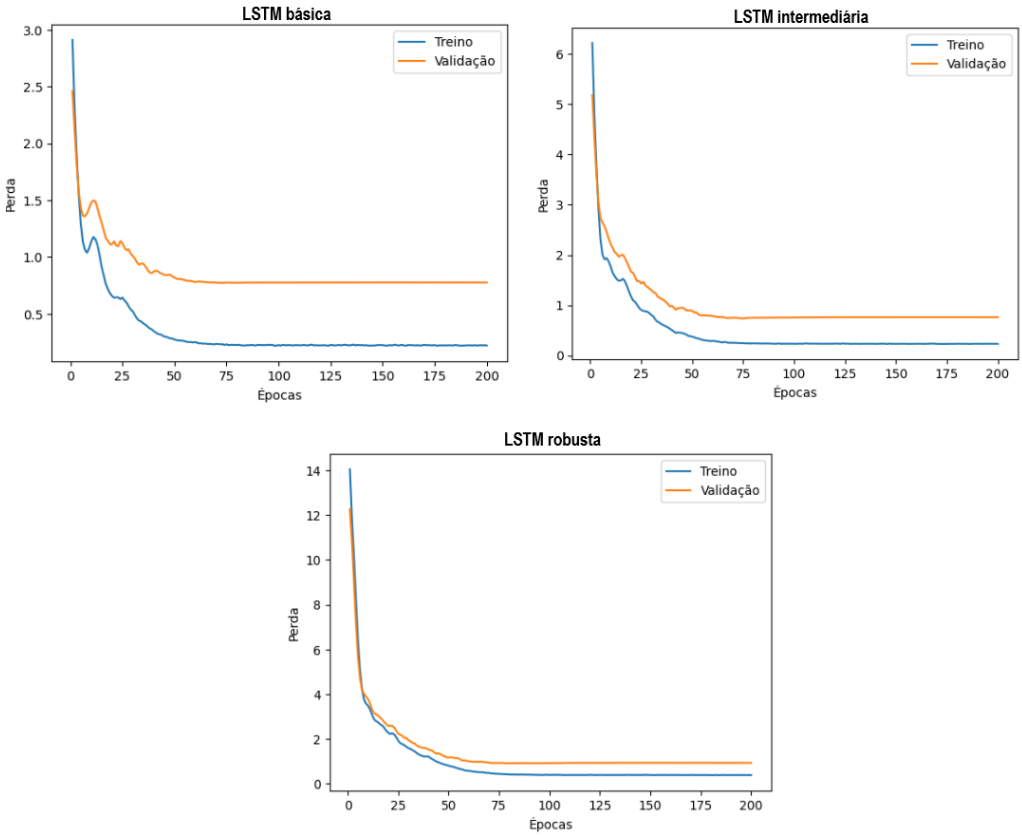


Fonte: Autor (2024)

LSTM. Em todos os casos, nota-se que a perda converge para um valor estável após cerca de 50 épocas, mantendo-se praticamente constante até o término do treinamento. Uma distinção relevante entre as arquiteturas é a discrepância entre as perdas de treinamento e validação, sendo que as arquiteturas mais complexas demonstraram diferenças menores entre esses dois conjuntos de dados. Esse fenômeno sugere que as arquiteturas mais robustas conseguiram generalizar melhor os padrões aprendidos durante o treinamento para dados não vistos.

A Figura 46 apresenta os gráficos de precisão durante o treinamento das três arquiteturas CNN, onde Observa-se comportamentos distintos entre as arquiteturas. A arquitetura básica alcançou estabilidade rapidamente, por volta da época 20, mantendo-se constante até o término do treinamento. Já a arquitetura intermediária teve um início de treinamento com perda significativa de precisão, que gradualmente se recuperou até atingir um platô por volta da época 100, mantendo-se estável até o final. Por sua vez, a arquitetura robusta também apresentou uma queda inicial na precisão, embora menos acentuada que a intermediária. Posteriormente, observou-se alguma variação, mas ela

