UNIVERSIDADE ESTADUAL DA PARAÍBA
CENTRO DE CIÊNCIAS EXATAS, SOCIAIS E APLICADAS
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

JHÔNATAN IZIDRO ALMEIDA

REDES NEURAIS CONVOLUCIONAIS APLICADAS AO RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA FALA

PATOS
2024

JHÔNATAN IZIDRO ALMEIDA

REDES NEURAIS CONVOLUCIONAIS APLICADAS AO RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA FALA

Trabalho de Conclusão de Curso apresentado ao Programa de Graduação em Ciência da Computação da Universidade Estadual da Paraíba.

Área de concentração: Inteligência Artificial

Orientador: Esp. Jaian Tales Gomes Santos

PATOS

2024

A447r Almeida, Jhonatan Izidro.
Redes neurais convolucionais aplicadas ao reconhecimento de emoções através da fala [manuscrito] / Jhonatan Izidro Almeida. - 2024.
61 f.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Ciência da computação) - Universidade Estadual da Paraíba, Centro de Ciências Exatas e Sociais Aplicadas, 2024.

"Orientação : Prof. Esp. Jaian Tales Gomes Santos, Coordenação do Curso de Computação - CCEA".

1. Redes neurais convolucionais. 2. Reconhecimento de emoções. 3. Mel-Frequency Cepstral Coefficients (MFCC). I. Título

21. ed. CDD 006.3

Elaborada por Kaliane Eveny Martins de Oliveira - CRB - 15/986 BSC7

JHONATAN IZIDRO ALMEIDA

REDES NEURAIS CONVOLUCIONAIS APLICADAS AO RECONHECIMENTO DE EMOÇÕES ATRAVÉS DA FALA

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Ciência da Computação da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação

Aprovada em: 22/11/2024.

BANCA EXAMINADORA

Documento assinado eletronicamente por:

- **Jaian Tales Gomes Santos** (***.796.864-**), em **02/12/2024 09:03:28** com chave **70cddaeab0a511efa2c21a7cc27eb1f9**.
- **José Aldo Silva da Costa** (***.862.334-**), em **02/12/2024 12:29:22** com chave **34a55012b0c211ef84691a1c3150b54b**.
- **Rosangela de Araujo Medeiros** (***.723.558-**), em **02/12/2024 15:13:19** com chave **1c34839cb0d911efa90406adb0a3afce**.

Documento emitido pelo SUAP. Para comprovar sua autenticidade, faça a leitura do QrCode ao lado ou acesse https://suap.uepb.edu.br/comum/autenticar_documento/ e informe os dados a seguir.
Tipo de Documento: Folha de Aprovação do Projeto Final
Data da Emissão: 20/12/2024
Código de Autenticação: 292e75

À minha família, que, incondicionalmente, sempre me apoiou e acreditou em mim, dedico este trabalho.

AGRADECIMENTOS

Eu quero expressar a minha gratidão a Deus pela saúde física e mental que me foram concedidas, permitindo não só a realização deste trabalho, mas todos os meus feitos até aqui.

Também quero agradecer, apesar de palavras não serem suficientes, o amor, cuidado, e apoio da minha família, especialmente dos meus pais que sempre, incondicionalmente, fizeram todo o possível para que eu pudesse ter uma boa educação e ser uma pessoa bem sucedida na vida e profissionalmente.

Expresso os meus sinceros e profundos agradecimentos ao meu orientador, Jaian Tales Gomes Santos que esteve sempre, atenciosamente, me auxiliando e guiando na jornada de elaboração deste trabalho. Além de agradecer, também, à professora Jannayna Domingues Barros Filgueira, que também me ajudou e foi importante na construção desse trabalho.

Quero expressar a meu reconhecimento e agradecer na pessoa do professor Jucelio Soares dos Santos, que sempre foi muito solicito e atencioso quando precisei de ajuda, pelo apoio da coordenação do curso com o qual sempre pude contar.

Saúdo e agradeço a cada um dos meus colegas com os quais tive o prazer de compartilhar todas as vivências e dificuldades da jornada acadêmica nesse período em que estivemos juntos.

Por fim, mas não menos importante, quero agradecer a todo o corpo docente da UEPB, que fazem desta universidade a grande e competente instituição de ensino que ela é, juntamente com todo o seu corpo técnico e administrativo.

“O futuro da tecnologia ameaça destruir tudo o que é humano no homem, mas a tecnologia não atinge a loucura; e nela então o humano do homem se refugia.”

Clarice Lispector.

RESUMO

Com a crescente onda de desenvolvimento tecnológico no campo da inteligência artificial, muitas outras áreas de conhecimento têm evoluído e se beneficiado dessa evolução, entre elas está a área de computação afetiva. No intuito de tornar a interação dos usuários com as máquinas ainda mais fluida e natural, pesquisas concernentes à computação afetiva têm sido realizadas no intuito de dotar as máquinas da capacidade de identificar, interpretar e sintetizar emoções. Porém, essa tarefa de classificação de emoções de forma automática enfrenta muitos desafios, dentre eles conseguir desenvolver um modelo que apresente bons resultados e que consiga generalizar bem a tarefa de classificação, além da limitada quantidade de bases de dados disponíveis para realizar o treinamento dos modelos de aprendizagem de máquina. A literatura apresenta modelos de redes neurais convolucionais treinadas a partir de características MFCC extraídas dos áudios de fala que apresentam resultados robustos na classificação automática de emoções. Portanto, neste trabalho foi utilizada esta técnica de reconhecimento de emoções através da fala, e assim foi desenvolvida uma arquitetura de rede neural convolucional que foi treinada e testada com os dados da base EmoUERJ. A arquitetura foi capaz de apresentar resultados robustos tanto na fase de treinamento quanto na fase de teste, com 93,73% e 95% de acurácia respectivamente, apesar de ter sido um projeto de pequena escala, com uma quantidade limitada de dados. Ainda foi realizado um experimento com áudios extraídos de vídeos do Youtube, a fim de testar a performance do modelo na classificação de áudios brutos de fora do contexto de treinamento, que apresentou no final do experimento uma acurácia de 65%, um resultado muito interessante vista a aplicação mais realista do modelo.

Palavras-chave: Redes Neurais Convolucionais; Reconhecimento de emoções; MFCC.

