Universidade Estadual Paulista
Instituto de Biociências, Letras e Ciências Exatas
Departamento de Ciências de Computação e Estatística

Daphne Lie Haranaka Pereira

Aprendizado de máquina para detecção e análise de emoções em sinais de voz

São José do Rio Preto - SP
2024

Daphne Lie Haranaka Pereira

Aprendizado de máquina para detecção e análise de emoções em sinais de voz

Monografia apresentada ao Curso de graduação em Ciência da Computação da UNESP para obtenção do título de Bacharel.

Orientador: Prof. Dr. Rodrigo Capobianco Guido

São José do Rio Preto - SP
2024

Pereira, Daphne Lie Haranaka
P436a Aprendizado de máquina para detecção e análise de emoções em sinais de voz / Daphne Lie Haranaka Pereira. -- São José do Rio Preto, 2024
70 p. : tabs.

Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (UNESP), Instituto de Biociências Letras e Ciências Exatas, São José do Rio Preto
Orientador: Rodrigo Capobianco Guido

1. Aprendizado do computador. 2. Processamento de sinais. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Dados fornecidos pelo autor(a).

Dedico este trabalho à minha família, que por toda minha vida me amou incondicionalmente.

E aos meus amigos, que todos os dias me amam autenticamente.

Agradecimentos

A todos os professores que compartilharam o seu conhecimento dentro e fora da sala ao longo desses quatro anos, em especial meu orientador e Prof. Rodrigo, por me ajudar e guiar para garantir que este trabalho fosse desenvolvido em um ambiente acolhedor. A UNESP e todos os funcionários do IBILCE, que sempre me fizeram sentir em casa dentro do câmpus. A minha família, principalmente minha mãe, pai, e irmã, e todas as amizades que foram formadas antes e durante o curso, todo esse apoio fazem o esforço valer a pena.

"'Hope' is the thing with feathers -

That perches in the soul -

And sings the tune without the words -

And never stops - at all -

And sweetest - in the Gale - is heard -

And sore must be the storm -

That could abash the little Bird

That kept so many warm -

I've heard it in the chillest land -

And on the strangest Sea -

Yet - never - in Extremity,

It asked a crumb - of me."

Emily Dickinson

Resumo

Atualmente, a tecnologia está cada vez mais integrada à vida cotidiana, tornando-se algo indispensável. Nesse contexto, surge uma crescente necessidade de tornar a interação entre humanos e máquinas mais natural. Para alcançar esse objetivo, é necessário que as máquinas possam identificar e interpretar eficientemente as diferentes emoções expressas pelos seres humanos. Esse campo de estudo é conhecido como SER (Reconhecimento de Emoção através da Fala), que pesquisa as diferentes abordagens aplicadas nas etapas de um modelo de aprendizagem de máquina, desde o pré-processamento, até a extração de características do áudio e a seleção dos atributos. A escolha desses métodos afeta diretamente a performance final e os resultados obtidos. Todo o processo realizado para a seleção dos algoritmos e seus respectivos resultados vão ser debatidos ao longo deste trabalho, sendo destacados resultados de pesquisas científicas anteriores que implementaram tais técnicas.

Palavras-chave: Processamento de sinais. Reconhecimento de emoção. Aprendizado de máquina. Extração de Atributos. Seleção de Atributos. Classificadores.

Abstract

Currently, technology is increasingly integrated into daily life, becoming indispensable. In this context, there is a growing need to make the human-machine interaction more natural. To achieve this goal, machines must be able to efficiently identify and interpret the different emotions expressed by humans. This field of study is known as SER (Speech Emotion Recognition), which researches the various approaches applied in each step of a machine learning model, from preprocessing to audio feature extraction and feature selection. The choosen methods have a direct effect on the final performance and the results obtained. The entire process for the selection of algorithms and their respective results will be discussed throughout this work, highlighting results from previous scientific studies that have implemented such techniques.

Keywords: Signal processing. Emotion Recognition. Machine Learning. Feature Extraction. Feature Selection. Classifiers.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

SER	*Speech Emotion Recognition*
DWT	*Discrete Wavelet Transform*
LPCC	*Linear Predictive Cepstrum Coefficients*
MFCC	*Mel Frequency Cepstral Coefficients*
RNA	Rede Neural Artificial
IEMOCAP	*Interactive Emotional Dyadic Motion Capture*
EMO-DB	*Berlin Database of Emotional Speech*
RAVDESS	*Ryerson Audio-Visual Database of Emotional Speech and Song*
CNN	*Convolutional Neural Network*
FSASL	*Feature Selection with Adaptive Structure Learning*
RFE	*Recursive Feature Elimination*
SFS	*Sequential Feature Selection*
FFT	*Fast Fourier Transform*
CWT	*Continuous wavelet transform*
RBF	*Radial basis function*
SVM	*Support Vector Machine*
WPT	*Wavelet Packet Transform*
ZCR	*Zero Crossing Rate*
MLP	*Multilayer Perceptron*
KNN	*K-Nearest Neighborhood*
RFE	*Recursive Feature Elimination*
RNN	*Recurrent Neural Network*

Sumário

Capítulo 1

Introdução

1.1 Considerações iniciais

A fala é especialmente uma fonte importante de conhecimento, permitindo a comunicação entre humanos. O sinal da fala revela informações cruciais não apenas sobre o estado emocional do locutor, mas também permite extrair outras características específicas, como gênero, idade, linguagem, dialeto e cultura (LANGARI; MARVI; ZAHEDI, 2020). Todas essas informações, embora tornem um possível modelo de máquina mais complexo, contribuem para um melhor desempenho de sistemas que realizam tomada de decisões.

Emoções têm uma influência direta nas decisões humanas, e acabam causando uma alteração no sinal da voz (TUNCER; DOGAN; ACHARYA, 2021). Embora humanos consigam facilmente reconhecer emoções pela fala, essa não é uma tarefa trivial para as máquinas. Ao identificar a emoção expressa, as máquinas podem tomar decisões mais assertivas e personalizadas, adaptando seu comportamento à emoção reconhecida (ANAGNOSTOPOULOS; ILIOU, 2010).

Atualmente, o procedimento mais utilizado para realizar o reconhecimento de emoções envolve a extração e seleção de características do sinal de fala, seguida da modelagem do classificador com as características definidas (LANGARI; MARVI; ZAHEDI, 2020). A performance do sistema pode variar bastante, dependendo da qualidade das características, do classificador implementado, e da base de dados usada para o treinamento do modelo (KERKENI et al.,

2019).

1.2 Objetivos

O principal objetivo desse trabalho é apresentar conceitos iniciais e a situação atual das pesquisas sobre reconhecimento de emoções, com foco no uso de aprendizado de máquina. Buscando identificar padrões nos processos aplicados nesse campo, destacando os algoritmos mais utilizados e aplicando-os de forma prática. E a partir disso analisar os resultados obtidos e apontar possíveis soluções para os problemas encontrados durante o desenvolvimento.

1.3 Justificativa e Motivação

Há uma crescente demanda por sistemas que permitam uma interação mais natural entre humanos e máquinas, melhorando a experiência do usuário com respostas adaptadas. Por ser um campo ainda pouco explorado, existem estratégias e técnicas a serem implementadas que podem aumentar o desempenho de sistemas de reconhecimento de emoções. Um dos principais desafios dessa área atualmente é a combinação de diferentes técnicas que equilibrem desempenho e custo computacional, além de enfrentar o problema da baixa generalização. Este trabalho visa expor soluções para esses desafios e levantar outras questões pertinentes a essa área de pesquisa. Aprofundando e abrangendo o conhecimento científico em uma área ainda pouco estudada dentro de processamento de sinais, e com grande potencial de impacto em setores como saúde, educação e segurança. Portanto, esse estudo contribui diretamente para o aumento e aprimoramento das metodologias aplicadas, visando melhorar a precisão na detecção de emoções em sinais de falas.

1.4 Metodologia

Inicialmente foi realizada uma pesquisa ao redor das bases de dados, pois elas contêm áudios já rotulados. Foi analisado o uso dessas bases em trabalhos acadêmicos, os idiomas em que estão disponíveis e seus volumes de dados. Em seguida, houve uma pesquisa abrangente sobre os trabalhos já publicados, identificando padrões de ferramentas e métodos utilizados. Na etapa de pré-processamento dos áudios, foram aplicados processos de padronização, normalização, além de filtros passa-baixa e passa-banda. Para a extração de características, utilizou-se as funções ZCR e RMSE, além dos coeficientes MFCCs. Na etapa da seleção de características foi aplicado o teste t de Welch, definindo as principais características que definem o sinal, otimizando a eficiência do classificador utilizado. Na etapa de classificação, foi empregado um modelo de CNN com múltiplas camadas. A eficácia do modelo foi medida por meio de métricas como acurácia, F1-score, precisão e *recall*.

1.5 Organização do trabalho

De forma a melhor organizar a apresentação e distribuição de informações ao longo deste trabalho, este foi separado em diferentes seções, estas são indicadas a seguir.

- No Capítulo 2 são apresentados alguns conceitos importantes de aprendizado de máquina, seguido pelo estado da arte, destacando a sua importância e impacto ao ser aplicado no mundo real. Após isso será abordado tópicos relacionados a definição e representação de emoções, as principais bases de dados utilizadas nesta área de pesquisa, e as etapas do modelo de aprendizado de máquina mais utilizados na classificação de emoções, sendo explicada e aprofundada cada uma delas.

- No Capítulo 3, são descritas as etapas do modelo de aprendizado de máquina aplicado, detalhando os algoritmos utilizados e o raciocínio por trás de sua aplicação.

- No Capítulo 4, os resultados obtidos são apresentados por meio de tabelas e gráficos.
- No Capítulo 5 foi feita uma conclusão breve, relatando os desafios enfrentados durante o desenvolvimento do trabalho e sugestões de melhorias a serem aplicados futuramente.

Capítulo 2

Revisão Bibliográfica

2.1 Introdução a *Deep Learning* e Rede Neural Convolucional

Atualmente técnicas de *deep learning* têm se mostrado efetivas na representação de características e reconhecimento de tarefas (KWON et al., 2021). A aprendizagem profunda faz parte de um conjunto abrangente de métodos baseados na aprendizagem de representações de dados (FERREIRA et al., 2017).

O aprendizado de máquina busca desenvolver métodos capazes de extrair conhecimentos a partir de amostras de dados. São utilizados algoritmos de aprendizado que empregam um conjunto de dados com classificações conhecidas para treinar o modelo. Isso permite que o modelo interprete e classifique novos dados da maneira mais apropriada. Existem três paradigmas que categorizam os algoritmos de aprendizado de máquina (FERREIRA et al., 2017):

- Aprendizado supervisionado: são utilizados dados rotulados para treinar o modelo.
- Aprendizado não supervisionado: os dados não são rotulados. São agrupados e classificados de acordo com os padrões encontrados pelo modelo.
- Aprendizado por reforço: aprendizado com base em recompensas, dependendo da performance do modelo.

2.1.1 *Deep Learning*

O aprendizado profundo foi desenvolvido com base no funcionamento do sistema nervoso, modelos são treinados para que sejam capazes de visualizar informações como humanos. Ao tentar atingir esse objetivo foram desenvolvidas Redes Neurais Artificiais (RNAs), que são máquinas projetadas para reconhecer padrões complexos de forma análoga ao cérebro. Uma rede neural é formada por um conjunto de neurônios que se comunicam por impulsos (FERREIRA et al., 2017).

Um neurônio é uma unidade que realiza o processamento de informações, e é uma parte essencial para o funcionamento de uma rede neural, os elementos mais comuns encontrados em uma arquitetura de RNA são apontados e descritos a seguir (FERREIRA et al., 2017):

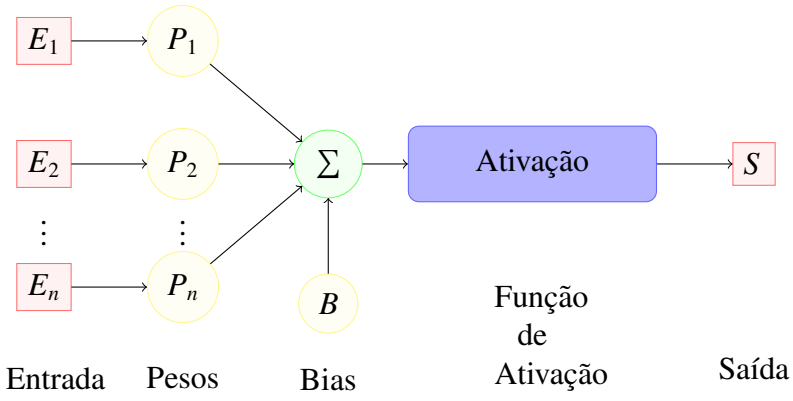

Figura 2.1 – Arquitetura básica de um neurônio.

Referência: (KONG et al., 2023)

- Entrada: camada que vai receber os dados que vão ser processados pelo neurônio.
- Sinapses: representam a conexão entre os neurônios.
- Bias: um termo de correção que permite a melhor adaptação da função de ativação, o seu valor influencia a intensidade da atividade do neurônio.
- Somador: realiza a soma das entradas, considerando os pesos das sinapses correspondentes.

- Função de ativação: função que limita o intervalo da amplitude de possíveis saídas do neurônio, garantindo que a resposta esteja dentro de um intervalo realista.

- Saída: valor produzido por um neurônio após o processamento.

O aprendizado profundo possibilita a construção de modelos mais complexos, compostos por múltiplas camadas de processamento, capazes de aprender representações de dados que possuem múltiplos níveis de abstração. Uma das arquiteturas de redes mais utilizadas de *Deep Learning* são as redes neurais convolucionais (FERREIRA et al., 2017).

2.1.2 Rede Neural Convolucional

Uma rede neural convolucional, ou CNN, é um tipo de algoritmo de aprendizado profundo amplamente utilizado para receber e classificar dados que possuem uma estrutura bidimensional, como imagens ou sinais de voz. Esse modelo possui múltiplas camadas que extraem as principais características dos dados de entrada, realizando convoluções sucessivas e redimensionamentos para determinar a classe apropriada dos dados analisados (FERREIRA et al., 2017). Os nós em uma CNN estão interconectados e associados a um peso e um limite. Se a saída de qualquer nó estiver acima do valor de limite, esse nó será ativado e enviará dados para a próxima camada. À medida que os dados percorrem as camadas, os volumes de entrada são analisados em uma escala cada vez maior, permitindo que, ao final, ocorra a classificação dos dados de entrada.

Uma CNN é composta por três camadas principais: camada convolucional, seguida de camada de *pooling*, e uma camada final chamada de camada totalmente conectada. Essa estrutura é ilustrada pela figura 2.2.