Figura 45 – Perdas nos treinamentos das arquiteturas LSTM



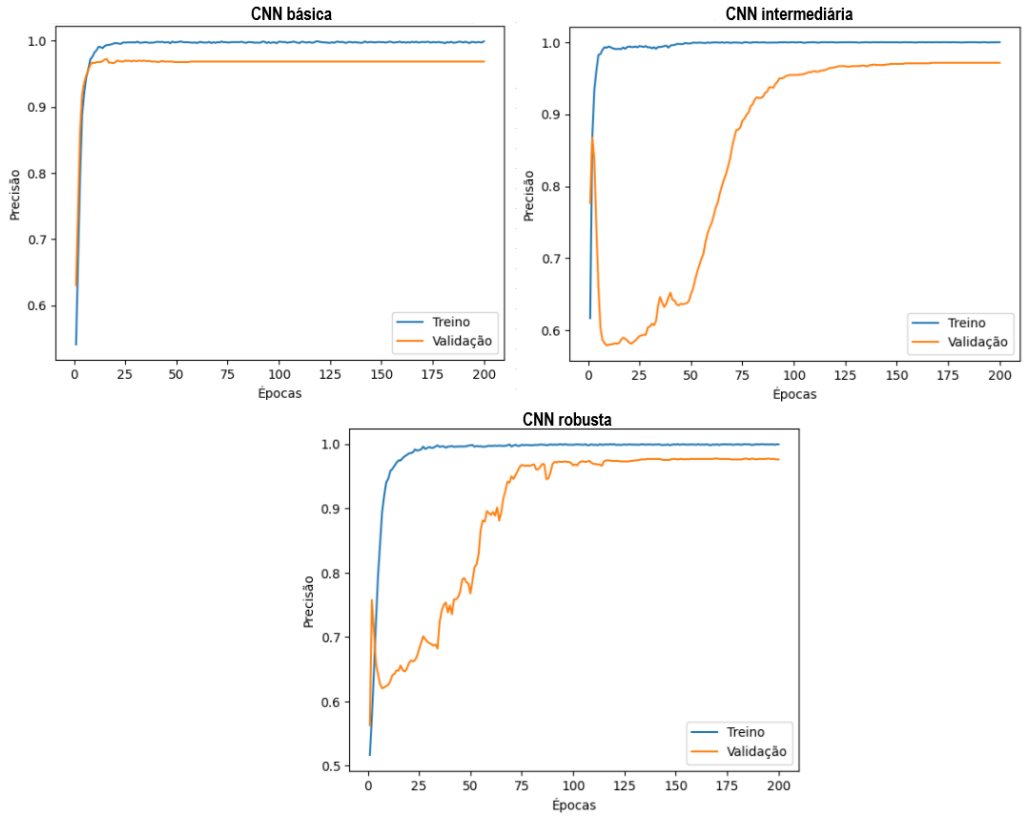
Fonte: Autor (2024)

alcançou um platô por volta da época 100, mantendo-se praticamente constante até o término do treinamento.

A Figura 47 apresenta os gráficos de perda durante os treinamentos das três arquiteturas CNN. Na arquitetura básica, observa-se que a perda se estabiliza rapidamente no início do treinamento, mantendo-se constante até o final. Já na arquitetura intermediária, a curva de perda é mais suave, alcançando o platô por volta da época 100. Por sua vez, a arquitetura robusta também atinge a estabilidade logo no início do treinamento, no entanto, diferencia-se da básica pela proximidade entre as perdas de treinamento e validação, enquanto na básica essa diferença é mais significativa.

Para analisar os recursos computacionais necessários para o treinamento e validação de cada uma das arquiteturas de redes neurais implementadas, foram coletadas informações sobre o tempo despendido e o consumo de memória gastos durante as etapas de treinamento e validação das redes. Essa análise é importante para a seleção dos modelos mais adequados para diferentes cenários de aplicação quanto à alocação eficiente de recursos computacionais, para garantir não apenas a eficiência e o desempenho

Figura 46 – Precisões nos treinamentos das arquiteturas CNN



Fonte: Autor (2024)

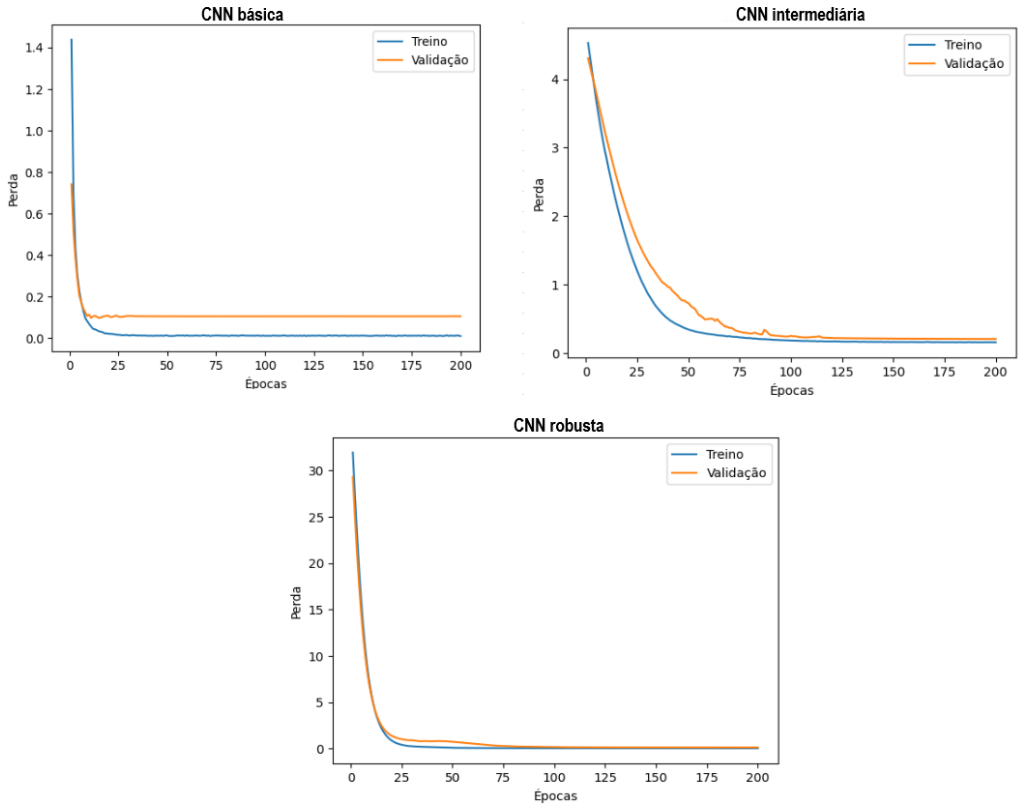
dos modelos, mas também para otimizar o uso dos recursos disponíveis. A Tabela 9 apresenta um comparativo sobre recursos computacionais utilizados por cada arquitetura implementada.