ABSTRACT

With the increasing wave of technological development in the Artificial Intelligence field, many other areas of knowledge have also evolved and benefited from that evolution, among them is the affective computing area. In order to make users' interactions with computational systems more fluid and natural, affective computing research has been conducted in order to provide the machines with the ability to identify, interpret and synthesize emotions. However, this task of automatic emotion classification deals with many challenges, among them the ability to develop a model that presents good results and can generalize well to the classification task, as well as the limited amount of available databases for training the machine learning models. The literature presents convolutional neural networks trained on MFCC features extracted from speech audio, which show robust results in emotion classification. Thus, in this paper, this technique of speech emotion recognition was utilized, and an architecture of a convolutional neural network was developed, which was then trained and tested with data from the EmoUERJ database. The architecture was able to achieve robust results in both the training and testing phases, with 93.73% and 95% accuracy, respectively, despite being a small-scale project with a limited amount of data. An experiment was also performed using audio extracted from YouTube videos to test the model's performance in classifying raw audio from outside the training context. The experiment ultimately achieved 65% accuracy, which is a very interesting result considering the more realistic application of the model.

Keywords: Convolutional Neural Networks; Emotion Recognition; MFCC.

LISTA DE ILUSTRAÇÕES

Figura 1 – Representação de neurônio.....25
Figura 2 – Representação de neurônio artificial.....25
Figura 3 – Representação de uma RNA com camadas executadas em diferentes kernels de uma GPU.....26
Figura 4 – Procedimento de convolução.....28
Figura 5 – Representação de uma camada totalmente conectada.....30
Gráfico 1 – Função Sigmoid.....31
Gráfico 2 – Função Tanh.....32
Gráfico 3 – Função ReLU.....33
Figura 6 – Exemplo de matriz de confusão.....34
Figura 7 – Representação do fluxo do processo de classificação das emoções.....38
Figura 8 – Jupyter notebook com código Hello World!.....39
Gŕafico 4 – Quantidade de áudios representando cada emoção.....40
Figura 9 – MFCCs de diferentes emoções.....41
Figura 10 – Código da CNN.....43
Figura 11 – Arquitetura da CNN.....44
Gráfico 5 – Acurácia durante o treinamento.....46
Gráfico 6 – Perda durante o treinamento.....46
Figura 12 – Matriz de confusão fase treinamento.....47
Gráfico 7 – Precisão fase treinamento.....48
Gráfico 8 – Sensibilidade fase treinamento.....49
Gráfico 9 – F1-Score fase treinamento.....49
Figura 13 – Matriz de confusão teste.....50
Gráfico 10 – Precisão fase teste.....51
Gráfico 11 – Sensibilidade fase teste.....52
Gráfico 12 – F1-Score fase teste.....53
Figura 14 – Matriz de confusão experimento antes do treino.....54
Figura 15 – Matriz de confusão experimento depois do treino.....55

LISTA DE ABREVIATURAS E SIGLAS

BFCC	Bark Frequency Cepstral Coefficients
BN	Bayesian Networks
CNN	Convolutional Neural Network
DL	Deep Learning
GMM	Gaussian Mixture Model
GTCC	Gammatone Cepstral Coefficients
HMM	Hidden Markov Models
IA	Inteligência Artificial
KNN	K-nearest Neighbors
LPCC	Linear Predictive Cepstral Coefficients
MFCC	Mel Frequency Cepstral Coefficients
ML	Machine Learning
MLP	Maximum Likelihood Principle
RNA	Rede Neural Artificial
SVM	Support Vector Machines

SUMÁRIO

1 INTRODUÇÃO..... 11
1.1 PROBLEMÁTICA..... 13
1.2 OBJETIVOS..... 14
1.2.1 OBJETIVOS GERAIS..... 14
1.2.2 OBJETIVOS ESPECÍFICOS..... 14
1.3 JUSTIFICATIVA..... 15
1.4 ORGANIZAÇÃO DO DOCUMENTO..... 16
2 REFERENCIAL TEÓRICO..... 17
2.1 Emoções..... 17
2.2 Processamento Digital de Sinais..... 18
2.3 Processamento da Fala..... 18
2.4 MFCC..... 20
2.5 Machine Learning..... 20
2.6 Deep Learning..... 23
2.7 Redes Neurais Artificiais..... 24
2.8 Redes Neurais Convolucionais..... 27
2.8.1 Camada de convolução..... 27
2.8.2 Camada de pooling..... 29
2.8.3 Camada totalmente conectada..... 29
2.8.4 Funções de ativação..... 30
2.8.5 Batch Normalization..... 33
2.8.6 Dropout..... 33
2.9 Métodos de avaliação..... 34
2.10 Trabalhos Relacionados..... 35
3 METODOLOGIA..... 38
3.1 Ambiente de desenvolvimento..... 38
3.2 Base de dados utilizada..... 39
3.2 Extração de MFCC..... 40
3.3 Modelo de Rede Neural Convolucional (CNN)..... 41
3.3.1 Arquitetura..... 41
3.3.2 Otimização (Fine Tuning)..... 42
3.3.3 Treinamento..... 45
4 RESULTADOS..... 45
4.1 Fase de treinamento..... 45
4.2 Fase de teste..... 50
4.3 Experimentos..... 53
5 CONCLUSÃO..... 56
REFERÊNCIAS..... 58

1 INTRODUÇÃO

Com a crescente quantidade no volume de dados e a necessidade de se realizar o armazenamento e o processamento desses dados a fim de se obter informações, o que é uma tarefa difícil e um tanto custosa em termos de complexidade (Peixoto e Linhares, 2023), os desenvolvimentos obtidos no campo da Inteligência Artificial (IA) em geral, especialmente no subcampo de *Machine Learning* (ML), do inglês, aprendizado de máquina.

A ML é definida como um conjunto de técnicas e ferramentas que tem como finalidade fazer com que os computadores possam extrair padrões dessas massivas quantidades de dados e possam aprender as relações entre as informações presentes nos dados, sem que tenham sido propriamente programados para aprender os padrões e relações em questão. Essas técnicas têm sido cruciais para aplicações que lidam com tarefas complexas nas mais diversas áreas do conhecimento humano (Santos, 2022).

Dentre as aplicações de classificação e extração de padrões de forma automática, está o reconhecimento de emoções dos indivíduos. Esta tarefa de reconhecimento foi proporcionada pelo desenvolvimento e aprimoramento de pesquisas na área de conhecimento da Computação Afetiva, que apresentam estudos relacionados às emoções humanas focados na identificação, interpretação ou síntese de emoções, como indica Silva (2022).

Os progressos significativos que têm sido alcançados no decorrer dos últimos anos no campo da Inteligência Artificial, a tarefa de reconhecimento e classificação das reações emocionais dos indivíduos têm sido aprimoradas e vêm apresentando resultados cada vez mais precisos. (Silva, 2022).

Há um grande interesse no reconhecimento das emoções humanas de forma automática, tendo em vista a grande variedade de possibilidades de aplicações que podem ser exploradas como, por exemplo, aprimoração dos assistentes virtuais que, a partir da identificação da emoção atual do usuário, poderão responder às demandas deste de forma mais precisa e personalizada (Peixoto e Linhares, 2023).