Camada Convolucional

Onde ocorre a maior parte dos cálculos, o principal objetivo dessa camada é extrair as principais características dos dados de entrada.

Nessa camada é realizada a operação de convolução, onde é feito o tratamento dos dados

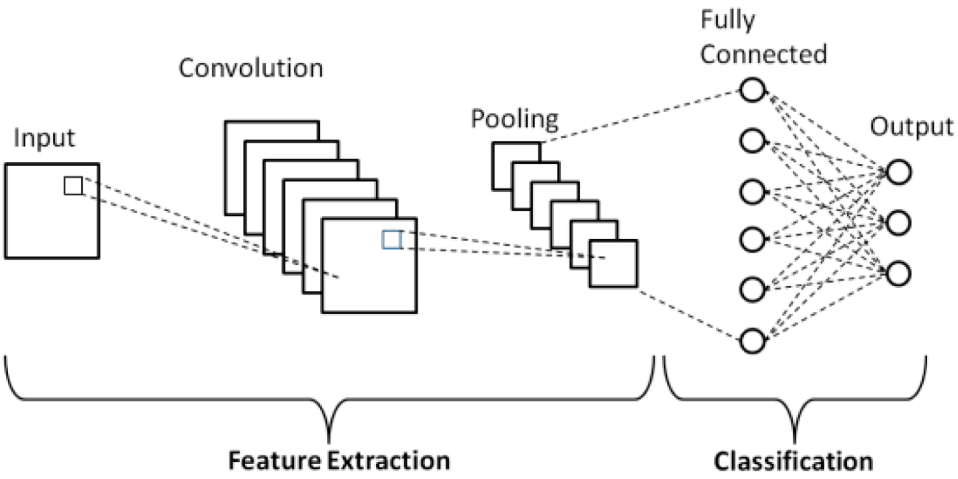

Figura 2.2 – Figura ilustrando as camadas de uma CNN

Fonte: (PHUNG; RHEE, 2019)

de entrada pelo *kernel*, que irá aplicar filtros nos dados para mapear e destacar as suas feições. Após isso é feito a somatória do produto ponto a ponto entre os valores da matriz *kernel* e cada posição da matriz dos dados de entrada (FERREIRA et al., 2017).

Existem três parâmetros que afetam o tamanho do volume resultante da camada convolucional, que precisam ser definidos antes do início do treinamento da rede neural:

- *Depth*: representa o número de filtros usados na operação de convolução (HUANG et al., 2014). O número de filtros tem um impacto na complexidade computacional, quanto maior o seu número, maior vai ser o número de características extraídas (FERREIRA et al., 2017).

- *Stride*: referente ao tamanho do filtro, ou seja, a distância que o *kernel* se move sobre a matriz de *input* (HUANG et al., 2014)(IBM, 2024).

- *Zero-padding*: processo de completar a matriz de entrada com zeros, geralmente é utilizada quando o filtro não se encaixa na imagem de *input* (HUANG et al., 2014).

Camada de Pooling

Responsável por reduzir a dimensionalidade dos dados, mantendo somente as informações mais importantes. Assim como na camada convolucional, um filtro é aplicado para todas as entradas, a diferença é que esses filtros não tem pesos. Existem diferentes tipos de operação de *pooling*, como máximo, média, soma, entre outros (FERREIRA et al., 2017).

No *pooling* máximo, à medida que o filtro passa pelo dado de entrada, ele vai selecionar o ponto de maior valor e o envia para a matriz de produção. Já no *pooling* médio, o filtro calcula o valor médio dentro do campo receptivo antes de enviar para a matriz de produção (IBM, 2024).

Camada *Fully Connected*

Diferentemente das outras camadas, cada nó na sua camada de saída é conectado diretamente a um nó da camada anterior, normalmente camadas totalmente conectadas utilizam uma função de ativação *softmax* para a classificação das entradas, produzindo um valor entre zero e um, representando uma probabilidade (IBM, 2024).

O principal objetivo dessa camada é utilizar os recursos extraídos dos dados de entrada em outras camadas para classificar os dados nas classes corretas, com base no conjunto de dados de treinamento (FERREIRA et al., 2017).

2.2 Estado da Arte

2.2.1 Definição, Representação e Aplicação de um Sistema de Classificação de Emoção

Um modelo de máquina capaz de classificar emoções tem aplicações diversas. Antes de abordar essas aplicações, é preciso discorrer sobre a definição de emoção. Uma das definições mais reconhecidas entre pesquisadores é a proposta por Scherer (ANAGNOSTOPOULOS; ILIOU, 2010), que afirma: “Emoções são episódios de mudanças coordenadas em vários componentes (incluindo a ativação neurofisiológica, expressão motora e sentimentos subjetivos) [...] em resposta a eventos externos ou internos de grande importância para o organismo”. Nesse contexto, pesquisadores desenvolvem diferentes métodos para representar as principais emoções humanas.

A aplicação de sistemas capazes de classificar emoções abrange uma variedade de áreas.

Em serviços de *call center*, essa tecnologia pode melhorar o atendimento ao identificar a emoção do cliente. Na área da saúde, é possível ser utilizada para analisar o estresse e depressão de um paciente. Em sistemas de carros inteligentes, aprimora a performance geral dos veículos (KWON et al., 2021). Na educação, é comprovado que o conhecimento do estado emocional do aluno pode melhorar a qualidade do ensino (KERKENI et al., 2019). Outras áreas que também se beneficiam com o uso de um classificador de emoções são: ciência forense, vigilância de áudio, entretenimento, bancos, e outros (KERKENI et al., 2019).

A aplicação dessa pesquisa em inúmeras e diferentes áreas justifica a importância do seu desenvolvimento e estudo, dado o seu impacto direto e significativo em diversos indivíduos. É importante destacar também que novos métodos são descobertos e aplicados constantemente, de forma a aumentar a eficiência de modelos e a sua performance, uma vez que ainda não existe um consenso sobre a melhor abordagem a ser utilizada.

O número de emoções definidas, bem como a forma que serão representadas e analisadas, influenciam a performance do modelo construído. A seguir, são abordadas duas formas de representação das emoções: em categorias, e valores contínuos.

Emoções Categorizadas

Essa representação de emoções baseia-se na teoria das emoções discretas, inspirada pelas ideias de Darwin. Ela sugere a existência de um conjunto de emoções, conhecidas como emoções básicas. Um exemplo de um conjunto de emoções básicas são aquelas da figura 2.3. Essas emoções são caracterizadas por estarem presentes em grande parte dos eventos importantes da vida e são sentidas com maior frequência. Pesquisas como a de Scherer (ANAGNOSTOPOULOS; ILIOU, 2010), agrupam nessa categoria emoções como raiva, medo, felicidade, tristeza, ansiedade, tédio e emoções neutras.

Um método comumente usado para diferenciar emoções é através da distinção entre polos positivos e negativos. Em estudos como o de Ekman (SINGH et al., 2021), são definidas seis emoções principais: raiva, nojo, medo, felicidade, tristeza, e surpresa, com felicidade e surpresa no polo positivo, e as demais no negativo. Outro estudo, liderado por Plutchik (SINGH et al., 2021), indica como emoções positivas: felicidade, surpresa, antecipação e confiança, enquanto raiva, nojo, tristeza e medo são consideradas emoções negativas. Já Latinjak (TUNCER; DOGAN; ACHARYA, 2021), combina emoções positivas e negativas, dividindo-as em

quatro categorias: fadiga, surpresa, calma e alerta.

É importante destacar que as classes de emoções utilizadas neste trabalho se limitam às emoções expressas pela fala, pois estudos indicam que o número de emoções dependem do meio em que são expressas. Como na pesquisa vista no trabalho de Singh et al. (SINGH et al., 2021), onde foram identificadas 24 emoções expressas por vocalização breve, 28 emoções com base em expressões faciais, e 12 emoções pela prosódia da fala.

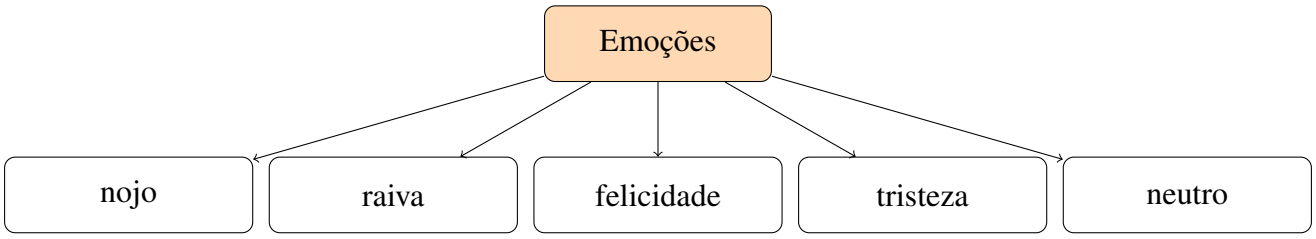

Figura 2.3 – Exemplos de Categorias de Emoções

Emoções como valores contínuos

Outros estudos utilizam a ideia de que as emoções podem ser representadas em um espaço de quatro dimensões, permitindo uma análise mais detalhadas das similaridades e diferenças entre elas. Os quatro eixos mais utilizados são os seguintes (ALBUQUERQUE, 2020):

- *Arousal*: refere-se à excitação causada pela intensidade das reações fisiológicas, variando de muito calmo para muito animado.
- Valência: representa a sensação agradável ou desagradável provocada pela experiência emocional.
- Potência: refere-se à sensação de estar no controle ou ser dominado, variando entre dominante e submisso.
- Surpresa: emoção sentida ao experienciar novidade e imprevisibilidade.

Dentre esses quatro eixos, a comunidade de pesquisa geralmente foca em um espaço bidimensional, composto apenas pelos eixos de valência e *arousal* (ANAGNOSTOPOULOS; ILIOU, 2010). Essas duas dimensões são então utilizadas como coordenadas para caracterizar diferentes emoções, conforme ilustrado na figura 2.4.

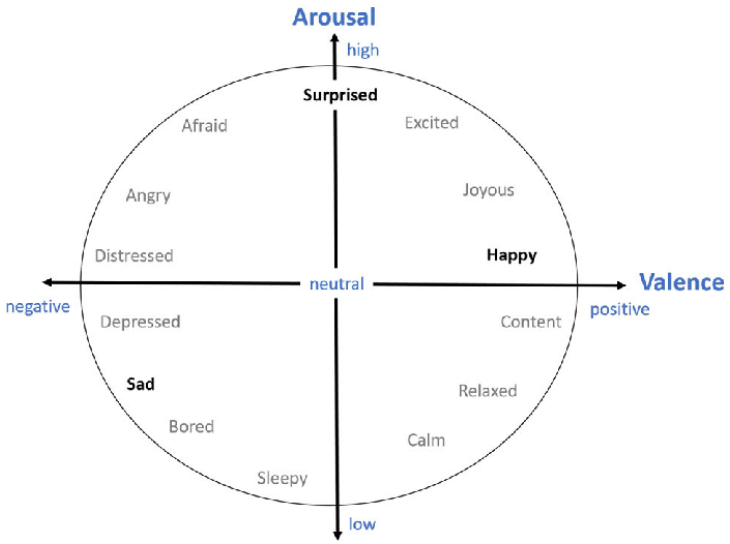

Figura 2.4 – Emoções representadas em 2 eixos

Fonte:
(TSIOURTI et al., 2019)

Informação Linguística e Paralinguística

Para ajudar na identificação de emoções, é possível recorrer tanto às informações linguísticas quanto às paralinguística (ANAGNOSTOPOULOS; ILIOU, 2010), ambas relacionadas à fala.

- Informações linguísticas: Envolve dados que identificam padrões qualitativos articulados pelo orador. Relacionando palavras a emoções, por exemplo, a palavra "feliz" pode ser associada ao sentimento de felicidade.

- Informações paralinguística: São informações extraídas utilizando métodos de processamento, medidas por características quantitativas que descrevem como palavras ou frases são pronunciadas. Exemplos incluem variações no tom e na intensidade da fala, parâmetros que não estão relacionados à identidade da palavra.

2.2.2 Base de Dados

Ao desenvolver um modelo que classifica qualquer tipo de dado, é necessário realizar o treinamento desse modelo, de modo a alcançar uma boa acurácia. O processo de treinamento

envolve fornecer tanto um algoritmo de *machine learning* quanto os dados de treinamento, que podem ser obtidos com uma base de dados. Essa base deve conter não apenas os dados de treinamento, mas também as respostas correspondentes, permitindo que o modelo identifique padrões entre os dados de treinamento e a resposta. Assim, ao fornecer novos tipos de dados, o modelo tem uma maior chance de retornar a resposta esperada.

Na área de classificação envolvendo áudio, aspectos como características de fala do locutor, tamanho, número de palavras na sentença, e língua, interferem diretamente na construção de um conjunto de dados (LANGARI; MARVI; ZAHEDI, 2020).

Um dos tipos de base de dados utilizadas são as chamadas de simuladas, onde as sentenças são faladas por atores profissionais ou estudantes de teatro, que conseguem interpretar diferentes tipos de emoções. Outro tipo de base de dados é a induzida, cujo conteúdo é coletado ao provocar certas reações em pessoas de forma artificial, utilizando vídeos, imagens, ou jogos. Um terceiro tipo envolve utilizar emoções em sua forma natural, utilizando gravações de telefone, entrevistas, e conversas. Esse último tipo de coleta de dados não possui a mesma qualidade que os outros, pois as gravações frequentemente possuem ruído de fundo, o que dificulta a obtenção apenas das vozes principais (KERKENI et al., 2019).

Pesquisas apontam que modelos possuem mais facilidade classificando bases de dados que foram atuadas (ANAGNOSTOPOULOS; ILIOU, 2010), pois as falas tendem a ser mais exageradas, facilitando o reconhecimento.

Para um melhor desempenho do modelo, é preciso escolher uma ou mais bases de dados de boa qualidade. Existem vários métodos utilizados para a construção dessas bases de dados, e várias bases estão disponíveis na internet. A seguir, são mencionadas as principais bases de dados usadas nesse ramo de pesquisa.

- IEMOCAP: uma base de dados multimodal que contém diálogos roteirizados e improvisados, gravados com 10 atores em 5 sessões separadas, somando ao total 12 horas de dados incluindo vídeos, transcrições de texto e capturas de movimentos do rosto. As emoções contidas dentro dessa base são as seguintes: felicidade, tristeza, neutro, raiva, surpresa, animação, frustração, nojo, medo, e outros (YILDIRIM; KAYA; KILIÇ, 2021)(BANDELA; KUMAR, 2020).