Tabela 9 – Recursos de tempo e memória utilizados no treinamento de validação das redes neurais artificiais

Arquitetura	**Tempo (s)**	**Memória (MB)**
MLP básica	27,85	8,84
MLP intermediária	31,48	14,51
MLP robusta	52,2	31,02
LSTM básica	77,89	26,78
LSTM intermediária	157,54	74,82
LSTM robusta	823,51	265,21
CNN básica	316,85	145
CNN intermediária	466,39	165,44
CNN robusta	984,93	200,79

A Figura 48 ilustra um gráfico comparativo do tempo despendido pelas diferentes arquiteturas durante o treinamento e validação. Observa-se claramente que as redes

Figura 47 – Perdas nos treinamentos das arquiteturas CNN

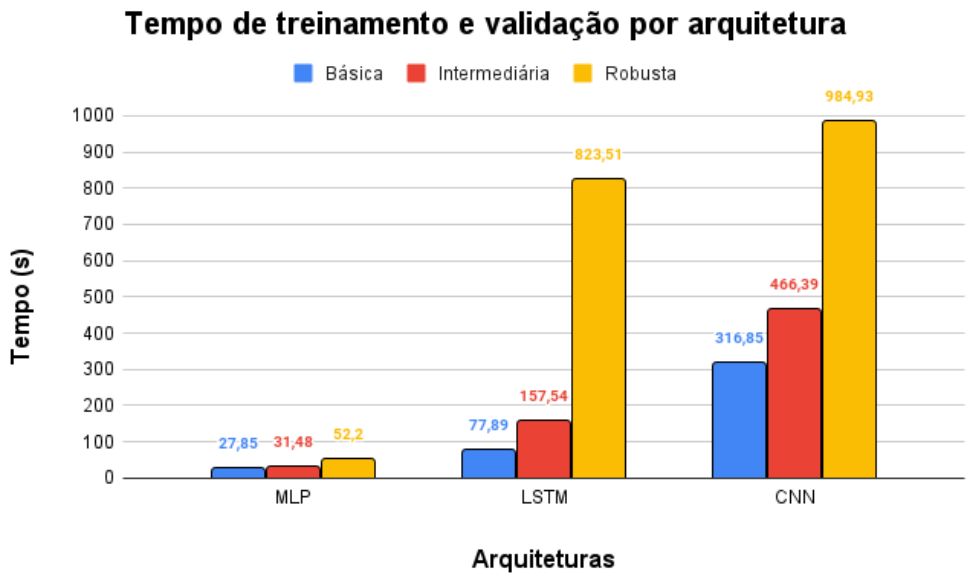


Fonte: Autor (2024)

MLP demandaram significativamente menos tempo em comparação com as redes LSTM e CNN. Por outro lado, as redes CNN demandaram mais tempo para o término do treinamento, o que era esperado devido à sua natureza computacionalmente complexa, especialmente devido à operação de convolução. Além disso, é notável que o tempo necessário aumentou à medida que a complexidade da rede aumentou. Este padrão evidencia uma relação direta entre a complexidade da arquitetura e o tempo exigido para o treinamento e validação.

A Figura 49 apresenta o gráfico comparativo do consumo de memória durante o treinamento das redes neurais. Os resultados revelam que as redes MLP requereram quantidades relativamente menores de memória, enquanto as redes CNN apresentaram, em média, um consumo mais elevado. No entanto, destaca-se que a arquitetura LSTM robusta foi a variante que registrou o maior consumo de memória durante o treinamento. Esse fenômeno pode ser atribuído à necessidade das redes LSTM de armazenar informações de estado ao longo do tempo durante a fase de retropropagação, o que demanda uma alocação adicional de memória. Assim como ocorreu para o recurso

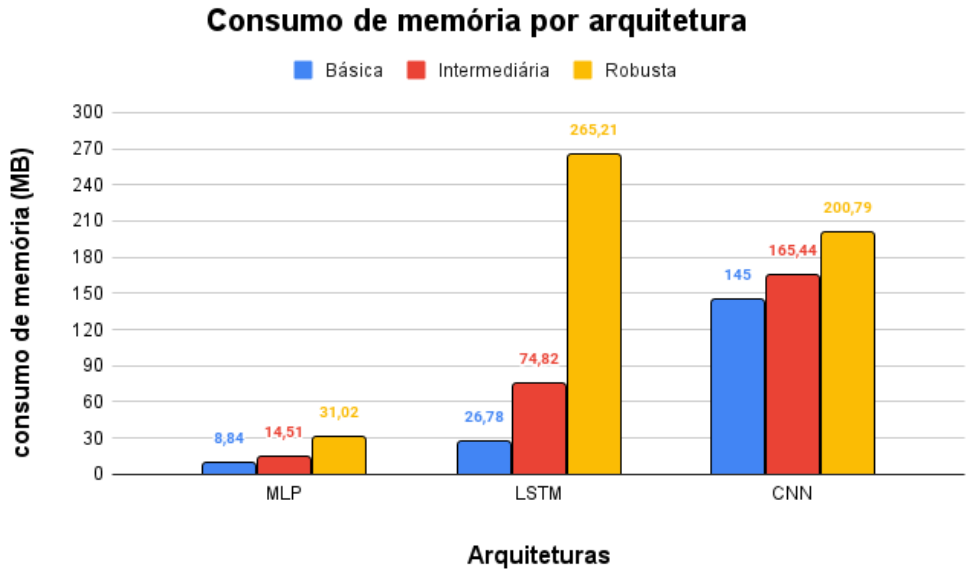
Figura 48 – Tempo de treinamento e validação das redes



Fonte: Autor (2024)

de tempo, o consumo de memória também aumentou conforme maior era a complexidade da rede.