Atualmente, existem várias técnicas de IA que estão sendo empregadas nessas tarefas de reconhecimento de emoções que se dão através de imagens, vídeos, ou mesmo da fala. Sendo exemplos destas técnicas os algoritmos de ML: *K-nearest Neighbors* (KNN), *Support Vector Machines* (SVM) e *Hidden Markov Models* (HMM).

A técnica KNN é um método utilizado em tarefas de classificação e de regressão, e por ser um método não paramétrico, é capaz de inferir relações mais complexas a partir dos dados fornecidos, uma vez que não fica limitado a atender as definições de parâmetros pré

definidos, como acontece nos métodos paramétricos. A partir dos vetores de treinamento, essa técnica identifica uma certa quantidade de pontos vizinhos mais próximos de um vetor de características desconhecido, cuja classe se deseja identificar.

A técnica SVM utiliza aprendizagem supervisionada e também pode ser aplicada nos problemas de classificação e regressão. Esta técnica consiste na plotagem de todos os itens presentes nos dados de treinamento em um espaço de n dimensões, sendo que a quantidade de dimensões depende da quantidade de características de cada item presente nos dados, para então encontrar uma linha (hiperplano) que separe esses pontos de acordo com as diferentes classes de itens presentes nos dados de treinamento.

O *Hidden Markov Models* (HMM) é um modelo gráfico probabilístico capaz de descrever a relação probabilística entre dois processo estocásticos interdependentes, o processo observável e o processo oculto de Markov, podendo fazer predições de observações futuras ou classificações de sequências a partir do processo oculto de Markov responsável por gerar os dados visíveis (Alzubi; Nayyar; Kumar, 2018; Pan et al., 2024).

Porém, as Convolutional Neural Networks (CNN), do inglês, Redes Neurais Convolucionais, tipo de Rede Neural Artificial (RNA) com múltiplas camadas, são os modelos de aprendizagem de máquina que têm apresentado os melhores resultados nas tarefas de classificação, visto que possuem camadas especializadas na extração de características por meio de filtros de convolução (Santos, 2022).

As CNNs podem aprender características relevantes associadas às emoções humanas através do espectrograma de áudios de fala (Silva, 2022). Assim, algumas técnicas de extração de características de emoções a partir de áudios têm sido associadas com as redes neurais convolucionais, e dentre essas técnicas, a de *Mel Frequency Cepstral Coefficients* (MFCC) vem apresentando bons resultados quando aplicada individualmente, como mostra Hilleshein (2018).

Esta técnica se baseia na audição humana para adquirir as características de um sinal de áudio, sendo a representação da resposta feita utilizando a escala de frequência *mel*. A escala *mel* foi desenvolvida para dimensionar as frequências para que correspondam à forma que o ouvido humano é capaz de perceber e interpretar essas frequências, fazendo com que os coeficientes *mel* estejam concentrados apenas na faixa de frequência que os humanos conseguem ouvir (Hilleshein, 2018; Badr; Mukherjee; Thumati, 2021).

Assim, essas características extraídas a partir de sinais de áudio contendo a fala das pessoas são utilizadas como entradas para que as redes neurais convolucionais possam extrair

padrões e criar as classes que servem para distinguir as diferentes características que representam as emoções quando são expressadas na fala.

E nesse trabalho será feito o uso dessa técnica de aprendizagem de máquina, a CNN, para classificar as emoções através da fala, utilizando como entrada as *features* MFCC, que representam as características dos áudios transformados em espectrogramas.

1.1 PROBLEMÁTICA

A área de conhecimento da Computação Afetiva que, segundo Mezencio (2022), pode ser definida como o campo de estudo no qual sistemas computacionais estão relacionados com emoções, ou ainda, quando ocorre o surgimento de emoções enquanto se manipula artefatos computacionais. Ela é considerada uma importante área para o desenvolvimento geral da inteligência artificial, visto que esta se baseia no processo de aprendizagem humana que é relacionada diretamente às emoções, aponta Silva (2022).

Modelos estatísticos e algoritmos de aprendizagem de máquina têm sido amplamente utilizados, desde o final do século XX, na tarefa de reconhecer emoções através da fala automaticamente (Peixoto e Linhares, 2023). Sendo os principais classificadores divididos nas classes de classificadores lineares e não lineares.

O Bayesian Networks (BN), o *Maximum Likelihood Principle* (MLP) e o *Support Vector Machine* (SVM), são exemplos de classificadores lineares que são muito utilizados para realizar a classificação de emoções. Enquanto os classificadores não lineares mais utilizados são o *Gaussian Mixture Model* (GMM) e o *Hidden Markov Model* (HMM), destaca Khalil et al. (2019).

Porém, pesquisas no campo da Aprendizagem Profunda, do inglês, *Deep Learning* (DL), têm mostrado que em tarefas de identificação de estruturas e características complexas e de difícil classificação os modelos de DL apresentam melhores resultados que os métodos tradicionais (Khalil et al., 2019)**.**

No campo de DL, as CNNs têm demonstrado resultados promissores na área de visão computacional, também no processamento de Linguagem Natural, porém, tem-se igualmente explorado essa arquitetura, com as devidas adaptações, em tarefas de reconhecimento de emoções através da fala (Gomes Junior, 2019).

Para que a aplicação da CNN seja utilizada no reconhecimento das emoções presentes na fala, é necessário que se represente esses sinais de fala em formato de imagem, que são obtidas no pré-processamento dos áudios utilizando extratores de características de sinais

sonoros (Gomes Junior, 2019), como os extratores *Mel Frequency Cepstral Coefficients* (MFCC**)**, os *Bark Frequency Cepstral Coefficients* (BFCC), os *Gammatone Cepstral Coefficients* (GTCC), e os *Linear Predictive Cepstral Coefficients* (LPCC), sendo os mais comuns (Santos e Reis, 2020).

Com base nos resultados encontrados na literatura (Hilleshein, 2018), e, observando os resultados dos testes realizados (Santos e Reis, 2020), avaliam que os extratores MFCC têm apresentado melhores resultados em comparação com os outros métodos quando aplicados individualmente. Alguns dos modelos desenvolvidos e apresentados na literatura utilizam modelos "híbridos" usando mais de um classificador e/ou mais de um extrator de características como o exemplo do modelo DEEP apresentado por Campos e Moutinho (2020).

Observando o atual contexto das aplicações de identificação de emoções através da fala e as diferentes abordagens utilizadas para desempenhar essa incubência, surge o questionamento: é possível obter bons resultados na tarefa de classificação de emoções através da fala utilizando um modelo de CNN simples empregando a técnica de extração de características MFCC obtidas a partir de um pequeno conjunto de dados de áudio?

1.2 OBJETIVOS

1.2.1 OBJETIVOS GERAIS

- Utilizar uma rede neural convolucional, em conjunto com a técnica de extração de características do sinal de voz MFCC, para realizar o reconhecimento de emoções humanas através da fala.