- EMO-DB: base de dados alemã, e uma das mais utilizadas em modelos de classificação

de emoções. Contém sentenças interpretadas por 10 atores profissionais, sendo 5 mulheres e 5 homens, na idade entre 25 a 35 anos. Interpretando emoções como felicidade, tristeza, raiva, tédio, medo, neutro e nojo. Resultando em 535 clipes diferentes de áudio (YILDIRIM; KAYA; KILIÇ, 2021)(BANDELA; KUMAR, 2020).

- CASIA: base de dados chinesa gravada pelo Instituto de Automação, na Academia Chinesa de Ciências. Gravada com 4 pessoas (2 homens e 2 mulheres), interpretando 300 sentenças em 5 emoções: surpresa, felicidade, tristeza, raiva, medo, além de neutro (LIU et al., 2018).

- Base de dados espanhol: contém 6041 declarações, o que a torna uma das maiores bases com esse tipo de conteúdo. Seus áudios foram gravados em 2 sessões com 2 atores (1 mulher e 1 homem), interpretando 6 emoções básicas: raiva, tédio, nojo, medo, felicidade e tristeza. E de formas variadas, diferenciando a altura e a velocidade com que as sentenças são faladas (KERKENI et al., 2019).

- RAVDESS: contém 24 vozes de atores profissionais (12 vozes femininas e 12 masculinas), falando 2 sentenças com o sotaque norte americano. As emoções contidas nessa base de dados são: felicidade, tristeza, raiva, com medo, calmo, nojo e surpresa. Todas essas são expressas em 2 níveis de intensidade emocional (forte e fraco) (HUANG et al., 2014).

- CREMA: *Crowd-sourced Emotional Multimodal Actors* (CREMA), é uma base de dados em inglês composta por 7442 clipes de áudio de 91 atores (48 homens e 43 mulheres), entre as idades de 20 e 74. Os atores falaram 12 sentenças expressando seis emoções: raiva, nojo, medo, felicidade, neutro e tristeza (MASHHADI; OSEI-BONSU, 2023).

- TESS: *Toronto Emotional Speech Set* (TESS) contém áudios em inglês de 2 atrizes, com idades de 26 e 64 anos. Essas atrizes falaram 200 palavras em 7 emoções: raiva, nojo, medo, felicidade, surpresa, tristeza e neutro. Essas palavras foram escolhidas na base de dados *Affective Norms for English Words* (ANEW) (MASHHADI; OSEI-BONSU, 2023).

- SAVEE: Inclui áudios em inglês interpretados por 4 atores entre 27 e 31 anos. Cada um

dos atores falaram 15 sentenças em 7 emoções diferentes: raiva, nojo, medo, felicidade, tristeza, surpresa e neutro (MASHHADI; OSEI-BONSU, 2023).

2.2.3 Modelo Básico de *Machine Learning* para Detecção de Emoção

Existe uma variedade de modelos utilizados na área de *machine learning*, cada uma com características distintas e complexidade computacional diferente. No entanto, na maioria dos casos, os modelos que realizam a classificação de sinais da fala seguem as etapas delineadas no diagrama apresentado na figura 2.5.

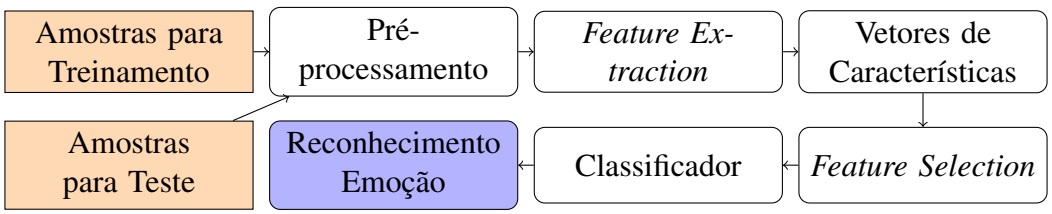

Figura 2.5 – Diagrama representando um modelo básico de aprendizado de máquina

A seguir serão comentados brevemente a função, os objetivos, e a importância de cada uma dessas etapas.

2.2.4 Pré-processamento

Nesse momento, é possível aplicar diferentes técnicas para tornar o sinal mais adequado para as próximas etapas. Como aplicar a normalização, para eliminar mudanças na amplitude que ocorreram durante a gravação, além de filtrar ruídos de fundo e segmentar os sinais de fala (LANGARI; MARVI; ZAHEDI, 2020).

Existem duas abordagens diferentes que podem ser seguidas para realizar o pré-processamento dos dados. O método tradicional inclui realizar a amostragem, *pre-emphasis*, *windowing* e *framing*, além de *denoising* (MADANIAN et al., 2023).

- Amostragem e Quantização: a amostragem (ou *sampling*) é o processo de conversão de um sinal analógico em um sinal digital. A taxa de amostragem refere-se ao número de amostras extraídas do sinal original por segundo, criando um sinal digital discreto. A quantização é realizada em conjunto com a amostragem para digitalizar o sinal, quantizando os valores de amostragem para o nível mais próximo.

- *Pre-emphasis*: é aplicada para maximizar os efeitos das frequências mais altas e mais baixas. As frequências mais altas serão aumentadas e as frequências mais baixas serão diminuídas, pois frequências mais altas tendem a conter informações de maior importância.

- *Framing*: o processo de *framing* divide os sinais em diferentes partes de tamanhos fixos, chamados de *frames*. Dessa forma os *frames* são analisados separadamente como vetores, sem que haja uma perda de informação.

- *Windowing*: após o *framing*, uma função janela pode ser aplicada em cada *frame* para eliminar possíveis descontinuidades no seu início e fim, além de ajustar infiltrações espectrais no sinal e intersecções devido a sobreposições.

- *Denoising*: remoção de ruído é importante, pois a presença de ruídos no sinal pode afetar negativamente os resultados. Já a remoção do silêncio permite que o modelo se concentre em analisar partes específicas do sinal.

Outro método utilizado inclui implementar as técnicas anteriormente mencionadas, e também transformar o sinal de áudio em espectrogramas e *wavelets*. Um espectrograma mostra a intensidade e a força de um sinal ao longo do tempo em diferentes frequências e tipos específicos de onda. Os *wavelets* conseguem prover uma análise do tempo, energia e frequências em um sinal de fala (MADANIAN et al., 2023).

- Transformada Fourier: a transformada de Fourier permite a análise do sinal no domínio da frequência, já que, anteriormente, os sinais estavam no domínio do tempo.

- Transformada *Wavelet*: a transformada *wavelet* soluciona um dos problemas encontrados ao usar apenas a transformada de Fourier, que não consegue capturar componentes de tempo e frequência simultaneamente. A transformada *wavelet* decompõe o sinal em componentes, dividindo-os com base na frequência (altas e baixas). Isso permite uma

boa resolução no tempo para componentes de alta frequências, e uma melhor resolução de frequência para componentes de baixa frequências (MADANIAN et al., 2023).

Algoritmos Pré-processamento

As principais operações feitas durante a etapa de pré-processamento são mostradas a seguir (SHINDE; PATIL, 2023):

- *Pre-emphasis filtering*: pode ser utilizado um filtro *high pass* de primeira ordem, de modo a aumentar o sinal das frequências mais altas (BANDELA; KUMAR, 2020). A equação do filtro é mostrada em 2.1.

$$H[n] = S[n] - \alpha * S[n-1] \tag{2.1}$$

Fonte: (SHINDE; PATIL, 2023)

Sendo $\alpha = 0.97$.

- *Framing*: o tamanho de um *frame* pode variar, um valor indicado por Madanian et al. (MADANIAN et al., 2023) é um *frame* de tamanho entre 10 a 20 ms. Já na pesquisa de Shinde (SHINDE; PATIL, 2023), o sinal é amostrado e dividido em *frames* de 25 ms com 10 ms de sobreposição.

- *Windowing*: um valor bastante utilizado é a janela de Hamming, representado pela equação 2.2 (LANGARI; MARVI; ZAHEDI, 2020).

$$v[n] = 0.54 - 0.46 * \cos\left(\frac{2\pi n}{N}\right), \text{onde } 0 \leq n \leq N \tag{2.2}$$

Fonte: (SHINDE; PATIL, 2023)

O sinal enviado para a fase de extração de atributo é resultante da multiplicação da amostra $H[n]$ por $v[n]$.

Por fim, é recomendado remover o sinal *trend*, pois pode causar alguns erros na análise de correlação no domínio do tempo, na análise espectral no domínio da frequência, e na autenticação no espectro de baixas frequências. Esse processo de remoção é baseado no método de detecção da taxa de *zero-crossing*. E é feito em dois passos, o primeiro passo é o cálculo do

sinal *trend*, determinando um *threshold*. Isso pode ser feito pelas seguintes fórmulas (MADANIAN et al., 2023):

$$T(t) = \sum IMF \tag{2.3}$$

Fonte: (MADANIAN et al., 2023)

Com $T(t)$ representando a parte *trend* do sinal. E IMF a Função do Modelo Intrínseco. Uma função de modelo intrínseco é aquela onde o número de extremos no sinal (máximos e mínimos) são iguais, ou diferem no máximo por um do número do *zero crossing*. Outra propriedade dessas funções é que a média dos envelopes definidos pelos máximo e mínimos locais sempre são zeros (KERKENI et al., 2019).
O segundo passo é a subtração do sinal *trend* do sinal original:

$$S(t) = X(t) - T(t) \tag{2.4}$$

Fonte: (MADANIAN et al., 2023)

$X(t)$ é o sinal original e $S(t)$ é o sinal derivado após a remoção do *trend*.

2.2.5 *Feature Extraction*

É importante eliminar diferenças individuais presentes no sinal da voz para que o sistema analise somente os dados com as informações essenciais, aumentando a acurácia e criando um sistema capaz de analisar dados universais (LIU et al., 2018).

Podendo ser chamada de extração de recursos, *feature extraction* é uma etapa fundamental, pois elimina ambiguidades nas informações extraídas, melhorando a performance do modelo (LANGARI; MARVI; ZAHEDI, 2020). Nessa etapa o sinal de entrada é reduzido para um vetor de tamanho fixo, de forma a facilitar análises posteriores. O objetivo é extrair características específicas e essenciais do sinal da voz, utilizando várias técnicas de processamento de sinal (BANDELA; KUMAR, 2020). A extração de recursos também tem como foco melhorar o

nível de similaridade e não similaridade entre diferentes classes, além de comprimir e reduzir a quantidade de dados, diminuindo o poder de processamento necessário.

É possível dividir os recursos presentes na fala em quatro categorias, duas delas relacionadas à dependência ou independência do locutor, e são explicadas a seguir:

Características dependentes do locutor

Estas características são influenciadas por fatores pessoais do locutor, como o ritmo da fala, onde mesmo comparando sinais de voz de duas pessoas em estados emocionais similares, ainda é possível perceber uma diferença na distribuição dos dados (LIU et al., 2018).

Características independentes do locutor

Para minimizar as diferenças de características individuais dentro da fala entre pessoas, é introduzido uma taxa de mudanças que reflete essas diferenças durante a fala. Para obter características da fala que são independentemente do locutor é aplicado a derivada (LIU et al., 2018).

Outra forma de categorizar esses recursos é basear-se nas diferenças entre as características do sistema do trato vocal e as características prosódicas.

Características do sistema do trato vocal

Normalmente são características que representam a distribuição do parâmetro energia ao longo de uma faixa de frequência no sinal da fala. São analisados coeficientes cepstrais, definidos como a transformada (inversa) de Fourier do logaritmo do espectro do sinal, representando as características do trato vocal e as características da fonte da voz (AGHAJAN; AUGUSTO; DELGADO, 2009), como Mel Frequency Cepstral Coefficients (MFCC), *Linear Predictive Cepstral Coefficients* (LPCC) e Formante (LANGARI; MARVI; ZAHEDI, 2020).

Características prosódicas

São características extraídas dos dados prosódicos, como tom, energia e duração (LANGARI; MARVI; ZAHEDI, 2020).

Alguns estudos apontam que o tom e a energia do sinal estão fortemente relacionados ao estado emocional do indivíduo (ANAGNOSTOPOULOS; ILIOU, 2010). E por isso são destacados como parâmetros fundamentais na extração de dados para diferenciar emoções. A

expressão de alegria ou raiva, por exemplo, geralmente tem valores de energia e tom mais altos comparados a emoções como tristeza.

O tom é considerado uma frequência fundamental, e estudos como o de Anagnostopoulos (ANAGNOSTOPOULOS; ILIOU, 2010) mostram que os dados relacionados ao tom transmitem informações importantes sobre o estado emocional. No entanto, o tom é altamente relacionado ao gênero do locutor, e isso deve ser levado em consideração para evitar classificações errôneas.

A energia, também entendida como volume ou intensidade da fala, contém informações essenciais. Com ela é possível diferenciar grupos de emoções, embora sua medida isolada não seja suficiente para diferenciar emoções básicas (ANAGNOSTOPOULOS; ILIOU, 2010).

Principais Algoritmos *Feature Extraction*

Mel Frequency Cepstral Coefficients (MFCC)

MFCC é um dos atributos espectrais, esses atributos são informações da voz derivadas do espectro do sinal de fala, e utilizados para extrair dados do trato vocal, modelando o padrão da intonação e o tom da frequência do locutor (LANGARI; MARVI; ZAHEDI, 2020).

Já o Mel é uma escala utilizada nas medidas que envolvem frequência e tom, simulando a frequência de resposta do ouvido humano (SHINDE; PATIL, 2023). Para realizar a conversão de medida entre frequência vs tom para a escala Mel é utilizada a fórmula:

$$\text{Mel value} = (2595) * \log_{10}\left[1 + \left(\frac{f}{700}\right)\right] \tag{2.5}$$

Fonte: (HEMA; MARQUEZ, 2023)

Onde f é a frequência.

MFCCs são baseadas no espectro de curto-prazo e características dominantes, tornando o seu valor mais próximo ao tom emocional do orador (HEMA; MARQUEZ, 2023). Os componentes de frequências mais baixas de um sinal contêm informações fonéticas mais claras. Ainda é possível utilizar uma escala de frequência Mel não linear do banco de filtros na etapa de *pre-emphasis*, para diminuir os componentes de frequências altas (LANGARI; MARVI; ZAHEDI, 2020).