Figura 49 – Consumo de memória durante o treinamento e validação das redes



Fonte: Autor (2024)

Para avaliar o desempenho das arquiteturas implementadas e determinar se as discrepâncias observadas nos resultados entre os tipos de redes neurais e arquiteturas

foram devidas às diferenças nas complexidades das arquiteturas ou apenas ao acaso, foi realizada uma análise de variância (ANOVA). Inicialmente, as três arquiteturas de cada tipo de rede neural foram analisadas separadamente e, em seguida, todas as arquiteturas foram examinadas em conjunto.

O teste de hipóteses foi conduzido utilizando o *F1-score* como a métrica de análise, devido à sua robustez em comparação com a precisão e a acurácia. Foram comparadas as médias do *F1-score* entre os grupos de arquiteturas do mesmo tipo. Portanto, as hipóteses formuladas são as seguintes:

- Hipótese Nula (H0): Não há diferença significativa entre as médias do *F1-score* entre as arquiteturas. Qualquer variação observada é atribuída ao acaso.
- Hipótese Alternativa (HA): Existe uma diferença 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 10 apresenta os resultados da ANOVA para os três modelos de arquiteturas dentro de cada tipo de rede neural.

Tabela 10 – Análise de variância por testes de Tukey entre os modelos de RN

Grupo vs Grupo	**Médias *F1-score***		**Diferença**	***P-value***
MLP básica vs MLP intermediária	86,72	88,61	1,89	0,143
MLP básica vs MLP robusta	86,72	90,07	3,35	0,005**
MLP intermediária vs MLP robusta	88,61	90,07	1,46	0,331
LSTM básica vs LSTM intermediária	80,58	83,10	2,52	0,035*
LSTM básica vs LSTM robusta	80,58	85,61	5,03	< 0,001***
LSTM intermediária vs LSTM robusta	83,10	85,61	2,51	0,034*
CNN básica vs CNN intermediária	96,88	97,82	0,94	0,158
CNN básica vs CNN robusta	96,88	98,76	1,88	0,002**
CNN intermediária vs CNN robusta	97,82	98,76	0,94	0,22

A análise comparativa entre os modelos de redes neurais revelou diferenças significativas em relação ao desempenho, conforme evidenciado pelos testes de Tukey. Nas arquiteturas MLP, foi observada uma melhoria estatisticamente significativa no desempenho da rede ao transitar da configuração básica para a robusta (*P-value* = 0,005). No entanto, não houve melhorias significativas entre as arquiteturas básica e intermediária, bem como entre a intermediária e a robusta. Por outro lado, para os modelos LSTM, todas as comparações entre os diferentes níveis de complexidade indicaram diferenças estatisticamente significativas, representando os benefícios de

aumentar a robustez da arquitetura. Quanto aos modelos CNN, apenas foi observada 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 mais poder de classificação.

Também foi realizada a análise estatística entre todas as arquiteturas implementadas com o objetivo de verificar quais arquiteturas alcançaram resultados estatisticamente melhores do que outras. A Tabela 11 apresenta os resultados da ANOVA, onde arquiteturas com de mesmas letras não apresentaram diferenças significativas entre suas médias.

Tabela 11 – Análise de variância por testes de Tukey para todas as arquiteturas implementadas

Arquitetura	**Média *F1-score***	**Grupo**
LSTM básica	80,58	a
LSTM intermediária	83,10	b
LSTM robusta	85,61	c
MLP básica	86,72	cd
MLP intermediária	88,61	de
MLP robusta	90,07	e
CNN básica	96,88	f
CNN intermediária	97,82	fg
CNN robusta	98,76	g

Os resultados revelam que as redes CNN superaram as demais, evidenciado pela superioridade estatística em todas as suas arquiteturas. Em contraste, entre as redes MLP e LSTM, apenas as variantes LSTM robusta e MLP básica não apresentaram diferenças estatisticamente significativas em suas médias. De maneira geral, as redes LSTM apresentaram um desempenho inferior às MLP, enquanto as CNN destacaram-se como as mais eficientes entre todas as arquiteturas analisadas.

Realizando um comparativo com os resultados encontrados na literatura, é possível destacar a utilização de redes neurais do tipo CNN para classificação de vocalizações nos trabalhos de Pandeya et al. (2022), Sattar (2022), Şaşmaz e Tek (2018), Vidana-Vila et al. (2023), Jung et al. (2021). Quanto à utilização de redes neurais

recorrentes, podem ser elencados os trabalhos de Huang et al. (2021), Duan et al. (2021). Também houve trabalhos que utilizaram as duas redes como Shorten (2023) e Gavojdian et al. (2023). A Tabela 12 apresenta um comparativo entre os resultados obtidos pela presente pesquisa e a literatura. Ainda que os resultados não possam ser comparados diretamente, visto que foram feitos com bases de dados distintas para outras aplicações, verifica-se que os resultados atingidos neste trabalho são compatíveis com aplicações de CNN em problemas de reconhecimento de som.