1.2.2 OBJETIVOS ESPECÍFICOS

- Investigar técnicas de processamento de voz;
- Selecionar uma base de dados com áudios representando diferentes emoções para treinamento e validação do modelo de rede neural;
- Desenvolver e treinar um modelo de rede neural convolucional para o reconhecimento de emoções, utilizando a técnica de extração de características MFCC;
- Avaliar o desempenho do modelo, validando sua acurácia e realizando comparações entre os resultados obtidos nas diferentes emoções reconhecidas.

1.3 JUSTIFICATIVA

Com a grande evolução tecnológica que vem acontecendo nas últimas décadas, a citar a evolução da IA, que vem proporcionando melhorias nos sistemas automatizados que auxiliam e facilitam a vida das pessoas com a automação de tarefas (Peixoto e Linhares, 2023), utilizando novos paradigmas de interação homem-máquina que buscam fazer com que as interações entre as pessoas e as máquinas sejam mais naturais e eficientes, levando em consideração a perspectiva emocional das pessoas nos diálogos (Suguino et al. 2022).

Assim, pesquisas na área da computação afetiva vêm sendo realizadas nas últimas décadas, visando a detecção automática de emoções, argumenta Suguino et al. (2022), que pode ser realizada através de expressões faciais, postura, fala, gestos, dados fisiológicos, sinais psicológicos, entre outros, exemplifica Libralon (2014). Contudo, a detecção através da fala tem se tornando muito comum, visto que a fala além de informações semânticas (Peixoto e Linhares, 2023), também evidencia informações sobre as emoções do falante.

Porém, ainda se tem muitos desafios no que diz respeito a interpretar, computacionalmente, as informações sobre as emoções que são expressadas pelos interlocutores durante as interações a fim de que a máquina reaja de acordo com o contexto da interação (Libralon, 2014).

Dentre os fatores que dificultam essa tarefa, Ferreira et al. (2022) citam o fato que cada palavra de um trecho que o usuário fala pode conter uma emoção distinta, bem como que cada pessoa pode expressar suas emoções de forma única de acordo com a sua personalidade, e ainda, em casos em que a fala contém um sotaque local específico que pode interferir na identificação das emoções.

Os estudos e pesquisas são desenvolvidos com o intuito de amadurecer os sistemas de reconhecimento de emoções para a aplicação nos mais diversos contextos de interação, como na detecção de emoções de usuário de telefones celulares, em call centers, elenca Peixoto e Linhares (2023).

Estes sistemas também podem ser utilizados em serviços prestados ao cliente de auto atendimento a fim de avaliar a sua satisfação, em sistemas de tutoria para avaliar a receptividade dos alunos aos conteúdos, na área de entretenimento indicando conteúdos que estejam compatíveis com a emoção dos usuários, acrescenta Libralon (2014), e ainda sua aplicação em estudos clínicos psiquiátricos, psicológicos e neurofisiológicos como o apresentado por Moraes (2020) no seu trabalho de detecção depressão através da fala.

Portanto, vista a vasta aplicabilidade bem como importância de tal método, buscará-se desenvolver e avaliar tal arquitetura de identificação de emoções através da fala, utilizando CNN e MFCC, a fim de poder contribuir com esse campo de pesquisa, que em muito poderá contribuir para a sociedade nos mais diversos problemas e aplicações cabíveis.

1.4 ORGANIZAÇÃO DO DOCUMENTO

Este trabalho é composto por esta seção introdutória, onde é feita uma breve explanação sobre o tema além de serem elencados os objetivos gerais e específicos, bem como, é apresentada a problemática da pesquisa e em seguida a sua justificativa. Também conta com a seção que explora o referencial teórico, que abarca as informações fundamentais que servirão de base para o desenvolvimento da pesquisa e do modelo de rede neural convolucional proposto, responsável por fazer o reconhecimento das emoções através da fala.

É apresentada, ainda, a seção que trata da metodologia utilizada para a conduzir o desenvolvimento da parte prática do trabalho. Em seguida, há o capítulo no qual serão explorados os resultados obtidos com o desenvolvimento do projeto proposto. Logo após, tem-se a seção de conclusão do trabalho que aborda uma visão geral dos resultados obtidos, bem como, traz considerações gerais sobre o desenvolvimento do trabalho.

2 REFERENCIAL TEÓRICO

2.1 Emoções

As emoções humanas são objeto de estudo de variadas áreas do conhecimento como filosofia, biologia, psicologia, psiquiatria, neurociência, entre outras (Silva, 2022). E apesar de não ter uma única definição concreta e objetiva, as emoções, geralmente, são tidas como um impulso neural, que pode ser resultante de uma interação com o mundo externo, ou de alguma antecipação de ações e comportamentos, ou ainda de mecanismos de defesa, que move um organismo a realizar determinada ação (Libralon, 2014).

Segundo Silva (2022), há três visões mais conhecidas que se debruçam sobre o estudo das emoções e tentam explicá-las, a Darwinista, a Jamesiana e a Cognitiva. Charles Darwin, nos seus estudos sobre evolução das espécies, introduz a perspectiva darwinista em relação às emoções e argumenta que as emoções ajudam os seres humanos na tarefa da manutenção da sobrevivência, e que, portanto, possuem um importante papel no processo evolutivo.

Já William James, na sua teoria jamesiana que versa sobre as emoções, diz que as emoções são produtos de determinados estímulos que causam reações fisiológicas, que resultam nas experiências conscientes dessas reações, ou seja, as emoções são, justamente, as respostas a essas reações, assim, sem que haja estímulos fisiológicos, consequentemente, não há emoções.

Enquanto a teoria cognitiva, argumenta que as emoções dependem da avaliação não consciente dos estímulos recebidos e portanto são inerentes, sendo a emoção constituída pela interpretação do estímulo que por conseguinte é associada a uma causa, existindo então um processo cognitivo na constituição das emoções.

O reconhecimento das emoções têm sido um assunto um tanto estudado dada a sua importância e potenciais aplicações, resultando na criação de classificadores automáticos de emoções utilizando diferentes abordagens para realizar o processo de reconhecimento, dentre as abordagens estão a classificação por meio de imagens, a partir da análise de expressões faciais, postura, ou mesmo gestos, bem como através de sons, como por meio da fala da pessoa, além do uso dos sinais vitais e neurológicos como fontes (Silva, 2022).

Assim, com a busca pela compreensão das emoções, as máquinas tentam entender as informações contextuais sobre os indivíduos e seus ambientes a fim de poder ter uma conclusão apropriada, o que é uma tarefa muito complexa, vista a diversidade dos contextos dos quais surgem as emoções (Libralon, 2014).