Os MFCCs de baixa ordem contém informações em relação ao envelope espectral, que so-

fre mudanças lentamente, enquanto que os MFCCs de ordens mais altas explicam as rápidas variações do envelope. Os filtros são espaçados linearmente em frequências mais baixas e logaritmicamente em frequências mais altas, essa organização permite a captura de características foneticamente importantes do áudio (THIRUVENGATANADHAN, 2018).
Os principais objetivos ao utilizar o MFCC são tornar as características extraídas independentes, e capturar a dinâmica do contexto (AGGARWAL; KUMAR, 2022).

Para obter o MFCC, é necessário realizar primeiramente o processo de *pre-emphasizing* e segmentação em *frames*, podendo também ser aplicado a janela Hamming em cada *frame*. Após isso, um espectro de amplitude é computado para cada um dos *frames* utilizando FFT. Para obter o espectro-mel, os coeficientes de Fourier passam pelo banco de filtro Mel, convertendo em um conjunto de saídas do banco de filtros na escala mel. Por fim, é aplicada a transformada discreta de cosseno no coeficiente logarítmico da frequência Mel para a extração dos MFFCs. Os procedimentos utilizados no trabalho de Langari et al. (LANGARI; MARVI; ZAHEDI, 2020) para realizar essa extração e as derivadas são mostrados através das equações (2.6), (2.7) e (2.8).

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nk}{N}}; 0 \leq k \leq N-1 \tag{2.6}$$

$$s(m) = \sum_{k=0}^{N-1} |X(k)|^2 H_m(k); 0 \leq m \leq M-1 \tag{2.7}$$

$$C(n) = \sum_{m=0}^{M-1} \log(s(m)) \cos\left(\frac{n\pi(m-0.5)}{M}\right); 0 \leq n \leq C-1 \tag{2.8}$$

Fonte: (LANGARI; MARVI; ZAHEDI, 2020)

Sendo:

$X(k)$: Coeficientes discretos de Fourier

$s(m)$: Escala no espectro-mel de Fourier

$C(n)$: MFCCs

N: número de amostras na computação da Transformada Discreta de Fourier

M: Número de bancos de filtros de Mel

C: Número de MFCCs

Além disso, $x(n)$ está relacionado ao sinal principal e $H_m(k)$ está relacionado ao banco de filtro

Mel.

Discrete Wavelet Transform (DWT)

Wavelets são oscilações não periódicas que existem em um pequeno intervalo de tempo. Já WTs são conversões matemáticas baseados em uma *wavelet* específica (GUIDO, 2022). Transformadas *wavelets* são divididas geralmente em duas categorias: contínua e discreta (MADANIAN et al., 2023).

- DWT: Utiliza amostras discretas de *wavelets*, se destaca por sua capacidade de capturar informações no domínio do tempo e da frequência.

- CWT: Permite uma representação completa do sinal, pois operam com todos os tamanhos de janelas e posições, de modo a cobrir porções específicas do sinal para análise, diferentemente da DWT, que usa apenas um subconjunto de possibilidades. A principal diferença entre ela e a DWT é que a primeira só utiliza a *wavelet* mãe, e a DWT utiliza a *wavelet* mãe e a *wavelet* pai.

É importante ressaltar que tanto a DWT quanto a CWT são utilizadas para analisar sinais contínuos.

Apesar da CWT conseguir uma representação que oferece boas resoluções no domínio de tempo e frequência, é necessário um custo computacional maior, e por isso a DWT é mais viável em mais aplicações.

Ao desenvolver uma análise sobre *wavelets*, é necessário discutir sobre a DTWT, que é uma variação da DWT, e muito importante e utilizada para analisar sinais discretos.

DTWT

Na DTWT são utilizados dois filtros: $h[\cdot]$, exibindo uma resposta em torno da frequência passa-baixa, e $g[\cdot]$, com uma resposta de frequência espelhada passa-alta (GUIDO, 2022). Diferentemente do CWT e DWT, a DTWT se baseia no processo de *filtering* e *downsampling*, aplicados utilizando o algoritmo de Mallat.

Antes de realizar esses procedimentos, é importante definir o nível de decomposição e o par de filtro ideais, além da uma análise ao redor da família do filtro.

Seguindo o princípio da incerteza de Heisenberg (GUIDO, 2017) informações de tempo e

frequências são opostas, ou seja, quanto melhor a primeira, pior a qualidade obtida da segunda, e vice-versa.

Em casos onde é preciso obter uma boa resolução espectral, é recomendado a decomposição até o último nível, e um filtro com o maior tamanho possível, como os filtros Daubechies.

De maneira oposta, quando se está a procura de uma melhor resolução temporal, é indicado realizar a decomposição em níveis mais baixos, e utilizar um filtro pequeno, tal como filtros Haar.

Por fim caso seja necessário obter informações úteis no domínio da frequência e tempo, é recomendado uma decomposição até um nível médio, e um tamanho de filtro intermediário (GUIDO, 2017).

Na pesquisa desenvolvida por Shinde (SHINDE; PATIL, 2023), foi utilizado o *wavelet* Daubechies4 (db4) como a função base. As variações em emoções podem ser identificadas por conta da decomposição em cada nível em frequência passa alta e passa baixa. A resolução DWT multi sub banda pode ser representada através da fórmula a seguir:

$$T_{m,n} = \int_{-\infty}^{\infty} x(t)\psi_{m,n}(t)dt \tag{2.9}$$

Fonte: (SHINDE; PATIL, 2023)

Onde ψ é a *wavelet* mãe, que nesse caso é a db4.

2.2.6 *Feature Selection*

O objetivo dessa etapa é remover informações redundantes que podem estar dentro do sinal, reduzindo o vetor de características e a complexidade do modelo, melhorando a performance e diminuindo o tempo de processamento de algoritmos de aprendizagem. As características extraídas na primeira etapa ainda contém informações desnecessárias e irrelevantes, que não contribuem positivamente na predição final, e por isso é necessário a seleção de características. Os dados resultantes após essa etapa vão ser os essenciais para ser possível realizar a classificação dos dados em diferentes classes (YILDIRIM; KAYA; KILIÇ, 2021)(KERKENI et al., 2019).

É possível separar as técnicas de *feature selection* em diferentes categorias, se baseando na rotulagem de dados como supervisionado, não supervisionado e semi supervisionado (BANDELA; KUMAR, 2020). Adicionalmente, o processo de *feature selection* possui três dimensões, que caracterizam o funcionamento dos métodos utilizados nessa etapa (SILVA; BISPO; TEIXEIRA, 2021). As dimensões são as seguintes:

- Direção da busca: é diretamente relacionada com o ponto de início definido. Por exemplo na *foward selection*, que utiliza uma estratégia de busca linear que começa com um subconjunto vazio, e seleciona características individualmente para serem adicionados. Já a eliminação *backward* começa com todas as características no subconjunto e as remove separadamente. A seleção bidirecional realiza duas buscas paralelamente por iteração, uma que adiciona as características e outra que remove, sendo útil quando não se sabe o número ótimo de características.

- Estratégia da busca: pode ter uma ação global ou local. Como exemplo a busca exaustiva, que analisa todas as possibilidades de combinação de características e seleciona o melhor conjunto.

- Critério de parada: um exemplo de um método de critério de parada é parar de remover e adicionar atributos quando nenhuma das alternativas melhoram a acurácia da classificação, ou quando o subconjunto selecionado já consegue separar em classes certas.

Ademais as técnicas utilizadas nesse processo também podem ser classificadas dentre os seguintes tipos: baseada em filtro (*filter*), baseada em embrulho (*wrapper*), métodos de incorporação (*embedded*), híbrido (*hybrid*) e *ensemble* (BANDELA; KUMAR, 2020). O modelo de funcionamento da seleção baseada em filtro, embrulho, por incorporação e pelo método híbrido são representados nos diagramas 2.6, 2.7, 2.8 e 2.9 respectivamente, destacados em vermelho são as etapas que os caracterizam.

Seleção Baseada em Filtro

São técnicas que utilizam alguma métrica estatística para encontrar atributos relevantes em um conjunto de dados, atribuindo a eles uma pontuação. Com base nessa pontuação, os atributos

são ranqueados. A partir desse ranqueamento, os atributos podem ser mantidos ou removidos do vetor de atributos original (BANDELA; KUMAR, 2020). Normalmente os atributos escolhidos são aqueles que possuem uma posição mais alta no ranqueamento.
Algumas das métricas utilizadas para realizar o cálculo dessa pontuação são o teste qui-quadrado e a correlação de Pearson.

- Qui-quadrado (2.10): calcula o quão distante os valores esperados estão do resultado real (LI PETER LU, 2023)

$$X^2 = \sum \frac{(O - E)^2}{E} \tag{2.10}$$

Fonte: (FRANKE; HO; CHRISTIE, 2012)

Sendo O a frequência observada e E a frequência esperada.

- Correlação de Pearson (2.11): calcula a intensidade da correlação entre duas variáveis (LI PETER LU, 2023)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2][\sum_{i=1}^{n}(y_i - \bar{y})^2]}} \tag{2.11}$$

Fonte: (LI PETER LU, 2023)

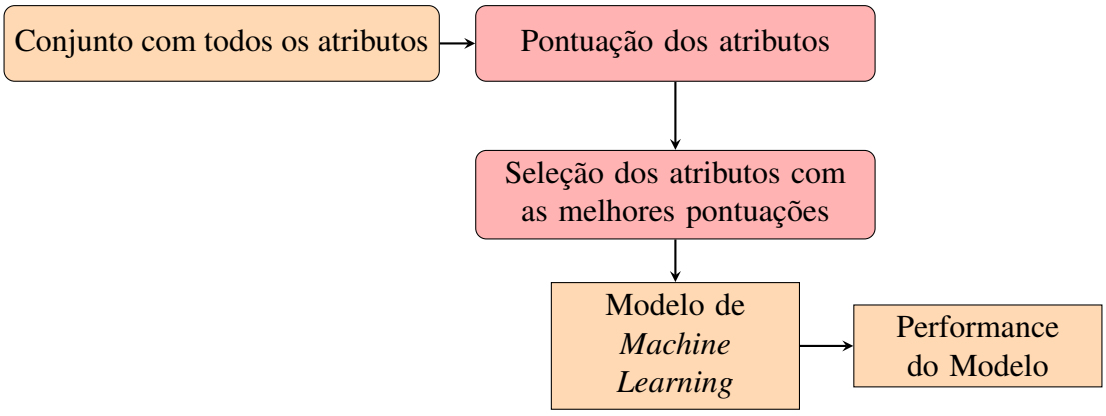

Figura 2.6 – Diagrama Seleção Baseada em Filtro

Referência: (DIAS, 2023)

Seleção baseada em embrulho

São considerados um conjunto de atributos e suas combinações de subconjuntos de atributos. Esses subconjuntos são comparados uns com os outros, e cada um é avaliado analisando seu impacto na performance do modelo, aqueles que obtiveram os melhores resultados são selecionados. Alguns algoritmos famosos para realizar essa seleção são Algoritmos Genéticos, Eliminação Recursiva de Características (RFE), e Seleção Sequencial de Atributos (SFS) (BANDELA; KUMAR, 2020).

Algumas desvantagens desse tipo de seleção são que apresentam um alto custo computacional, além de ser suscetível ao *overfitting*, casos em que o modelo se adéqua estritamente ao conjunto de dados de treinamento, resultando na perda da capacidade de generalização (KOBAYASHI; NAKAJI; YAMAMOTO, 2022). Já uma vantagem é a sua precisão (BANDELA; KUMAR, 2020).

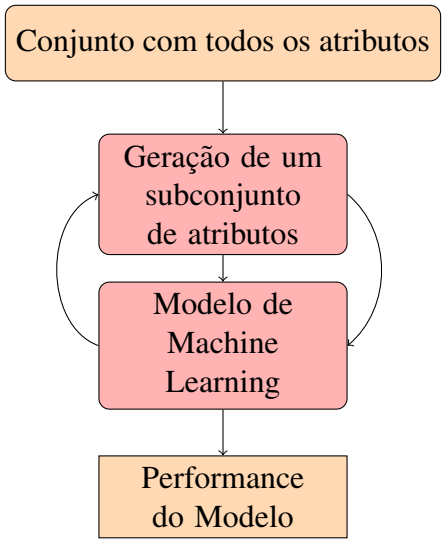

Figura 2.7 – Diagrama Seleção Baseada em Embrulho

Referência: (DIAS, 2023)

Seleção por métodos de incorporação

Nesse caso a seleção de atributos é feita durante a criação do modelo no processo de treinamento, são selecionados os melhores atributos que podem ser usados para aumentar a acurácia. Técnicas de regularização são as mais utilizadas nesse tipo de seleção, como LASSO, FSASL, Regressão de Ridge, *Elastic Net*, entre outras (BANDELA; KUMAR, 2020).

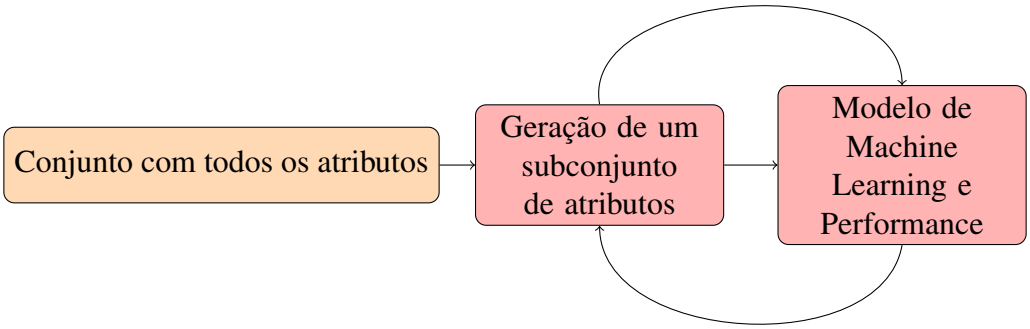

Figura 2.8 – Diagrama Seleção por métodos de incorporação

Referência: (DIAS, 2023)

Método Híbrido

É feito uma combinação de dois ou mais métodos de seleção de atributos. O principal objetivo é obter as vantagens das duas técnicas, resultando em um aumento na eficiência, performance na predição e uma possível redução da complexidade computacional (BANDELA; KUMAR, 2020).