Tabela 12 – Comparação entre os resultados obtidos e a literatura

Trabalho	**RN utilizada**	**Característica analisada**	**Animal**	**Resultados**
Presente pesquisa	MLP, LSTM, CNN	MFCC	Bovinos	Acc = 90,24%, 85,04%, 98,74% Rec = 88,35%, 88,93%, 99,11% F1 = 90,05%, 85,60%, 98,74%
Sattar (2022)	CNN	MFCC	Bovinos	Acurácia = 84%
Jung et al. (2021)	CNN	MFCC	Bovinos	Acurácia = 94,18%
Şaşmaz e Tek (2018)	CNN	MFCC	Diferentes animais	Acurácia = 75%
Pandeya et al. (2022)	CNN	Espectrograma	Bovinos	F1 = 70,90%
Shorten (2023)	CNN	Espectrograma	Bovinos	Acurácia = 96,2%
Vidana-Vila et al. (2023)	CNN	Espectrograma	Bovinos	F1 = 61,7%
Gavojdian et al. (2023)	CNN + GRU	23 parâmetros vocais	Bovinos	F1 = 89,4%
Duan et al. (2021)	LSTM	MFCC	Ovinos	F1 = 97,41%
Huang et al. (2021)	LSTM	STZ (*short time zero*) STE (*short time energy*)	Aves	Revocação = 97%

Ao analisar a tabela, destaca-se que a maioria dos estudos encontrados na literatura empregou redes neurais convolucionais (CNN) para a classificação, além de utilizar os coeficientes MFCC como características de análise acústica. Também é observa-se o uso do espectrograma como uma característica de interesse para o modelo de classificação.

Os resultados obtidos nesta pesquisa são corroborados por estudos anteriores que adotaram abordagens semelhantes. Destacam-se os trabalhos de Sattar (2022), Jung et al. (2021) e Duan et al. (2021), que alcançaram resultados significativos na

classificação de sons. Esses estudos registraram taxas de acerto de 84%, 94,18% e 97,41%, respectivamente. Vale ressaltar que o estudo de Duan et al. (2021) se concentrou na identificação do comportamento ruminante em ovinos, apresentando um contexto diferente em relação à proposta desta pesquisa.

Outro estudo que merece destaque é o de Gavojdian et al. (2023), que empregou uma rede neural híbrida combinando uma rede convolucional com uma recorrente, atingindo 89,4% na métrica *F1-score*. No trabalho foram utilizados como entrada para a rede muitos dos parâmetros analisados na seção 6.1, tais como a média, máximo e mínimo da frequência fundamental (F0), os formantes F1-F4 e a medida de harmonia. Esse estudo ressalta a viabilidade desses parâmetros na classificação de vocalizações bovinas.

Os resultados encontrados pela pesquisa 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 para a classificação, as redes CNN se mostraram mais promissoras em comparação com as MLP e LSTM, onde até mesmo a arquitetura mais básica implementada alcançou melhores resultados. Essa tendência pode ser um indício para a preferência predominante na literatura pelo uso de redes CNN na análise acústica de vocalizações animais.

Em comparativo entre as arquiteturas de mesma rede, todas as variantes robustas demonstraram resultados estatisticamente superiores. 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 em comparação com as redes menos complexas. Portanto, ao escolher uma arquitetura de rede neural é importante avaliar esses aspectos computacionais, principalmente em sistemas de tempo real, onde as limitações de *hardware* podem ser restritivas para o uso eficaz de redes neurais mais robustas.

7 CONCLUSÃO

A pecuária bovina de corte é uma das principais fontes de renda no Brasil. No entanto, enfrenta desafios para aprimorar sua produtividade, sendo um ponto crucial a crescente exigência dos consumidores e dos países exportadores quanto à qualidade da carne bovina e a garantia de bem-estar dos animais, demandando produtos de maior qualidade.

O estresse bovino é um dos principais fatores causadores da perda de qualidade do produto final. Situações estressantes podem desencadear reações fisiológicas que resultam em carne escura e dura, reduzindo o valor agregado do produto. Em resposta a essas preocupações, 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.

Inicialmente, esta pesquisa visava desenvolver um protótipo para coletar dados sobre o estresse animal durante o transporte em caminhões. No entanto, devido a restrições das empresas de transporte, esse objetivo não foi alcançado. Assim, a pesquisa mudou de abordagem, concentrando-se na análise do estresse animal em diferentes contextos e na avaliação do desempenho de diferentes arquiteturas de redes neurais, incluindo *Multilayer Perceptron* (MLP), *Long Short-Term Memory* (LSTM) e *Convolutional Neural Network* (CNN), para a detecção de estresse em bovinos.

Dessa maneira, foram coletadas 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. No total foram coletadas 357 vocalizações em confinamento e 186 em manejo.

As vocalizações foram tratadas e filtradas, e a técnica de *Mel Frequency Cepstrum Coefficients* (MFCC) foi empregada para a extrair suas características. Pelos resultados obtidos pelas redes neurais atestou-se a eficácia do MFCC em capturar as características essenciais das vocalizações, possibilitando a distinção entre momentos de estresse e não estresse.

O estudo não apenas avaliou a viabilidade de redes neurais da identificação de estresse, mas também conduziu uma análise estatística dos parâmetros acústicos das vocalizações. Nas análises, foram considerados os principais parâmetros frequentemente investigados na literatura para distinguir vocalizações em diferentes contextos, incluindo frequência fundamental, formantes espectrais, *shimmer*, *jitter*, harmonia e intensidade. Os

resultados dessa análise acústica, em concordância com estudos na literatura, revelaram que bovinos submetidos a condições psicologicamente estressantes tendem a produzir vocalizações com parâmetros significativamente diferentes daqueles em condições não estressantes.