Os métodos de classificação de emoções por meio da fala realizados pelos computadores, hoje em dia, apresentam uma taxa de até, em média, 80 por cento de acerto (Peixoto e Linhares, 2023), contra a média de 65 por cento das classificações feitas pelos seres humanos, o que apresenta um equilíbrio e um resultado aceitável dos resultados dos modelos computacionais (Libralon, 2014).

Os modelos computacionais, no entanto, não são capazes de analisar os dados da fala, os áudios que a contém, de forma bruta, sendo necessário realizar a extração de características acústicas e/ou espectrais presentes nesses áudios que podem ser utilizadas para diferenciar padrões existentes na manifestação de diferentes emoções (Santos, 2022), e assim realizar a classificação.

No entanto, para que essa extração aconteça, os áudios devem passar por um processo de transformação para que sejam representados na forma digital, como uma sequência de bits, permitindo que os computadores possam processar esses sinais, e esse processo de transformação se dá no processamento de sinais digitais (Nunes, 2021).

2.2 Processamento Digital de Sinais

Calheiro (2021) define um sinal digital como sendo uma quantidade que varia em um determinado tempo. Os sinais digitais representam os sinais analógicos através de sequências de números quantizados que são formadas por amostras do sinal analógico capturadas regularmente com uma certa periodicidade (Silva, 2022).

Para que os sinais analógicos possam ser computados e armazenados por computadores, necessitam de serem convertidos para a sinais digitais, e esse processo é feito por um conversor analógico-digital, que fornece ao computador a sequência numérica discreta no tempo. E para que essa conversão aconteça são realizados os processos de amostragem e quantização (Calheiro, 2021).

No processo de amostragem do sinal analógico, nesse caso, do áudio de fala que se deseja processar, este é representado como uma sequência de valores amostrais retirados periodicamente do sinal analógico, sendo esses valores, números que representam o valor da onda acústica medida em determinado ponto específico no tempo, e então esses valores de onda são aproximados, quantizados, para o patamar mais próximo na escala de amplitude que está sendo utilizada, sendo esse o processo de quantização (Nunes, 2021).

2.3 Processamento da Fala

A voz é produzida por vibrações que ocorrem com as vibrações das cordas vocais. O ar vindo dos pulmões passa pela traqueia, então segue pela laringe, onde o ar é modulado pelas cordas vocais formando ondas acústicas quase periódicas, características da produção vocálica que apresenta ondas regulares mas não perfeitamente periódicas devido a diversos aspectos do falante como fisiológicos, emocionais, psicológicos, entre outros; então essas ondas acústicas entram na cavidade oral e são modificadas ao passarem pela boca e pelo nariz (Silva, 2022).

Esse sinal de voz produzido pelos seres humanos é analógico, pontua (Silva, 2006), e para que possa ser processado pelos computadores, precisam ser convertidos do analógico para o digital, conclui. Os sinais de voz são capturados por um microfone que converte as variações na pressão do ar, que são causadas pela fala humana, em variações de tensão elétrica, aponta (Silva, 2006).

Após esse processo de amostragem, é realizada a quantização do sinal amostrado que trata da discretização da intensidade do sinal para permitir a sua representação por uma quantidade finita de bits, e quanto maior a quantidade de bits mais fiel ao original é o sinal quantizado, e assim se obtém um sinal digital (Silva, 2006).

Para realizar o pré-processamento dos sinais de voz, geralmente se faz o processamento desses sinais em segmentos de duração entre 20 e 40 milissegundos, chamados de janelas de áudios, pelo fato de que esses sinais variam constantemente (Silva, 2022), assim podem ser extraídas as informações de cada segmento individualmente, conclui.

Então, com a conversão dos sinais de voz para sinais digitais, e da realização do janelamento dos áudios, segue-se para a extração das características dos sinais de voz. Dentre as técnicas de extração de características de voz a quem tem sido muito utilizada e apresenta uma boa eficiência (Hilleshein, 2018), é a dos coeficientes mel-cepstrais, do inglês *mel-frequency cepstrum coefficients* (MFCC).

A técnica de MFCC se baseia na audição humana, que possui uma percepção linear das frequências até 1000 Hz, e, acima disso, a percepção passa a ser logarítmica, e para aproximar-se dessa percepção foi criada a escala *mel* (Mendoza, 2009). Então, a extração das características são realizadas tendo como entrada um sinal de áudio no domínio do tempo e resulta nos coeficientes no domínio cepstral mel, explica (Hilleshein, 2018), que serão utilizados para o treinamento dos modelos de aprendizagem de máquina.

2.4 MFCC

Os Coeficientes cepstrais de frequência Mel (MFCC, a sigla em inglês), são recursos de parâmetros que estão sendo fortemente utilizados em aplicações que envolvem nas atividades de reconhecimento de voz, e áreas afins (Helali; Hajaiej; Cherif, 2024),bem como para o tratamento de áudio para o reconhecimento de emoções através da fala.

Os MFCCs são características de campos de frequências baseados na escala da audição humana, essas oferecem mais acurácia do que os recursos de domínio de frequência baseados no tempo, porque podem apresentar a mesma quantidade de informações em menos coeficientes, se tornando mais compactos, segundo (Helali; Hajaiej; Cherif, 2024).

A extração do MFCC se dá nas seguintes etapas: é realizada a pré-ênfase dos sinais de fala para amplificar as frequências altas; então são realizados processos de enquadramento e janelamento; após é aplicada a transformada discreta de *Fourier* (DTF) que é calculada para cada janela do espectro; também é realizado um filtro de espectro de potência com uma matriz do banco de filtros de frequência mel; em seguida é aplicada uma função logarítmica para cada elemento da matriz; e por fim é calculada a transformada discreta de cosseno na matriz (Vreča; Pilipović; Biasizzo, 2024).

2.5 Machine Learning

A fim de proporcionar a capacidade de executar as mais diversas tarefas que o ser humano é capaz de desempenhar aos computadores de uma forma ainda mais eficiente, o campo de ML tem sido amplamente fomentado (Taye, 2023). Dentre tais atividades estão: detecção de fraudes, sistemas de recomendação baseados no perfil dos usuários, negociação nos mercados financeiros, análise de risco e de crédito, detecção de objetos, reconhecimento de pessoas pelas mais características pessoais, diagnósticos médicos, entre muitas outras.