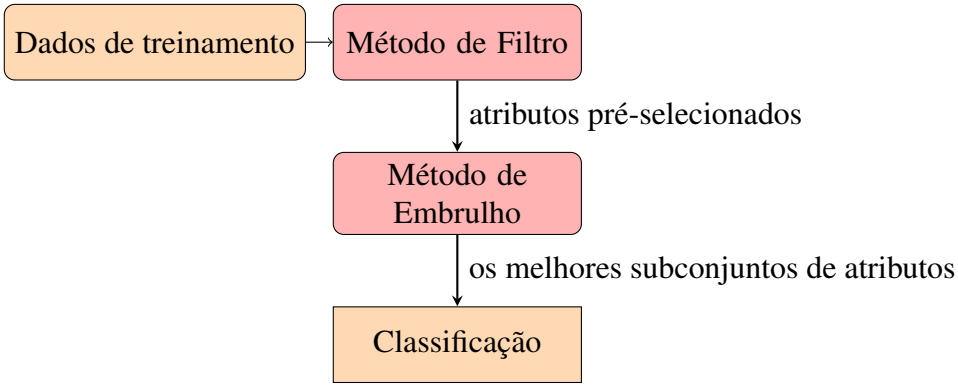

Figura 2.9 – Diagrama Seleção por método híbrido

Referência: (DIAS, 2023)

Ensemble

Nesse método, é construída uma coleção de subconjuntos de atributos, e um resultado agregado é produzido a partir do grupo, com o objetivo de criar o subconjunto ideal (BANDELA; KUMAR, 2020). Uma aplicação pode ser vista na figura 2.10.

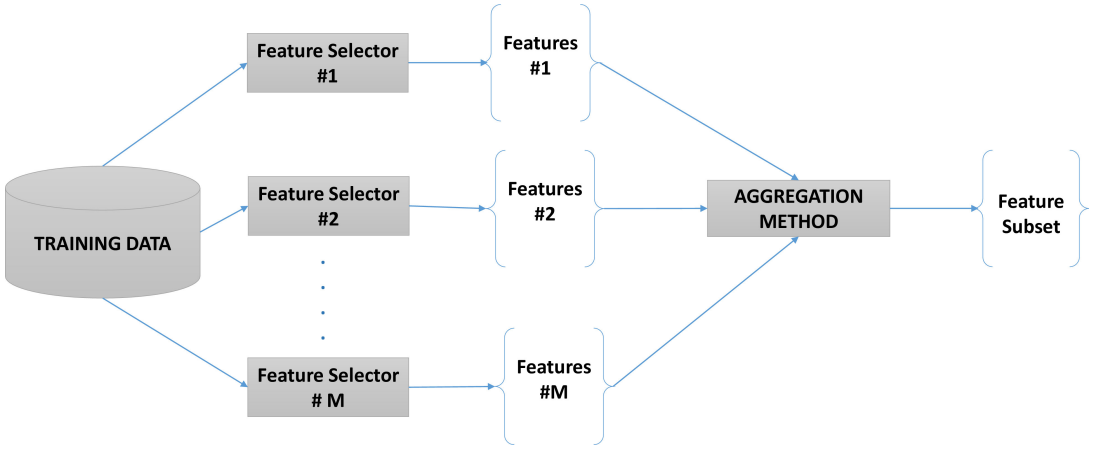

Figura 2.10 – Exemplo de projeto que utiliza o método Ensemble

Fonte: (BOLÓN-CANEDO; ALONSO-BETANZOS, 2019)

Principais Algoritmos *Feature Selection*

Pearson's Linear Correlation Coefficient

Um método de seleção de características que filtra dados com base na análise individual de suas características e na avaliação de sua correlação. Um bom subconjunto precisa ter uma baixa correlação entre os seus elementos, e uma correlação alta com a saída, que é a classe correspondente.

O estudo feito por Silva et al. (SILVA; BISPO; TEIXEIRA, 2021) emprega um método de seleção progressiva sequencial que utiliza uma correlação linear para avaliar a importância de cada parâmetro. No início do algoritmo, o espaço de busca está vazio e, a cada iteração, um parâmetro é analisado. O critério de parada é alcançado quando todos os atributos foram incluídos. A acurácia é calculada e armazenada em cada iteração para avaliar o desempenho do modelo. O subconjunto que alcança a maior acurácia é então selecionado.

Multilinear Regression Analysis

Esse método começa com um conjunto inicial de pesos e procede comparando a capacidade explicativa de modelos progressivamente maiores e menores. Em cada etapa, é calculado o valor-p de uma estatística F para testar modelos com e sem uma determinada característica candidata. Se a característica não está atualmente no modelo, a hipótese nula é que essa característica teria um coeficiente zero se adicionada. Por outro lado, se há evidências suficientes para rejeitar essa hipótese, a característica é incluída. Se uma característica já está no modelo,

a hipótese nula é que o coeficiente dessa característica é zero. Se não houver evidências suficientes para rejeitar essa hipótese, a característica é removida. O modelo multilinear pode ser representado pela equação:

$$Y_c = a + b_1 x_1 + b_2 x_2 + ... + b_k x_k \tag{2.12}$$

Fonte: (SILVA; BISPO; TEIXEIRA, 2021)

Onde a corresponde a interceptação no eixo y, $b_i x_i$ corresponde ao peso do atributo de número i, e k corresponde ao número de características independentes, e Y_c é a saída, ou a variável dependente.

A determinação da relevância das características é baseada em uma análise de regressão multilinear. Em cada fase, é calculado o valor-p de um teste de estatística F de modo a avaliar o desempenho do modelo. Para um atributo ser considerado relevante, ele deve apresentar alta significância estatística. O algoritmo utiliza como critério padrão para a inclusão de um parâmetro no modelo um valor-p menor que 0,05 e para remoção do modelo um valor maior que 0,10 (SILVA; BISPO; TEIXEIRA, 2021).

Teste t de Welch

Esse teste é uma alternativa ao teste t clássico, que normalmente avalia conjunto de características que possuem variações próximas. De forma geral esse teste compara a média de duas amostras de dados e verifica as diferenças estatísticas entre elas. A fórmula geral para calcular o valor de t, que mede a diferenças entre as características, é definida a seguir:

$$t = \frac{(u_1 - u_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{2.13}$$

Fonte: (SILVA; BISPO; TEIXEIRA, 2021)

Onde: u_1 e u_2 são as médias das duas amostras, S_1 e S_2 os desvios padrão das amostras, e n_1 e n_2 o tamanho das amostras.

Além disso é preciso calcular o grau de liberdade, que ajuda a determinar a forma da distribuição t. Sua fórmula é a seguinte, utilizando a aproximação de Satterthwaite:

$$v = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{\left[\frac{s_1^2}{n_1}\right]^2}{n_1 - 1} + \frac{\left[\frac{s_2^2}{n_2}\right]^2}{n_2 - 1}} \tag{2.14}$$

Fonte: (SILVA; BISPO; TEIXEIRA, 2021)

A partir disso é consultado a tabela de distribuição-t, localizando o valor crítico t_c, caso $t > t_c$, é possível concluir que há uma diferença significativa entre as médias das duas amostras, e portanto essas características são relevantes para realizar a distinção entre emoções (SILVA; BISPO; TEIXEIRA, 2021).

2.2.7 Classificadores

A classificação está relacionada ao aprendizado supervisionado, onde o computador aprende a mapear dados de treinamento em um espaço de atributos para determinados rótulos. As amostras podem pertencer a uma classe binária, resultando em apenas duas possíveis saídas, ou uma multi-classe, onde o modelo pode retornar diferentes respostas (KERKENI et al., 2019).

Existem vários tipos de algoritmos de classificação, alguns deles são: Regressão Linear, Árvore de decisão, SVM, *Naive Bayes*, KNN, *Random Forest*, Rede Neural, Modelo de Mistura Gaussiana, Modelo Oculto de Markov, Redes Neurais Recorrentes, e outros (KERKENI et al., 2019).

Principais Classificadores

SVM

O SVM constrói um hiperplano entre múltiplas classes (SHINDE; PATIL, 2023), formando um modelo linear baseado nos vetores de suporte para estimar a função de decisão. Quando os dados de treinamento são linearmente separáveis, o SVM encontra o hiperplano ótimo que separa os dados. Os vetores de suporte são cruciais para esta tarefa, pois definem o hiperplano ideal para a classificação (THIRUVENGATANADHAN, 2018).

A função kernel gera os produtos internos para a construção de máquinas com diferentes

superfícies de decisão não lineares no espaço de entrada. Ao utilizar diferentes funções de kernel, o método SVM transforma o espaço de dados de entrada no espaço de características de alta dimensão (SHINDE; PATIL, 2023).

Existe uma variedade de funções kernel, como linear, polinomial, base radial (RBF) e sigmóide. Em outros estudos é utilizado o kernel polinomial, definido pela seguinte equação (KERKENI et al., 2019):

$$K(x_i, x_j) = (\langle x_i, x_j \rangle + c_0)^d \tag{2.15}$$

Fonte: (KERKENI et al., 2019)

Onde d representa o grau polinomial.

Em contrapartida, existem outras pesquisas onde é utilizada a função kernel base radial gaussiana para classificar os sinais de fala em sete emoções (LANGARI; MARVI; ZAHEDI, 2020).

O SVM destaca-se especialmente quando há um número de dados para treinamento limitado, pois demonstra uma boa performance mesmo nesses casos (KERKENI et al., 2019).

CNN

Uma estrutura CNN pode ser aplicada utilizando uma maneira similar à pesquisa realizada por Huang et al. (HUANG et al., 2014), onde foi desenvolvido um modelo de classificação de emoções utilizando uma CNN com quatro principais camadas: camada convolucional, camada de ativação utilizando ReLU, uma camada de *Max Pooling*, e uma camada densa, separada em quatro módulos.

No primeiro módulo é feito a extração de características do áudio e a sua visualização, utilizando a biblioteca Librosa.

No segundo módulo acontece o treinamento do modelo e o cálculo da sua acurácia, utilizando uma função MFCC e a biblioteca Librosa para determinar a taxa de amostragem.

O terceiro módulo envolve a implementação da CNN, o sinal da fala foi representado bidimensionalmente com três camadas, permitindo à CNN realizar predições e aprender a identificar palavras e suas intonações.

Por fim, no módulo 4 ocorre a classificação do áudio em emoções.

2.2.8 Critérios para avaliação dos resultados

A escolha de um bom critério de avaliação é crucial para analisar a performance e o resultado do modelo. Existem diversas fórmulas que podem ser aplicadas para realizar essa tarefa, porém a mais popular utilizada em diversas pesquisas é a acurácia (MADANIAN et al., 2023), por ser a medida mais popular e com uma boa performance. Seu valor é obtido com a seguinte fórmula:

$$acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.16}$$

Fonte: (MADANIAN et al., 2023)

Onde VP e VN indicam se o modelo previu corretamente o resultado esperado, é realizado o cálculo de quantas respostas foram corretas, levando em conta o total de previsões.
Existem variantes desse cálculo, como a acurácia com e sem peso. A acurácia com peso considera o número de amostras em cada classe, dando maior peso às classes mais representadas. Já na acurácia sem peso, todas as classes têm o mesmo peso (YAM, 2015), ela é mais utilizada na avaliação da performance de um modelo SER que possui uma base de dados não balanceada (MADANIAN et al., 2023).

Em outros trabalhos, é utilizado o coeficiente de correlação de concordância (CCC) (MADANIAN et al., 2023). Essa métrica calcula o nível de concordância entre duas variáveis com base na correlação linear, variando de -1 a 1. Um valor perto de 1 indica uma proximidade da completa concordância, -1 representa discordância total, e 0 indica nenhuma correlação entre as variáveis (MADANIAN et al., 2023).

Métricas Adicionais para Avaliação da Performance

De forma a complementar o resultado da acurácia, também podem ser aplicados uma matriz de confusão, e medidas como precisão, sensibilidade, especificidade e *recall*.

- Matriz de confusão: dentro dela podem ser encontrados o número de previsões para cada classe, um exemplo dessa matriz é a Tabela 2.1.

Tabela 2.1 – Matriz de confusão.

	Valor Predito	
Valor Real	Negativo	Positivo
Negativo	VN	FP
Positivo	FN	VP

- Precisão: é um cálculo baseado no número de vezes que o modelo acertou com base no número de vezes que ele afirmou ser aquela resposta.

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{2.17}$$

Figura 2.11 – Fórmula de Precisão

Fonte: (SHUNG, 2018)

- Sensibilidade: calcula a taxa de acertos de resultados que são positivos.

$$sensibilidade = \frac{VP}{VP + FN} \tag{2.18}$$

Fonte: (CARVAJAL; ROWE, 2010)

- Especificidade: a taxa de acertos de resultados que são negativos.

$$especificidade = \frac{VN}{FP + VN} \tag{2.19}$$

Fonte: (CARVAJAL; ROWE, 2010)

- *Recall*: é a taxa de detecção, de todos os dados que são positivos, quantos foram corretamente preditos.

$$recall = \frac{VP}{VP + FN} \tag{2.20}$$

Fonte: (SHUNG, 2018)

- F1-score: pode ser utilizado caso a base de dados não for balanceada, ele calcula a média harmônica entre a precisão e o *recall*.

$$F1 - score = 2 * \frac{P * R}{P + R} \tag{2.21}$$

Fonte: (SILVA; BISPO; TEIXEIRA, 2021)

2.3 Trabalhos Correlatos

Foram criadas as tabelas 2.2, 2.3 e 2.4 com as pesquisas citadas ao longo desse capítulo, mostrando os métodos utilizados em cada etapa, a taxa de reconhecimento, e o número de emoções consideradas.