No estudo das redes neurais, todas as arquiteturas implementadas alcançaram uma acurácia satisfatória, variando de 80,59% a 98,74%. A análise revelou que as redes CNN apresentaram melhor desempenho na detecção do estresse bovino usando as características extraídas pelo MFCC, enquanto as redes LSTM obtiveram os resultados mais baixos. Além disso, o estudo também evidenciou que redes mais complexas conseguiram generalizar melhor os dados, resultando em melhores desempenhos. Contudo, mesmo com resultados estatisticamente melhores, é importante ressaltar que o aumento da complexidade das redes também resultou em um aumento de recursos computacionais.

Em comparação com a literatura, a pesquisa obteve resultados melhores no que diz respeito ao emprego de redes CNN na tarefa de identificação de vocalizações. No entanto, enquanto em relação às redes LSTM os resultados foram inferiores. Vale ressaltar que nem todos trabalhos correlatos eram em contextos de bovinos, bem como os procedimentos, as técnicas e as ferramentas utilizados para coleta e processamento dos dados foram distintos, o que pode ser um fator impactante na distinção entre os resultados alcançados pelos estudos encontrados na literatura e o presente trabalho.

Os resultados da pesquisa confirmaram a hipótese levantada de que os sons emitidos por animais estressados são distintos dos emitidos por animais tranquilos. Tanto a análise acústica, que verificou diferenças significativas nos parâmetros acústicos, quanto o treinamento das redes neurais, que tiveram êxito em distinguir as vocalizações, evidenciaram essa diferença, reforçando a validade da hipótese inicial.

Em relação aos objetivos da pesquisa, com exceção da coleta de dados de transporte, todos os demais foram alcançados com sucesso. Apesar da impossibilidade em coletar vocalizações durante o transporte, foi possível coletar, processar e analisar vocalizações de estresse e não estresse durante o confinamento e manejo. Esses dados se mostraram confiáveis e adequados para a realização da pesquisa, proporcionando descobertas relevantes sobre o estresse bovino e sua detecção por meio de análise acústica e redes neurais.

O objetivo de conduzir um estudo comparativo de redes neurais para identificação de estresse em bovinos também foi alcançado. A pesquisa evidenciou a eficácia

das redes neurais nessa tarefa e permitiu a comparação do desempenho de diferentes arquiteturas. Isso pode 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.

Em relação às oportunidades de melhorias, é importante destacar a necessidade de expandir a base de dados de vocalizações, tanto em condições de estresse quanto de não estresse. A inclusão de vocalizações durante o transporte também é essencial para uma avaliação mais abrangente dos resultados das redes neurais, proporcionando maior robustez aos resultados obtidos. Além disso, vale destacar a oportunidade de explorar outras arquiteturas de redes neurais, incluindo a combinação de diferentes tipos de arquiteturas, conforme observado em estudos anteriores na literatura. Essas abordagens podem oferecer observações adicionais e aprimorar ainda mais a capacidade dos classificadores de vocalizações bovinas em identificar o estresse com maior precisão e confiabilidade.

Como trabalho futuro, pretende-se integrar os equipamentos desenvolvidos em um caminhão de transporte de gado, visando verificar a viabilidade de identificação e alerta do estresse animal em tempo real durante o transporte. Essa abordagem tem o potencial de contribuir para a garantia do bem-estar animal durante essa importante etapa da produção, permitindo a implementação de medidas para mitigar e resolver situações que possam estar causando estresse aos animais. Além disso, as bases de dados serão tornadas públicas, dentro dos princípios FAIR de ciência aberta.

Em suma, esta pesquisa ofereceu uma contribuição significativa para o entendimento e a aplicação de técnicas de análise acústica e de redes neurais na identificação de estresse em bovinos. Os resultados obtidos confirmaram a eficácia das redes neurais, especialmente as CNNs, na detecção de estresse a partir de vocalizações bovinas processadas através do MFCC. Além disso, a análise estatística dos parâmetros acústicos reforçou a distinção entre vocalizações de estresse e não estresse. Embora os resultados tenham sido promissores, há espaço para melhorias futuras, incluindo a expansão da base de dados para incluir vocalizações em condições adicionais e a exploração de outras arquiteturas de redes neurais. Essas iniciativas podem aprimorar ainda mais a precisão e a generalização dos modelos, consolidando assim o avanço no desenvolvimento de métodos não invasivos para monitoramento do bem-estar animal na indústria pecuária.

REFERÊNCIAS

ARKSEY, H.; O'MALLEY, L. Scoping studies: Towards a methodological framework. **Int. J. Social Research Methodology**, Routledge, v. 8, n. 1, p. 19–32, 2005.

BISONG, E.; BISONG, E. Google colaboratory. **Building machine learning and deep learning models on google cloud platform: a comprehensive guide for beginners**, Springer, p. 59–64, 2019.

BOERSMA, P. Praat: doing phonetics by computer [computer program]. **http://www. praat. org/**, 2011.

CHELOTTI, J. O. et al. A real-time algorithm for acoustic monitoring of ingestive behavior of grazing cattle. **Computers and Electronics in Agriculture**, Elsevier, v. 127, p. 64–75, 2016.