As Aplicações de ML proporcionam às máquinas a capacidade de estarem constantemente aprendendo de forma adaptável e profunda no intuito de conseguir aumentar progressivamente a sua experiência e poder entender toda a complexidade do problema ao qual se dedica, com poucas intervenções humanas sendo necessárias (Alzubi; Nayyar; Kumar, 2018). Existem diferentes tipos de problemas, que são classificados como:

- Problema de classificação:
  Nesse tipo de problema, o algoritmo de aprendizagem de máquina só pode apresentar uma entre um número limitado de saídas possíveis dadas para o

problema ao qual o algoritmo propõe classificar. O problema de classificação pode ser categorizado como um problema de classificação binária, que é quando o algoritmo vai ser responsável por avaliar os dados e fornecer apenas uma classificação entre duas opções de classes possíveis, e pode ainda ser categorizado em um problema de classificação multiclasses, que diz respeito à classificação de um determinado conjunto de dados em uma classe dentre um número maior que duas classes.

- Problema de detecção de anomalia:
  Quanto tratando desse tipo de problema, o algoritmo de aprendizagem de máquina vai ser responsável analisar e detectar mudanças ou anomalias que possam se mostrar dentro do conjunto de dados que fujam dos padrões anteriormente detectados, o que acontece quando modelos de dados vão além ou divergem dos modelos de dados que são usualmente apresentados.

- Problema de regressão:
  Esse tipo de problema, geralmente, busca fazer predições, projeções ou estimativas baseadas em um conjunto de dados que possam apresentar padrões que podem ser úteis para realizar mensurações de valores que possam vir a se concretizar posteriormente. Esse tipo de problema é visto, por exemplo, no mercado financeiro onde tenta-se prever os preços futuros dos ativos baseando-se em uma série de diferentes informações que podem estar relacionados a diversos outros fatores passíveis de interferir nos preços dos ativos correlacionados, portanto analisa-se esses dados e utiliza-se os algoritmos de aprendizagem de máquina para realizar a predição dos preços futuros desses ativos a partir dos padrões presentes nos dados.

Vistos os diferentes tipos de problemas, os quais os algoritmos de ML buscam solucionar, existem diferentes formas de treinar, diferentes abordagens utilizadas no processo de aprendizagem desses modelos de IA para que eles sejam capazes de encontrar soluções para determinado problema a partir dos dados que lhe são apresentados. Dentre as diferentes categorias de aprendizagem de ML, as principais são:

- Aprendizagem supervisionada:

  Nessa categoria de aprendizagem, os algoritmos de ML são treinados a partir de dados rotulados, consistindo nos dados que serão fornecidos ao algoritmo juntamente com a saída desejada na classificação a partir dos dados fornecidos. Portanto, nesse processo os resultados desejados já são conhecidos, e para alcançar tal resultado a algoritmos de aprendizagem supervisionada usam funções de classificação quando as saídas desejadas são saídas discretas e funções de regressão quando as saídas são contínuas.

  A detecção de objetos a partir de imagens ou vídeos, é um exemplo de um problema de classificação no qual a aprendizagem supervisionada é aplicada. Já no de problema de regressão, por exemplo, utiliza-se os algoritmos de ML para prever preços de ativos financeiros a partir de uma série de dados distintos que podem estar relacionados com os preços desses.

- Aprendizagem não supervisionada:

  Para realizar o treinamento de algoritmos que implementados para esse tipo de aprendizagem são necessárias grandes quantidades de dados, vistos que os dados fornecidos ao algoritmo, diferentemente da aprendizagem supervisionada, não são rotulados, e cabe ao algoritmo encontrar os padrões que consegue inferir a partir dos dados fornecidos. Para realizar tais inferências os algoritmos tendem a reconhecer certas características presentes nos dados que difere uma ocorrência da outra, fazendo com que ele seja capaz de criar categorias e associar os diferentes tipos de ocorrências presentes nos dados aos diferentes tipos de categorias criadas.

- Aprendizagem semi supervisionada:

  Visto que a disponibilidade de conjuntos de dados que estejam devidamente rotulados é bem limitada e tendem a ter um custo um tanto elevado, dada a necessidade de despender mais esforço a fim de analisar ‘manualmente’ os dados a fim de categorizá-los, algumas aplicações de aprendizagem de máquina usam uma quantidade menor de dados rotulados juntamente com uma quantidade consideravelmente maior de dados não rotulados, sendo estes menos custosos financeiramente e mais passíveis de serem conseguidos sem maiores esforços. Assim, geralmente são feitas fusões dos dois tipos de modelos, usando as técnicas aplicadas em cada um dos tipo de aprendizagem, a fim de realizar o treinamento do algoritmo.

- Aprendizagem por reforço:

  A técnica de aprendizagem por reforço geralmente é aplicada na resolução de problemas há poucas ou inconsistentes informações e em situações onde o ambiente do problema é considerado volátil, sendo necessário que o julgamento do sistema de aprendizagem seja constantemente atualizado. Para isso, os algoritmos que utilizam esse método de aprendizagem são treinados de forma a aprenderem de acordo diretamente com as suas ações e como elas afetam o ambiente no qual está inserido. Assim, é definido um objetivo para o algoritmo, ao passo que o sistema realiza uma ação ele atualiza as informações que mudaram no ambiente por meio de suas ações e avalia se estão mais pertos ou longe do seu objetivo, e assim é recompensado ou punido, em forma de pontos, de acordo com o sucesso ou falha de suas ações. Nesse processo de tentativa e erro, o algoritmo aprende ao passo que avalia o resultado que as suas ações causam no ambiente e pondera se tal ação foi benéfica no alcance do seu objetivo.

Há problemas a serem resolvidos que são mais complexos e demandam um maior nível de abstração durante o processamento dos dados, envolvendo também uma maior quantidade de características a serem inferidos a partir desses dados, fazendo com que os modelos comuns de ML não atendam à necessidade de uma resolução aceitável do problema.

Por isso, muito tem-se estudado nas últimas décadas a respeito do Deep Learning (DL), do inglês, Aprendizagem Profunda, que apresenta modelos com uma maior quantidade de camadas de aprendizagem, conseguindo realizar um maior nível de abstração do que os demais modelos de ML, conseguindo lidar com uma quantidade maior de dados, inclusive tende a apresentar uma melhor performance à medida que a quantidade de dados fornecidos ao modelo aumenta (Taye, 2023).

2.6 Deep Learning

Um dos campos de Machine Learning que tem se provado muito bem sucedido em diversos tipos de problemas, tanto de regressão como de classificação, a serem solucionados com o auxílio da IA, é o campo de Deep Learning. Como o nome sugere, é capaz de proporcionar às máquinas a capacidade de performar uma aprendizagem mais específica para cada problema, se adaptando dinamicamente aos diferentes desafios e tipos de dados (Ottoni; Ottoni; Cerqueira, 2023).

Os métodos de DL se baseiam fundamentalmente nas redes neurais artificiais, que são formadas por várias pequenas unidades de processamento que se conectam umas às outras, conhecidos como neurônios artificiais. As redes neurais artificiais simulam o funcionamento do cérebro humano no que diz respeito à forma como ele lida com os dados que lhe são apresentados.