Pesquisa/autor	**Base de dados**	**Pré-processamento**	**Feature Extraction**
LIU et al., 2018	CASIA	não mencionado	Características Dependentes e Independentes do Locutor
LANGARI; MARVI; ZAHEDI, 2020	EMO-DB	Normalização do sinal	Coeficientes LPCC, MFCC e Fourier
ANAGNOSTOPOULOS; ILIOU, 2010	EMO-DB	Framing, Hamming Window	Pitch, MFCC, Energia, Formante
HUANG et al., 2014	RAVDESS	Normalização do sinal	Librosa e CNN
AGGARWAL; KUMAR, 2022	não mencionado	Conversão A/D, Preemphasis, Windowing	MFCC
HEMA; MARQUEZ, 2023	RAVDESS e CREMA-D	Filtro ruído de fundo, normalização e segmentação de sinais	MFCC
SHINDE; PATIL, 2023	EMO-DB	Preemphasis, Framing e Windowing	Librosa, MFCC, Chroma, ZCR, Mel Spectogram e DWT
YILDIRIM; KAYA; KILIC, 2021	IEMOCAP e EMO-DB	não mencionado	OpenSmile toolbox
KERKENI et al., 2019	EMO-DB e Spanish Database	Normalização e decomposição do sinal com EMD e TKEO	MS, MFF, SMFCC, ECC, EFCC

Tabela 2.2 – Métodos de pré-processamento e extração de características

Pesquisa/autor	**Feature Selection**	**Classificador**
LIU et al., 2018	Correlation Analysis e Fisher Criterion	ELM decision tree
LANGARI; MARVI; ZAHEDI, 2020	Algoritmo Genético e Cuckoo Search	SVM
ANAGNOSTOPOULOS; ILIOU, 2010	WEKA data mining tool	Multilayered Perceptron e SVM
HUANG et al., 2014	não mencionado	CNN
AGGARWAL; KUMAR, 2022	não mencionado	CNN
HEMA; MARQUEZ, 2023	não mencionado	CNN
SHINDE; PATIL, 2023	Teste estatístico ANOVA	SVM e MLP
YILDIRIM; KAYA; KILIC, 2021	ReliefF, Modified Binary Cuckoo Search e Modified NSGA-II	kNN, TreeBagger e SVM
KERKENI et al., 2019	RFE	SVM e RNNs

Tabela 2.3 – Métodos de seleção de características e classificadores

Pesquisa/autor	**Taxa de Reconhecimento**	**Emoções Classificadas**
LIU et al., 2018	89.60%	6
LANGARI; MARVI; ZAHEDI, 2020	92.75%	6
ANAGNOSTOPOULOS; ILIOU, 2010	MP: 86.6% SVM: 78%	7
HUANG et al., 2014	71%	7
AGGARWAL; KUMAR, 2022	não mencionado	8
HEMA; MARQUEZ, 2023	78%	8
SHINDE; PATIL, 2023	SVM: 89.70% MLP: 87.32%	4
YILDIRIM; KAYA; KILIC, 2021	EMODB, M-Cuckoo e SVM: 87.66% IEMOCAP, M-Cuckoo e SVM: 69.30%	EMODB: 7 IEMOCAP: 4
KERKENI et al., 2019	EMODB: SVM e SMFCC+MFF+MS= 85.63% Spanish database: SVM e SMFCC+MFF+MS+EFCC+Sem normalização = 90.69%	EMODB: 7 Spanish Database: 7

Tabela 2.4 – Resultados e número de emoções classificadas

2.4 Desafios e Tendências

Um desafio nesse ramo de pesquisa é a falta de métodos eficientes para descrever estados emocionais complexos, devido ao fato de que os estudos envolvendo SER ainda são recentes (LIU et al., 2018). Isso justifica a existência de poucas aplicações no mundo real, apesar de sua importância.

Outro desafio é em relação à baixa generalidade dos sistemas desenvolvidos. Muitos modelos apresentam bons resultados e alta acurácia apenas com os dados da base de dados utilizada para treinamento. Sua performance decai significativamente ao usar dados externos, algoritmos de *feature extraction* são utilizados para reduzir esse problema, mas ainda existe uma lacuna entre os resultados obtidos utilizando os dados externos e os dados da base de treinamento.

Para que sistemas SER sejam mais aplicáveis no cenário real, há uma necessidade crescente de se utilizar mais critérios de avaliação, de modo a garantir seu funcionamento. Parâmetros como privacidade, segurança e equidade estão sendo considerados para serem utilizados juntamente ao valor da acurácia. A privacidade avalia a capacidade do modelo de proteger informações sensíveis dos proprietários dos dados, enquanto segurança verifica se o modelo consegue produzir resultados confiáveis resistindo a possíveis ataques. A equidade é calculada com base na existência de uma discrepância entre a performance do modelo para indivíduos ou grupos específicos (MADANIAN et al., 2023).

Outros estudos avaliam a eficiência com base na otimização de recursos, pois a complexidade de um sistema limita a sua aplicação e implementação prática. São utilizados parâmetros como estabilidade, complexidade computacional e eficiência de dados (MADANIAN et al., 2023).

Capítulo 3

Detalhamento do Trabalho

3.1 Base de dados

As bases de dados utilizadas neste trabalho foram as seguintes: RAVDESS, CREMA, TESS, SAVEE.

Alguns dos principais motivos para a escolha dessas bases são o fato de estarem em inglês, o que facilita sua aplicação em outros projetos, já que é a língua predominante nesta área de pesquisa. Além disso, elas cobrem uma ampla gama de emoções, e ao combiná-las, obtém-se mais de 12.000 amostras de voz, que é uma ótima base inicial para o treinamento do modelo.

3.2 Bibliotecas Utilizadas

Biblioteca	**Descrição**
Pandas	Manipulação e análise de dados em estruturas de dados, como a criação de DataFrames.
Numpy	Execução de operações matemáticas e computação numérica.
Os	Oferece funcionalidade para interagir com o sistema operacional.
Sys	Manipulação de avisos e exceções relacionados ao interpretador Python.
Librosa	Auxilia na análise e manipulação de áudio.
Seaborn e matplotlib.pyplot	Visualização de dados, como a criação de gráficos.
Sklearn.preprocessing	Fornece ferramentas para realizar o pré-processamento de dados.
Sklearn.metrics	Utilizada para avaliar o desempenho final do modelo de classificação, gerando a matriz de confusão e calculando métricas como precisão, *recall* e *F1 score*.
Sklearn.model_selection	Utilizada para dividir o conjunto de dados para treino e teste.
Keras e tensorflow.keras	Fornece ferramentas para realizar a construção, o treinamento e a avaliação de redes neurais.
Scipy.signal	Implementação de filtros de áudio.
Multiprocessing e joblib	Permite a aceleração do processamento ao realizar as tarefas de forma paralela, utilizando os múltiplos núcleos do processador.
Timeit	Usado para medir o tempo de execução do código.
Scipy.stats	Aplica a técnica de seleção de características.

Tabela 3.1 – Descrição de bibliotecas e ferramentas utilizadas

3.3 Especificações da máquina

Ao realizar os testes necessários para garantir que o modelo está funcionando corretamente, o processamento foi realizado utilizando um computador com as seguintes especificações:

- Processador Intel Core i5-9400F
- Memória RAM 16GB, 2666MHz, DDR4
- Placa de vídeo GeForce GTX 1660 Super

3.4 Algoritmo

O diagrama 3.1 ilustra o processo geral que o algoritmo percorre até a classificação das emoções.

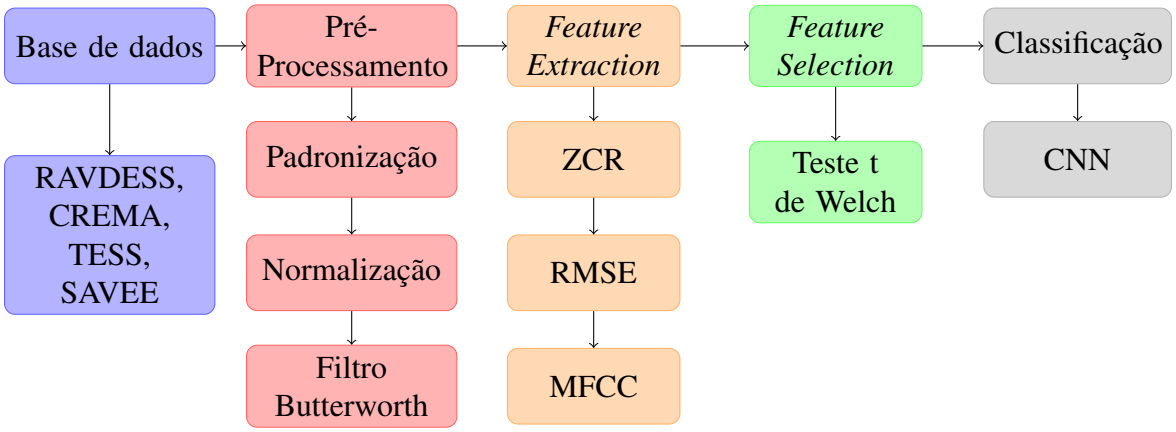

Figura 3.1 – Diagrama Completo do Modelo de Máquina

3.4.1 Carregar e rotular áudios

Cada base de dados nomeou os arquivos de forma diferente, então foi necessário tratar cada uma delas separadamente para que estejam em um padrão antes de processar os áudios.

RAVDESS

Os nomes dos arquivos dessa base de dados são compostos por 7 números, separados pelo caractere "-", sendo cada número o representante de alguma característica do áudio. Mais detalhado a seguir:

- Modalidade (01 = áudio e vídeo, 02 = apenas vídeo, 03 = apenas áudio).
- *Vocal channel* (01 = fala, 02 = música).

- Emoção (01 = neutro , 02 = calmo, 03 = felicidade, 04 = tristeza, 05 = raiva, 06 = medo, 07 = nojo, 08 = surpreso).
- Intensidade Emocional (01 = normal, 02 = forte)
- Frase (01 = *"Kids are talking by the door"*, 02 = *"Dogs are sitting by the door"*).
- Repetição (01 = primeira repetição, 02 = segunda repetição).
- Ator (01 até 24, números ímpares são atores, números pares são atrizes).

O processo seguido para retirar as informações dos áudios foram:

1. Primeiramente o diretório dos áudios de cada ator foi carregado.
2. A lista de diretórios percorrida.
3. A emoção que está sendo expressa no arquivo é identificada.
4. O nome da emoção é colocado em um vetor, e o diretório do arquivo em outro.
5. É criado um *dataframe*, concatenando os dados de emoções e os caminhos para cada arquivo.
6. Por fim as emoções são transformadas de números para texto (1 e 2: 'neutral', 3: 'happy', 4: 'sad'...).

CREMA-D

1. Novamente é carregado a lista de arquivos contido no diretório do *dataset*.
2. O terceiro elemento do nome de cada arquivo contém o código da emoção, elas então são mapeadas para rótulos de texto ('ANG':'angry', 'SAD':'sad', 'DIS':'disgust', 'FEA':'fear', 'HAP':'happy', 'NEU':'neutral').
3. Um *dataframe* é criado com os caminhos de cada arquivo de áudio.

TESS

1. Os diretórios do *dataset* são carregados, cada um correspondente a uma emoção.

2. A emoção contida no áudio é extraída.

    (a) Se o arquivo contém ‘ps’, a emoção é mapeada para surpresa; nos outros casos, o nome da emoção já está codificado no arquivo.

3. Um *dataframe* é criado para armazenar as emoções e o seu caminho.

SAVEE

1. Os arquivos do *dataset* são carregados.

2. A emoção é extraída a partir da letra abreviada do título do arquivo.

    (a) ‘a’: ‘angry’, ‘d’: ‘disgust’, ‘f’: ‘fear’, ‘h’: ‘happy’, ‘n’: ‘neutral’, ‘sa’: ‘sad’

    (b) Caso não seja nenhum desses a emoção é a surpresa.

3. As emoções e os caminhos dos arquivos são colocados em listas, utilizadas para criar um *dataframe* que contém essas informações.

No fim desse processo é criado novamente um *dataframe*, porém composto pelos 4 *dataframes* anteriormente criados para cada base de dados, contendo o nome da emoção e o seu caminho.

Após essas etapas a base de dados final utilizada possui a seguinte quantidade de áudios, separados por emoção:

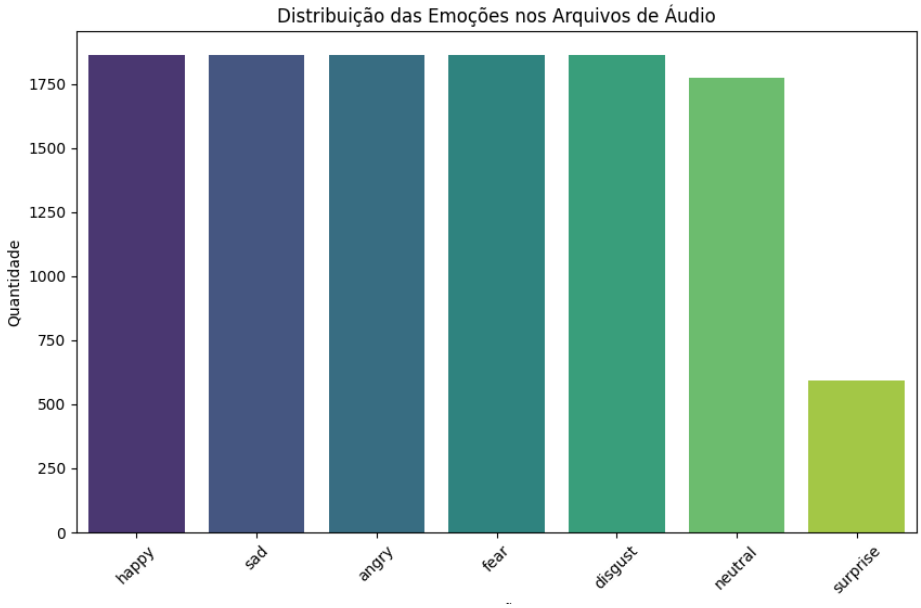

Figura 3.2 – Quantidade de áudios para cada classe

3.4.2 Procedimentos pré-processamento

1. Padronização: cada áudio é carregado e padronizado em termos de duração e *offset*, garantindo consistência entre todos os arquivos.

2. Normalização: os valores de amplitude são divididos pelo valor máximo absoluto, mantendo os dados dentro de um intervalo padrão e removendo variações de amplitude nas gravações.

3. Filtro *Butterworth*: primeiramente é aplicado um filtro passa-baixa, permitindo a passagem de frequências que estão abaixo de 8000 Hz. Em seguida é utilizado um filtro passa-banda, permitindo frequências entre 100 Hz e 8000 Hz.

3.4.3 *Feature Extraction*

1. Primeiramente é aplicada a função ZCR, que calcula a taxa de cruzamento por zero, ou seja, quantas vezes o sinal cruza o eixo zero.
2. Função RMSE, determinando a energia do sinal em cada *frame*.
3. Foi aplicado o MFCC, que calcula os coeficientes cepstrais de Mel, obtendo informações relevantes do espectro de frequências.
4. As características extraídas são armazenadas em X, e as emoções correspondentes em Y.
5. Após todas as tarefas paralelas forem concluídas, os resultados são reunidos novamente em duas listas: X (características) e Y (emoções).
6. É criado um *dataframe* que contém todas as características e emoções.
7. É verificado se existe valores ausentes, e caso encontrado, essas lacunas são preenchidas com 0.
8. Por fim, o *dataframe* é dividido em X (coluna de características) e Y (rótulos das emoções).

3.4.4 *Feature Selection*

Para a seleção de características, foram testados diversos métodos, como a correlação de Pearson, RFE e *Multilinear Regression Analysis*. No entanto, o método que apresentou melhores resultados foi o teste t de Welch. Os parâmetros utilizados são:

- Matriz de características X.
- Rótulos de destino Y.
- P_value_threshold, que define o limite do valor de p para a seleção das características.