CHRISTENSEN, M. G. **Introduction to Audio Processing**. [S.l.]: Springer, 2019.

CHUNG, Y. et al. Automatic detection of cow's oestrus in audio surveillance system. **Asian-Australasian journal of animal sciences**, Asian-Australasian Association of Animal Production Societies (AAAP), v. 26, n. 7, p. 1030, 2013.

CLAPHAM, W. M. et al. Acoustic monitoring system to quantify ingestive behavior of free-grazing cattle. **Computers and Electronics in Agriculture**, Elsevier, v. 76, n. 1, p. 96–104, 2011.

COSTA, M. J. Paranhos da. Ambiência na produção de bovinos de corte a pasto. **Anais de Etologia**, v. 18, p. 26–42, 2000.

DAMTEW, A. et al. The effect of long distance transportation stress on cattle: a review. **Biomedical Journal**, v. 2, p. 5, 2018.

DAVE, N. Feature extraction methods LPC, PLP and MFCC in speech recognition. **International journal for advance research in engineering and technology**, v. 1, n. 6, p. 1–4, 2013.

DELLER, J. R.; PROAKIS, J. G.; HANSEN, J. H. Discrete-time processing of speech signals. In: INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. [S.l.], 2000.

DESHMUKH, O. et al. Vocalization patterns of dairy animals to detect animal state. In: IEEE. **Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)**. [S.l.], 2012. p. 254–257.

DUAN, G. et al. Short-term feeding behaviour sound classification method for sheep using LSTM networks. **International Journal of Agricultural and Biological Engineering**, v. 14, n. 2, p. 43–54, 2021.

GAVOJDIAN, D. et al. Bovinetalk: Machine learning for vocalization analysis of dairy cattle under negative affective states. **arXiv preprint arXiv:2307.13994**, 2023.

GERHARD, D. et al. **Pitch extraction and fundamental frequency: History and current techniques**. [S.l.]: Department of Computer Science, University of Regina Regina, SK, Canada, 2003.

GÉRON, A. Hands-on machine learning with scikit-learn. **Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems**, O'Reilly Media, v. 1, 2019.

GREEN, A. C. **Decoding Holstein-Friesian dairy cattle vocalisations: Applications for welfare assessment**. Tese (Doutorado) — University of Sydney, 2020.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007.

HOFFMAN, L. C.; LÜHL, J. Causes of cattle bruising during handling and transport in namibia. **Meat Science**, Elsevier, v. 92, n. 2, p. 115–124, 2012.

HOPKINS, D. L.; BRUCE, H.; LI, D. Final report–causes and contributing factors to dark cutting: Current trends and future directions. 2016.

HUANG, J. et al. An intelligent method for detecting poultry eating behaviour based on vocalization signals. **Computers and Electronics in Agriculture**, v. 180, p. 105884, 2021. ISSN 0168-1699.

IBGE. Indicadores da produção pecuária. 2018.

IKEDA, Y.; ISHII, Y. Recognition of two psychological conditions of a single cow by her voice. **Computers and electronics in agriculture**, Elsevier, v. 62, n. 1, p. 67–72, 2008.

JAHNS, G. Call recognition to identify cow conditions—a call-recogniser translating calls to text. **Computers and electronics in agriculture**, Elsevier, v. 62, n. 1, p. 54–58, 2008.

JORQUERA-CHAVEZ, M. et al. 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**, Elsevier, v. 156, p. 11–22, 2019.

JUNG, D.-H. et al. Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering. **Animals**, MDPI, v. 11, n. 2, p. 357, 2021.

KITCHENHAM, B. **Procedures for Performing Systematic Reviews**. Keele, 2004.

LACERDA, D. P. et al. Design Science Research: método de pesquisa para a Engenharia de Produção. **Gest. Prod.**, São Carlos, v. 20, n. 4, p. 741–761, 2013.

LEE, J. et al. Stress detection and classification of laying hens by sound analysis. **Asian-Australasian journal of animal sciences**, Asian-Australasian Association of Animal Production Societies (AAAP), v. 28, n. 4, p. 592, 2015.

LEE, J. et al. Formant-based acoustic features for cow's estrus detection in audio surveillance system. In: IEEE. **2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)**. [S.l.], 2014. p. 236–240.

LINHART, P. et al. Expression of emotional arousal in two different piglet call types. **PloS one**, Public Library of Science San Francisco, CA USA, v. 10, n. 8, p. e0135414, 2015.

MAIGROT, A.-L.; HILLMANN, E.; BRIEFER, E. F. Encoding of emotional valence in wild boar (sus scrofa) calls. **Animals**, MDPI, v. 8, n. 6, p. 85, 2018.

MANTEUFFEL, G.; SCHÖN, P. C. Measuring pig welfare by automatic monitoring of stress calls. **Agrartechische Berichte**, v. 29, n. 1, 2002.

MATTHES, E. **Python crash course: A hands-on, project-based introduction to programming**. [S.l.]: no starch press, 2023.

MCLOUGHLIN, I. **Applied speech and audio processing: with Matlab examples**. [S.l.]: Cambridge University Press, 2009.

MCLOUGHLIN, M. P.; STEWART, R.; MCELLIGOTT, A. G. Automated bioacoustics: methods in ecology and conservation and their potential for animal welfare monitoring. **Journal of the Royal Society Interface**, The Royal Society, v. 16, n. 155, p. 20190225, 2019.