As redes neurais artificiais são capaz de autonomamente aprender características, juntamente com as suas representações hierárquicas, em múltiplos níveis de complexidade, sem que seja necessárias intervenções humanas para definir diretamente as regras para modelo, diferentemente de como acontece quando tratando de algoritmos de ML tradicionais (Taye, 2023).

Alguns dos problemas do mundo real nos quais são aplicados modelos de DL são: no processamento de imagens, processamento de linguagem natural, análise de séries temporais, jogos, geração de imagens, análise de emoções, reconhecimento de voz , entre muitas outras (Alzubi; Nayyar; Kumar, 2018). Para que um modelo de DL seja performático, geralmente, é necessário que este seja exposto a grandes quantidades de dados para que possam conseguir generalizar bem o problema ao que se propõe solucionar (Taye, 2023).

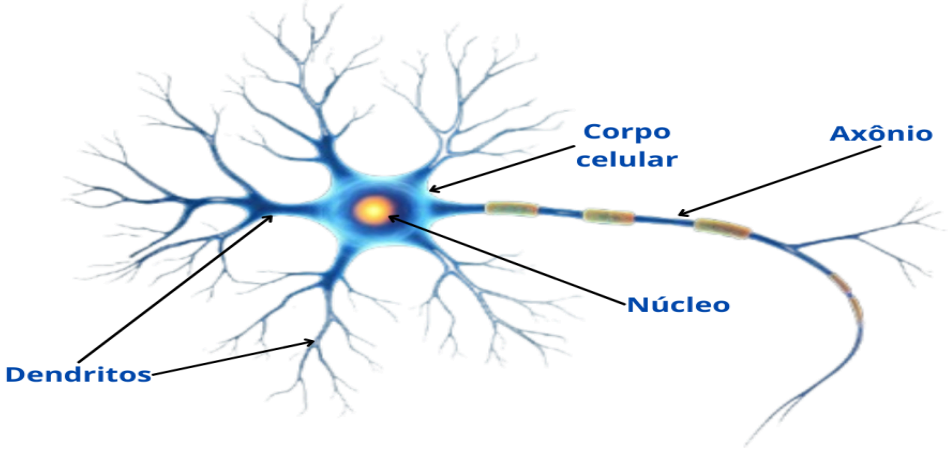
2.7 Redes Neurais Artificiais

As redes neurais artificiais foram desenvolvidas para simular computacionalmente a estrutura e o funcionamento do cérebro. Assim, os neurônios do cérebro (FIGURA 1) são representados pelas unidades de processamento das redes neurais (FIGURA 2), também chamadas de neurônios artificiais, que são responsáveis pelo processamento das informações que são recebidas pelas camadas de entrada da rede (Mendoza, 2009).

Já os dendritos, que são responsáveis por receber os sinais elétricos vindos de outros neurônios, são representados como as interconexões entre as unidades de processamento das redes neurais. Enquanto as sinapses cerebrais, que consistem no ponto de contato entre um axônio de um neurônio e o dendrito de outro, são representadas como os pesos presentes nos neurônios das redes neurais (Mendoza, 2009).

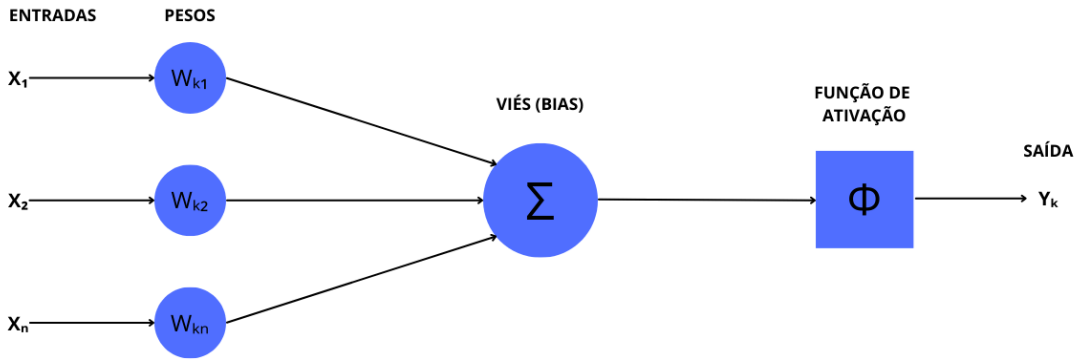
Por fim, os axônios, que são as terminações dos neurônios do cérebro que transferem o sinal de um neurônio para outro, são representados pelos terminais de saída dos neurônios artificiais (Mendoza, 2009).

Figura 1 – Representação de neurônio



Fonte: Imagem gerada pelo Gemini e editada pelo autor (2024).

Figura 2 – Representação de neurônio artificial



Fonte: Elaborado pelo autor, baseado no modelo de McCulloch-Pitts (2024).

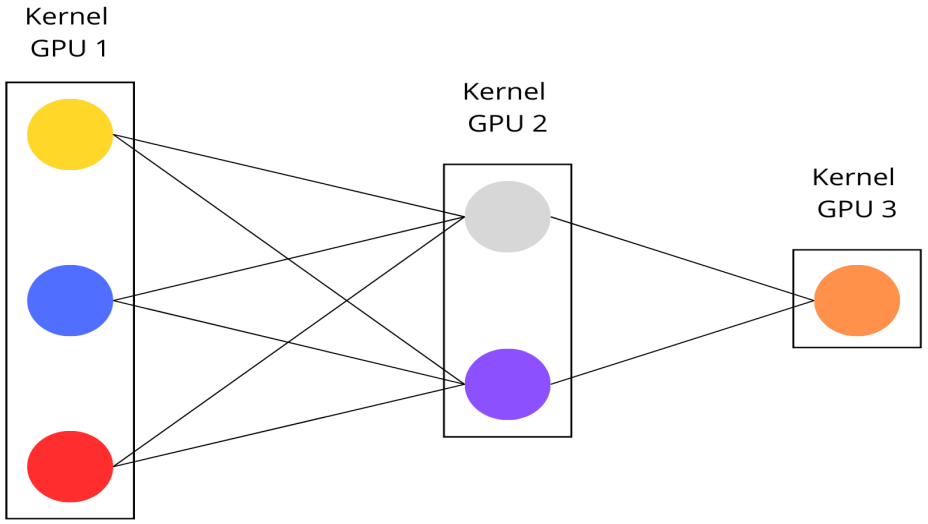
A estrutura mais conhecida das redes neurais artificiais possuem uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída (Silva, 2022). Algumas das propriedades das redes neurais artificiais são a não-linearidade, a adaptabilidade, a tolerância a falhas e dados ruidosos, a capacidade de generalização e apresentam o paralelismo.