O processo de seleção das características foi a seguinte:

1. Decodificação de Y: Conversão dos rótulos para classe inteira.
2. É criada uma lista de índices das características selecionadas.
3. Iteração sobre cada característica:
    - É aplicado o teste t de Welch para todas as combinações de pares de classes.
    - Comparação das médias das amostras de duas classes.
    - O p-value resultante é armazenado.
4. Após realizar os cálculos de *p-value* para cada característica em todos os pares de classes, verifica-se se o menor *p-value* é inferior ao limiar definido. Caso seja, a característica é adicionada à lista de selecionados.
5. São geradas novas matrizes de treino e teste, com as características selecionadas.

3.4.5 Classificação

CNN

A arquitetura construída para a função de classificador foi de uma CNN com múltiplas camadas, ilustrada na figura 3.3, e detalhada a seguir:

1. Divisão dos dados: 80% dos dados foram separados para o conjunto de treinamento, e o conjunto teste foi composto por 20% restante dos dados.
2. Normalização dos dados, para garantir que todos os dados estejam na mesma escala.
3. Redimensionamento dos dados para o formato tridimensional, para ser compatível com processamento feito por uma CNN.
4. Arquitetura da Rede Neural Convolucional (CNN):

(a) Cinco camadas de convolução (Conv1D) com ativação ReLU, essas camadas vão extrair as características dos dados e detectar padrões.

   i. Duas camadas com 512 filtros, kernel de tamanho cinco.

   ii. Duas camadas com 256 filtros, kernel de tamanho cinco e três.

   iii. Uma camadas com 128 filtros, kernel de tamanho três.

(b) Seis camadas de normalização por lote (*BatchNormalization*).

   i. Uma após cada camada de convolução, normalizando as ativações da camada anterior.

(c) Cinco camadas de *pooling* máximo (MaxPool1D).

   i. Essa camada reduz a dimensionalidade da saída da camada convolucional anterior.

(d) Três camadas de *dropout*.

   i. Após algumas das camadas de *pooling*, onde 20% dos neurônios são desativados, prevenindo *overfitting*.

(e) Duas camadas densas.

   i. Uma com 512 neurônios e ativação ReLU, totalmente conectada, recebendo a saída de todas as camadas anteriores.

   ii. Uma camada de saída com sete neurônios e ativação *softmax*, correspondendo às sete classes possíveis.

(f) Uma camada *flatten*.

   i. Transforma saída da rede convolucional (multidimensional) em um vetor unidimensional antes das camadas densas.

5. A compilação do modelo é feita com o otimizador Adam, com perda categórica (por ser um problema de classificação multiclasse), e a métrica de avaliação sendo a acurácia.

6. O treinamento é realizado por 10 épocas, com um tamanho de lote de 32, utilizando o conjunto de treinamento e validando o desempenho com o conjunto de teste.

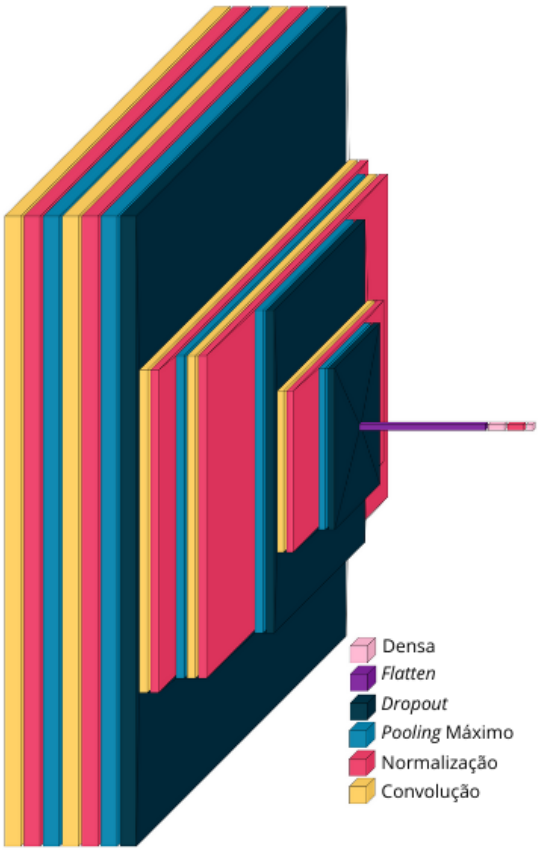

Figura 3.3 – Modelo CNN utilizado no trabalho

3.4.6 Métricas de Avaliação

- Matriz de confusão: avalia o desempenho do modelo, comparando as previsões com os rótulos reais.
- Relatório de Classificação: apresenta métricas como precisão, *recall* e *F1-score*.
- Acurácia: uma das métricas mais importantes utilizadas para avaliar a performance geral do modelo, o valor é exibido ao final do treinamento, indicando a acurácia do modelo no conjunto de testes.
- Gráficos: foram gerados gráficos que comparam a perda e a acurácia ao longo do treinamento e teste, o que permite verificar se o modelo apresenta sinais de *overfitting*.

Capítulo 4

Resultados

Foram realizados cinco testes com o mesmo algoritmo para avaliar a consistência dos resultados. A seguir, são apresentados o melhor e o pior caso entre os múltiplos testes feitos, mostrando tabelas com relatórios de classificação após o treinamento do modelo, e os dados referentes a cada classe. Além disso, são mostrados o tempo total de execução do processamento, a matriz de confusão, e um gráfico que permite visualizar a perda e a acurácia nos conjuntos de teste e treino.

Ao final deste capítulo, estão apresentados os relatórios de classificação de todos os testes, de forma a facilitar a comparação dos resultados.

4.1 Pior Resultado

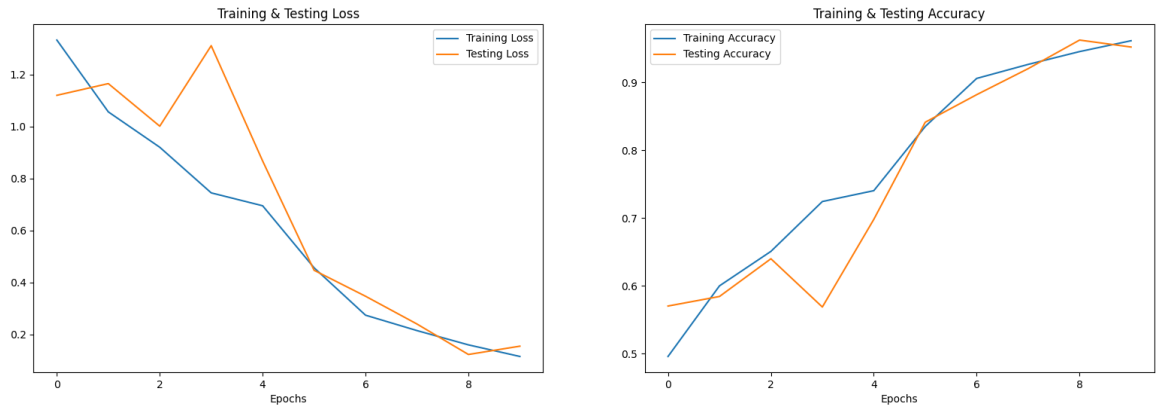

Figura 4.1 – Perda e acurácia ao longo das épocas na primeira instância

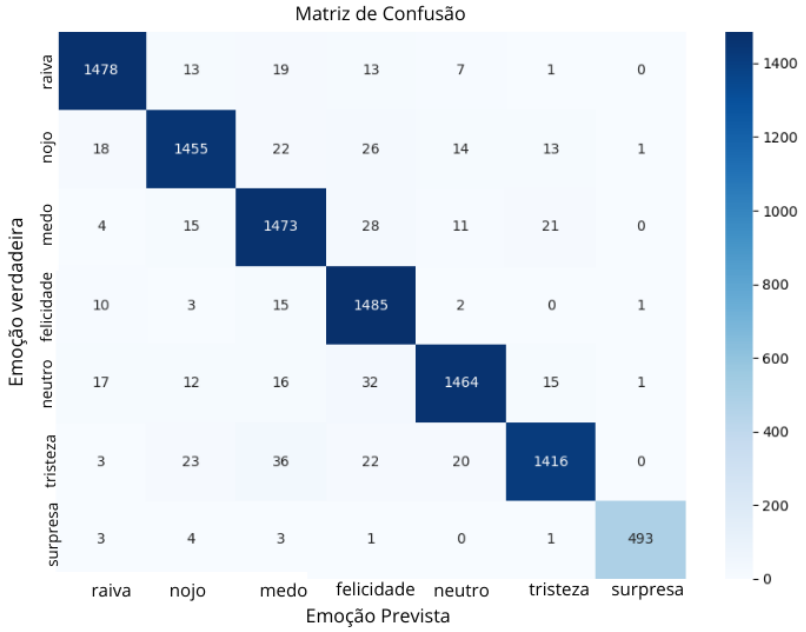

Figura 4.2 – Matriz de confusão no primeiro teste

Esses resultados foram obtidos a partir do primeiro teste, além de ser o caso com os piores resultados, também foi a instância com o maior tempo de execução, levando 84 minutos para ser concluída. A acurácia no conjunto de teste foi de 95%. Dentre as emoções, a surpresa apresentou a maior precisão, com 99%, enquanto a felicidade teve o menor desempenho, com

92%.

Ao analisar a Figura 4.2, é possível identificar os pares de classe em que o modelo mais confundiu. O maior número de confusões ocorreu entre as emoções tristeza e medo, sendo a tristeza classificada como medo em 36 ocasiões.

Emoção	**Precisão**	**Recall**	**F1-Score**	**Support**
Raiva	0.96	0.97	0.96	1531
Nojo	0.95	0.94	0.95	1549
Medo	0.93	0.95	0.94	1552
Felicidade	0.92	0.98	0.95	1516
Neutro	0.96	0.94	0.95	1557
Tristeza	0.97	0.93	0.95	1520
Surpresa	0.99	0.98	0.99	505
Acurácia Teste	0.95			9730
Macro avg	0.96	0.95	0.96	9730
Weighted avg	0.95	0.95	0.95	9730
Tempo de Execução	84 minutos			

Tabela 4.1 – Relatório de Classificação para o Pior Caso

4.2 Melhor Resultado

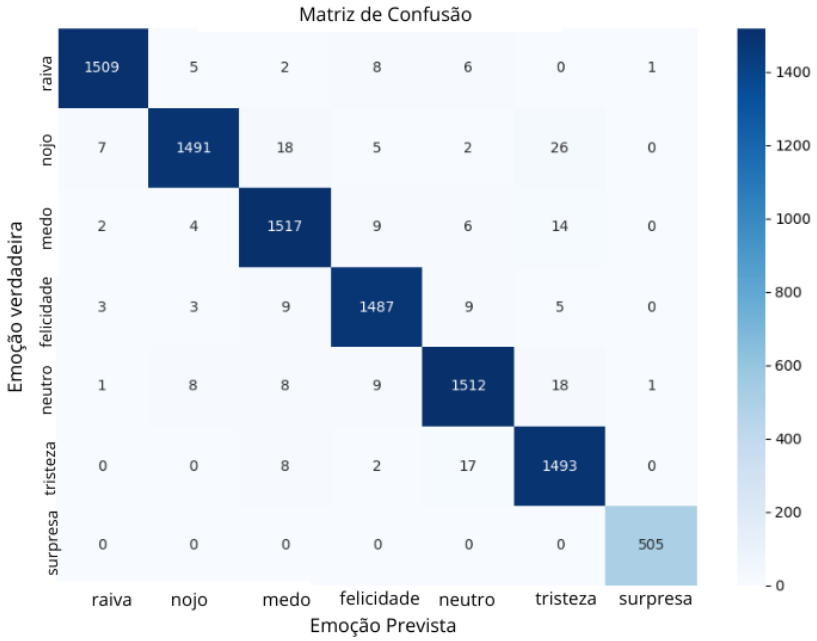

Figura 4.3 – Matriz de confusão no terceiro teste

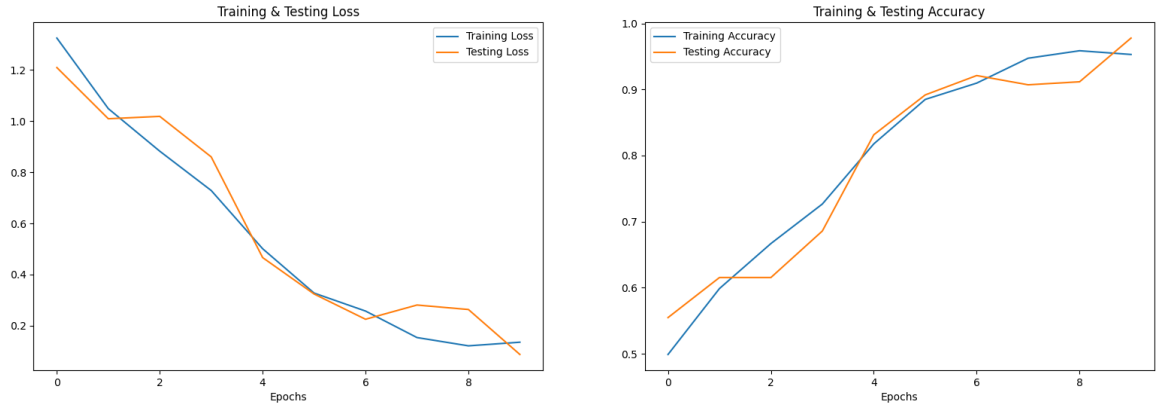

Figura 4.4 – Perda e acurácia ao longo das épocas na terceira instância

Esse foi o terceiro teste feito, e alcançou uma acurácia no conjunto de teste de 98%. A emoção com maior precisão foi novamente a surpresa, que obteve 100%, e aquela com a pior precisão foi a tristeza, com 96%.

A análise da Figura 4.3 revela que a emoção mais difícil de diferenciar foi o nojo, frequentemente confundido com tristeza, com 26 ocorrências. Também é possível apontar que no gráfico

apresentado pela Figura 4.4, as linhas que representam a perda e a acurácia mantêm-se mais próximas entre si quando comparadas ao gráfico do pior caso, ilustrado na Figura 4.1.