MEEN, G. et al. Sound analysis in dairy cattle vocalisation as a potential welfare monitor. **Computers and Electronics in Agriculture**, Elsevier, v. 118, p. 111–115, 2015.

MINKA, N.; AYO, J. Effects of loading behaviour and road transport stress on traumatic injuries in cattle transported by road during the hot-dry season. **Livestock Science**, Elsevier, v. 107, n. 1, p. 91–95, 2007.

MOLENTO, C. F. M. Bem-estar e produção animal: aspectos econômicos-revisão. **Archives of Veterinary Science**, v. 10, n. 1, 2005.

MOURA, D. et al. Real time computer stress monitoring of piglets using vocalization analysis. **Computers and Electronics in Agriculture**, Elsevier, v. 64, n. 1, p. 11–18, 2008.

MUNN, Z. et al. Systematic review or scoping review? guidance for authors when choosing between a systematic or scoping review approach. **BMC Medical Research Methodology**, v. 18, 2018. Disponível em: https://doi.org/10.1186/s12874-018-0611-x.

NETO, A. P. et al. Perdas econômicas ocasionadas por lesões em carcaças de bovinos abatidos em matadouro-frigorífico do norte de mato grosso. **Pesquisa Veterinária Brasileira**, SciELO Brasil, v. 35, n. 4, p. 324–328, 2015.

PANDEYA, Y. R. et al. A monophonic cow sound annotation tool using a semi-automatic method on audio/video data. **Livestock Science**, Elsevier, v. 256, p. 104811, 2022.

PROAKIS, J. G. **Digital signal processing: principles algorithms and applications**. [S.l.]: Pearson Education India, 2001.

RAO, K. S.; MANJUNATH, K. **Speech recognition using articulatory and excitation source features**. [S.l.]: Springer, 2017.

ROCCHESSO, D. **Introduction to sound processing**. [S.l.]: Mondo estremo, 2003.

RÖTTGEN, V. et al. Automatic recording of individual oestrus vocalisation in group-housed dairy cattle: development of a cattle call monitor. **animal**, Cambridge University Press, v. 14, n. 1, p. 198–205, 2020.

ŞAŞMAZ, E.; TEK, F. B. Animal sound classification using a convolutional neural network. In: IEEE. **2018 3rd International Conference on Computer Science and Engineering (UBMK)**. [S.l.], 2018. p. 625–629.

SATTAR, F. A context-aware method-based cattle vocal classification for livestock monitoring in smart farm. **Chemistry Proceedings**, v. 10, n. 1, 2022. ISSN 2673-4583.

SCHWARTZKOPF-GENSWEIN, K. et al. Road transport of cattle, swine and poultry in north america and its impact on animal welfare, carcass and meat quality: A review. **Meat science**, Elsevier, v. 92, n. 3, p. 227–243, 2012.

SCHWARTZKOPF-GENSWEIN, K.; GRANDIN, T. et al. Cattle transport by road. **Livestock handling and transport. Wallingford, UK: Cabi**, p. 143–173, 2014.

SHORTEN, P. R. Acoustic sensors for detecting cow behaviour. **Smart Agricultural Technology**, v. 3, p. 100071, 2023. ISSN 2772-3755.

SILAPARASETTY, V. **Deep Learning Projects Using TensorFlow 2**. [S.l.]: Springer, 2020.

SIMON, H. A. **The Sciences of the Artificial**. 3. ed. Cambridge, MA: MIT Press, 1996. 240 p. ISBN 9780262193740.

SIQUEIRA, J. K. Reconhecimento de voz contínua com atributos mfcc, ssch e pncc, wavelet denoising e redes neurais. **Sistemas Maxwell-PUC-RIO, Certificação Digital no**, v. 912874, 2011.

ST, L.; WOLD, S. et al. Analysis of variance (anova). **Chemometrics and intelligent laboratory systems**, Elsevier, v. 6, n. 4, p. 259–272, 1989.

STRAPPINI, A. et al. Bruises in culled cows: when, where and how are they inflicted? **Animal**, Cambridge University Press, v. 7, n. 3, p. 485–491, 2013.

TIWARI, V. Mfcc and its applications in speaker recognition. **International journal on emerging technologies**, Citeseer, v. 1, n. 1, p. 19–22, 2010.

TORRE, M. P. de la et al. Acoustic analysis of cattle (bos taurus) mother–offspring contact calls from a source–filter theory perspective. **Applied Animal Behaviour Science**, v. 163, p. 58–68, 2015. ISSN 0168-1591.

VIDANA-VILA, E. et al. Automatic detection of cow vocalizations using convolutional neural networks. 2023.

YAJUVENDRA, S. et al. Effective and accurate discrimination of individual dairy cattle through acoustic sensing. **Applied Animal Behaviour Science**, Elsevier, v. 146, n. 1-4, p. 11–18, 2013.

YE, F.; YANG, J. A deep neural network model for speaker identification. **Applied Sciences**, MDPI, v. 11, n. 8, p. 3603, 2021.

YEON, S. C. et al. Acoustic features of vocalizations of korean native cows (bos taurus coreanea) in two different conditions. **Applied animal behaviour science**, Elsevier, v. 101, n. 1-2, p. 1–9, 2006.

ZHENG, F.; ZHANG, G.; SONG, Z. Comparison of different implementations of mfcc. **Journal of Computer science and Technology**, Springer, v. 16, p. 582–589, 2001.