Emoção	**Precisão**	**Recall**	**F1-Score**	**Support**
Raiva	0.99	0.99	0.99	1531
Nojo	0.99	0.96	0.97	1549
Medo	0.97	0.98	0.97	1552
Felicidade	0.98	0.98	0.98	1516
Neutro	0.97	0.97	0.97	1557
Tristeza	0.96	0.98	0.97	1520
Surpresa	1.00	1.00	1.00	505
Acurácia Teste	0.98			9730
Macro avg	0.98	0.98	0.98	9730
Weighted avg	0.98	0.98	0.98	9730
Tempo de execução	80 minutos			

Tabela 4.2 – Relatório de Classificação para o Melhor Caso

4.3 Média dos Resultados

Emoção	**Precisão**	**Recall**	**F1-Score**	**Support**
Raiva	0.984	0.972	0.974	1531
Nojo	0.976	0.962	0.974	1549
Medo	0.954	0.968	0.960	1552
Felicidade	0.964	0.970	0.968	1516
Neutro	0.970	0.962	0.966	1557
Tristeza	0.954	0.972	0.962	1520
Surpresa	0.994	0.990	0.994	505
Acurácia Teste	0.968			9730
Macro avg	0.975	0.971	0.975	9730
Weighted avg	0.968	0.968	0.968	9730
Tempo de Execução	73.6 minutos			

Tabela 4.3 – Tabela de Médias

A Tabela 4.3 permite extrair informações importantes, como o fato de que a emoção que o modelo tem mais facilidade em identificar é a emoção surpresa, com 99,4% de precisão, apesar de ser a emoção com o menor volume de áudios. Por outro lado, as emoções com o pior desempenho são medo e tristeza, com 95,4% de precisão. Apesar dessas variações, a precisão

ainda é extremamente satisfatória, demonstrando que o modelo possui uma baixa taxa de falsos positivos e falsos negativos.

De modo geral, o desempenho do modelo foi excelente, com nenhuma métrica abaixo de 90%. Além disso, o alto F1-score reflete a capacidade do modelo de manter um equilíbrio entre precisão e *recall*, mostrando ser eficaz tanto em minimizar falsos positivos quanto em identificar a maioria dos verdadeiros positivos. No entanto, o custo computacional foi relativamente alto, com a GPU atingindo 100% de uso durante o treinamento do modelo, e o tempo de execução do programa demorado, ultrapassando a marca de uma hora em todos os testes.

4.4 Relatórios de Classificação de todos os testes

Emoção	**Precisão**	**Recall**	**F1-Score**	**Support**
Raiva	0.96	0.97	0.96	1531
Nojo	0.95	0.94	0.95	1549
Medo	0.93	0.95	0.94	1552
Felicidade	0.92	0.98	0.95	1516
Neutro	0.96	0.94	0.95	1557
Tristeza	0.97	0.93	0.95	1520
Surpresa	0.99	0.98	0.99	505
Acurácia Teste		0.95		9730
Macro avg	0.96	0.95	0.96	9730
Weighted avg	0.95	0.95	0.95	9730
Tempo de Execução	84 minutos			

Tabela 4.4 – Caso 1

Emoção	**Precisão**	**Recall**	**F1-Score**	**Support**
Raiva	0.99	0.98	0.98	1531
Nojo	0.98	0.97	0.98	1549
Medo	0.97	0.97	0.97	1552
Felicidade	0.97	0.98	0.98	1516
Neutro	0.98	0.97	0.97	1557
Tristeza	0.96	0.98	0.97	1520
Surpresa	1.00	0.99	1.00	505
Acurácia Teste		0.98		9730
Macro avg	0.98	0.98	0.98	9730
Weighted avg	0.98	0.98	0.98	9730
Tempo de Execução	67 minutos			

Tabela 4.5 – Caso 2

Emoção	**Precisão**	**Recall**	**F1-Score**	**Support**
Raiva	0.99	0.99	0.99	1531
Nojo	0.99	0.96	0.97	1549
Medo	0.97	0.98	0.97	1552
Felicidade	0.98	0.98	0.98	1516
Neutro	0.97	0.97	0.97	1557
Tristeza	0.96	0.98	0.97	1520
Surpresa	1.00	1.00	1.00	505
Acurácia Teste		0.98		9730
Macro avg	0.98	0.98	0.98	9730
Weighted avg	0.98	0.98	0.98	9730
Tempo de execução	80 minutos			

Tabela 4.6 – Caso 3

Emoção	**Precisão**	**Recall**	**F1-Score**	**Support**
Raiva	1.00	0.95	0.97	1531
Nojo	0.99	0.97	0.98	1549
Medo	0.96	0.98	0.97	1552
Felicidade	0.96	0.98	0.97	1516
Neutro	0.97	0.97	0.97	1557
Tristeza	0.96	0.98	0.97	1520
Surpresa	0.98	1.00	0.99	505
Acurácia Teste		0.97		9730
Macro avg	0.97	0.98	0.98	9730
Weighted avg	0.97	0.97	0.97	9730
Tempo de execução	70 minutos			

Tabela 4.7 – Caso 4

Emoção	**Precisão**	**Recall**	**F1-Score**	**Support**
Raiva	0.98	0.97	0.97	1531
Nojo	0.97	0.97	0.97	1549
Medo	0.94	0.96	0.95	1552
Felicidade	0.99	0.93	0.96	1516
Neutro	0.97	0.96	0.97	1557
Tristeza	0.92	0.99	0.95	1520
Surpresa	1.00	0.98	0.99	505
Acurácia Teste		0.96		9730
Macro avg	0.97	0.97	0.97	9730
Weighted avg	0.96	0.96	0.96	9730
Tempo de execução	67 minutos			

Tabela 4.8 – Caso 5

Capítulo 5

Conclusão

Para que haja uma interação entre humanos e máquinas de forma mais natural, é necessário que as máquinas sejam capazes de interpretar da melhor forma as emoções expressas, possibilitando tomadas de decisões mais assertivas e personalizadas. Neste trabalho foi apresentado um modelo de máquina capaz de analisar a voz e identificar a emoção expressa pelo locutor, contribuindo diretamente para pesquisas no ramo de SER.

A parte prática e experimental do estudo utilizou as base de dados RAVDESS, CREMA, RESS e SAVEE, cujos dados foram agrupados e submetidos a um pré-processamento antes de serem alimentados no modelo de *machine learning*. Foram aplicadas funções de extração de características, como ZCR e MFCC, e a função T de Welch para selecionar as características, antes da classificação dos áudios em classes de emoções. Os experimentos mostraram que o modelo criado foi bem-sucedido em distinguir as falas e classificar corretamente as emoções expressas. Essa validação foi realizada por meio de vários testes práticos utilizando o algoritmo desenvolvido em um ambiente real. O modelo alcançou até 98% de acurácia no conjunto de testes, demonstrando maior precisão ao predizer emoções como surpresa, enquanto encontrou mais dificuldades com medo e tristeza. A média de acurácia observada foi de 96,8%. Comparando os resultados com outros estudos, conclui-se que o modelo é bastante promissor, superando outros modelos reportados em diversos trabalhos, como aqueles feitos por Huang (HUANG et al., 2014) e Hema (HEMA; MARQUEZ, 2023), ambos utilizaram o classificador CNN, porém alcançaram taxas de reconhecimento de 71% e 78%, respectivamente, bem abaixo da taxa vista neste trabalho.

Apesar dos resultados positivos, o modelo ainda apresenta desafios, como o tempo de processamento necessário para a execução do algoritmo. Mesmo utilizando o poder computacional da GPU para o processamento, o tempo médio de execução foi de 73 minutos, um aspecto que pode ser otimizado futuramente. Outro problema discutido é a questão da generalização, embora o modelo tenha sido treinado com dados de várias bases, essas ainda não englobam toda a diversidade de vozes e emoções existentes. Portanto não é possível garantir que o modelo seja igualmente eficaz na identificação de emoções mais complexas ou utilizando áudios que foram gravados em diferentes ambientes. Todos os áudios das bases utilizadas foram encenados, então pode haver uma queda no desempenho ao predizer emoções contidas em áudios de situações reais. Dessa forma, para o avanço do projeto, seria recomendável incluir mais bases de dados que abranjam uma maior variedade de fontes de áudios, incluindo gravações fora de estúdios e com ruídos, para tornar o modelo mais robusto. Também seria pertinente a realização de testes com diferentes algoritmos de extração e seleção de características para analisar e avaliar variações no desempenho.

De forma prática, após os ajustes e melhorias, o modelo poderá ser aplicado em áreas como a educação, auxiliando alunos; no comércio eletrônico, com a recomendação de produtos para clientes; ou em aplicativos de música, realizando recomendações personalizadas para os usuários, impactando positivamente não apenas essas, mas também outras áreas da vida cotidiana.

Referências Bibliográficas

AGGARWAL, S.; KUMAR, S. Speech emotion recognition and analysis using mfcc & cnn. *Journal of Pharmaceutical Negative Results*, 2022.

AGHAJAN, H.; AUGUSTO, J. C.; DELGADO, R. L.-C. *Human-centric interfaces for ambient intelligence*. [S.l.]: Academic Press, 2009.

ALBUQUERQUE, E. S. G. Teste de percepção de dimensões emocionais do medo (tpde-m): um estudo de construção e evidências de validade. Universidade Federal de Pernambuco, 2020.

ANAGNOSTOPOULOS, C.-N.; ILIOU, T. Towards emotion recognition from speech: definition, problems and the materials of research. *Semantics in Adaptive and Personalized Services: Methods, Tools and Applications*, Springer, 2010.

BANDELA, S. R.; KUMAR, T. K. Speech emotion recognition using unsupervised feature selection algorithms. *Radioengineering*, v. 29, n. 2, 2020.

BOLÓN-CANEDO, V.; ALONSO-BETANZOS, A. Ensembles for feature selection: A review and future trends. *Information fusion*, Elsevier, v. 52, 2019.

CARVAJAL, D.; ROWE, P. Sensitivity, specificity, predictive values, and likelihood ratios. *Pediatrics in review / American Academy of Pediatrics*, v. 31, p. 511–3, 12 2010.

DIAS, M. *Importância da Seleção de Atributos em Modelos de Machine Learning*. 2023. ⟨https://aquare.la/importancia-da-selecao-de-atributos-em-modelos-de-machine-learning/⟩ [Accessed: 6 de Jul. de 2024.D].

FERREIRA, M. V. dos S. et al. Deep learning: Uma introdução às redes neurais convolucionais. 2017.

FRANKE, T. M.; HO, T.; CHRISTIE, C. A. The chi-square test: Often used and more often misinterpreted. *American journal of evaluation*, Sage Publications Sage CA: Los Angeles, CA, v. 33, n. 3, p. 448–458, 2012.

GUIDO, R. C. Effectively interpreting discrete wavelet transformed signals [lecture notes]. *IEEE Signal Processing Magazine*, IEEE, v. 34, n. 3, 2017.

GUIDO, R. C. Wavelets behind the scenes: Practical aspects, insights, and perspectives. *Physics Reports*, Elsevier, v. 985, 2022.

HEMA, C.; MARQUEZ, F. P. G. Emotional speech recognition using cnn and deep learning techniques. *Applied Acoustics*, Elsevier, v. 211, 2023.

HUANG, Z. et al. Speech emotion recognition using cnn. In: *Proceedings of the 22nd ACM international conference on Multimedia*. [S.l.: s.n.], 2014.

IBM. *O que são redes neurais convolucionais?* 2024. ⟨https://www.ibm.com/br-pt/topics/convolutional-neural-networks⟩ [Accessed: 6 de Jul. de 2024.].

KERKENI, L. et al. Automatic speech emotion recognition using an optimal combination of features based on emd-tkeo. *Speech Communication*, Elsevier, v. 114, 2019.

KOBAYASHI, M.; NAKAJI, K.; YAMAMOTO, N. Overfitting in quantum machine learning and entangling dropout. *Quantum Machine Intelligence*, Springer, v. 4, n. 2, p. 30, 2022.

KONG, F. et al. A practical non-profiled deep-learning-based power analysis with hybrid-supervised neural networks. *Electronics*, MDPI, v. 12, n. 15, 2023.

KWON, S. et al. Mlt-dnet: Speech emotion recognition using 1d dilated cnn based on multi-learning trick approach. *Expert Systems with Applications*, Elsevier, v. 167, 2021.

LANGARI, S.; MARVI, H.; ZAHEDI, M. Improving of feature selection in speech emotion recognition based-on hybrid evolutionary algorithms. *International Journal of Nonlinear Analysis and Applications*, Semnan University, v. 11, n. 1, 2020.

LI PETER LU, v.-c. B. *Seleção de recursos baseada em filtro*. 2023. ⟨https://learn.microsoft.com/pt-br/azure/machine-learning/component-reference/filter-based-feature-selection?view=azureml-api-2⟩ [Accessed: 6 de Jul. de 2024.].

LIU, Z.-T. et al. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, Elsevier, v. 273, 2018.

MADANIAN, S. et al. Speech emotion recognition using machine learning—a systematic review. *Intelligent systems with applications*, Elsevier, 2023.

MASHHADI, M. M. R.; OSEI-BONSU, K. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. *Plos one*, Public Library of Science San Francisco, CA USA, v. 18, n. 11, p. e0291500, 2023.

PHUNG, V. H.; RHEE, E. J. A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, MDPI, v. 9, n. 21, 2019.

SHINDE, A. S.; PATIL, V. V. Performance improvement in speech based emotion recognition with dwt and anova. *Communications in Mathematics and Applications*, RGN Publications, v. 14, n. 3, 2023.

SHUNG, K. P. *Accuracy, Precision, Recall or F1?* 2018. ⟨https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9⟩ [Accessed: 10 de Out. de 2024.].

SILVA, L.; BISPO, B.; TEIXEIRA, J. P. Features selection algorithms for classification of voice signals. *Procedia Computer Science*, Elsevier, v. 181, 2021.

SINGH, G. et al. Fine-grained emotion prediction by modeling emotion definitions. In: IEEE. *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. [S.l.], 2021.

THIRUVENGATANADHAN, R. Speech recognition using svm. *International Research Journal of Engineering and Technology (IRJET)*, v. 5, n. 09, 2018.

TSIOURTI, C. et al. Multimodal integration of emotional signals from voice, body, and context: Effects of (in) congruence on emotion recognition and attitudes towards robots. *International Journal of Social Robotics*, Springer, v. 11, 2019.

TUNCER, T.; DOGAN, S.; ACHARYA, U. R. Automated accurate speech emotion recognition system using twine shuffle pattern and iterative neighborhood component analysis techniques. *Knowledge-Based Systems*, Elsevier, v. 211, 2021.

YAM, C. *Data preparation: The balancing act*. 2015. ⟨https://devblogs.microsoft.com/ise/data-preparation-the-balancing-act⟩ [Accessed: 6 de Jul. de 2024.].

YILDIRIM, S.; KAYA, Y.; KILIÇ, F. A modified feature selection method based on metaheuristic algorithms for speech emotion recognition. *Applied Acoustics*, Elsevier, v. 173, 2021.