A não-linearidade faz com que as redes neurais possam operar funções complexas não lineares de transformação de dados. Já a adaptabilidade proporciona às redes neurais a capacidade de poder adaptar os seus pesos de acordo com as variações ambientais no qual está inserido.

As redes neurais ainda são tolerantes a falhas e dados ruidosos, que faz com que as redes neurais sejam consideradas robustas. Também possuem a capacidade de generalização, que faz com que as redes sejam capazes de generalizar novos padrões e não apenas memorizar os dados treinados. E, por fim, as redes neurais artificiais possuem a propriedade de paralelismo, que é ideal para as implementações destas redes em computadores que suportam

o processamento paralelo (FIGURA 3), para se obter uma melhor performance (Mendoza, 2009).

Figura 3 – Representação de uma RNA com camadas executadas em diferentes kernels de uma GPU



Fonte: Elaborado pelo autor (2024).

Os neurônios das redes neurais artificiais são as unidades de processamento simples, que estão altamente interconectadas, processando determinadas funções matemáticas, sendo que, na maioria dos modelos de redes, as conexões entre as camadas formadas pelos neurônios estão associadas a pesos que fazem o armazenamento do conhecimento representado no modelo, que é usado para realizar a ponderação da entrada recebida por cada neurônio da rede (Libralon, 2014).

E, finalmente, a saída do processamento realizado nos neurônios associados é transformada com uma função de ativação, sendo essa função escolhida de acordo com a natureza dos dados com os quais será realizado o treinamento da rede neural artificial e o tipo de problema, pontua Silva (2022).

Assim, a aprendizagem realizada pela a rede funciona a partir de uma sequência de exemplos que lhes são apresentados, sendo que, no processo de aprendizagem supervisionado, cada exemplo fornecido é composto por uma entrada e um rótulo que representa a classe correta associado a ela, descreve Silva (2022).

Enquanto que no processo de aprendizagem não supervisionada, não existe os rótulos de saída desejada, e a rede é treinada através de excitações ou padrões de entrada organizando-os, arbitrariamente, em categorias, assim, para cada entrada aplicada à rede, será fornecida uma resposta indicado a classe da entrada (Mendoza, 2009).

2.8 Redes Neurais Convolucionais

As redes neurais convolucionais, do inglês, *Convolutional Neural Networks* (CNN), nasceram a partir de estudos relacionados ao córtex visual do cérebro e têm sido aplicadas principalmente em tarefas de reconhecimento de imagens, não obstante, há muitas aplicações no reconhecimento de voz e processamento de linguagem natural (Calheiro, 2021).

A primeira CNN criada foi a LeNet desenvolvida em 1998, que era aplicada no reconhecimento de dígitos escritos à mão, sendo treinada utilizando uma base de dados contendo 60 mil imagens utilizadas para treinamento e outras 10 mil para teste, todas elas em escala de cinza de dimensões 28x28, apresentado um resultado melhor que as outras técnicas que eram empregadas na época (Moraes, 2020).

Todas as camadas das CNNs se conectam a camada seguinte, e para a realização dos ajustes dos pesos no decorrer do treinamento, segue-se a lógica do *back-propagation*, nela é dado uma amostra de entrada para obtenção da saída para que seja comparada com a saída desejada, e então é calculado o erro que é propagado da saída para a entrada e os pesos e o limiar são atualizados de acordo com a regra delta generalizada, que faz o uso do gradiente descendente, fazendo com que os erros sejam diminuídos gradativamente (Moraes, 2020).

A arquitetura básica das redes neurais convolucionais possuem três camadas: a camada convolucional, a camada de *pooling* e a camada totalmente conectada (Leite, 2022). Porém, podem apresentar camadas adicionais que podem fornecer algum tipo de processamento complementar à rede neural.

2.8.1 Camada de convolução

A camada de convolução tem como responsabilidades a identificação e a extração das características da imagem de entrada por meio de filtros convolucionais reduzidos e repassa para a próxima camada na forma de mapa de características (Leite, 2022). Os neurônios desta camada não se conectam a todos os pixel da imagem, se atêm apenas àqueles que estão presentes no seu campo de visão, fazendo com que as primeiras camadas tenham informações mais de baixo nível, enquanto as camadas seguintes lidam com informações de mais alto nível (Calheiro, 2021).

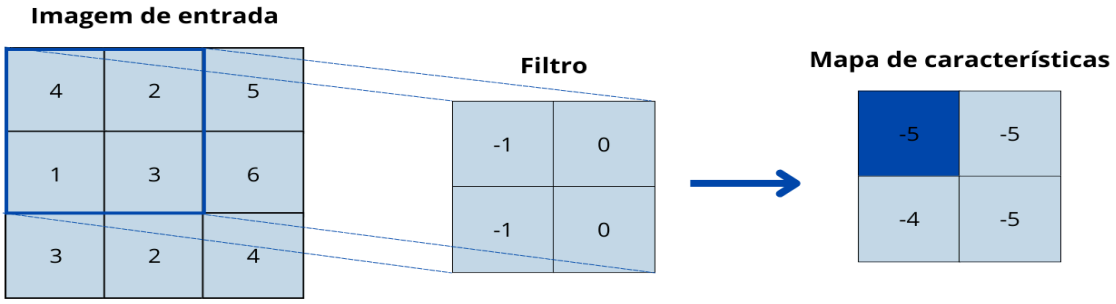
Os filtros convolucionais responsáveis pela extração de características dos dados dados como entradas à rede, são conhecidos também como núcleos convolucionais. Esses núcleos convolucionais apresentam três dimensões, a dimensão de largura, a de altura e a de

profundidade, que geralmente diz respeito aos filtros de cores das imagens, sendo os principais tamanhos utilizados para os núcleos sendo de 3 x 3 ou 5 x 5 (Zhao et al., 2024).

Os núcleos são formados por um conjunto de valores discretos que são usados para realizar o processo de convolução (FIGURA 4), que consiste na ação de deslocar o núcleo convolucional sobre os dados de entrada, e calcular o produto escalar, de acordo com a fórmula 1, de cada posição do núcleo com a respectiva posição dos dados de entrada produzindo um conjunto de mapas de características como saída da camada convolucional (Krichen, 2023).

$$a \bullet b = \sum_{i=1}^{n} a_i b_i \qquad (1)$$

Figura 4 – Procedimento de convolução



Fonte: Elaborado pelo autor (2024).

No processo de convolução, além das dimensões do filtro de convolução, há também alguns outros parâmetros que estão envolvidos no processo da extração de características que definem o tamanho e a resolução do mapa de características que é obtido como saída (Krichen, 2023). São eles:

- *Padding*: Que consiste no parâmetro que define a quantidade de linhas e colunas que serão preenchidas com valores nulos e adicionados ao redor dos dados de entrada antes que o processo de convolução seja iniciado. Usado para controlar o tamanho da saída do mapa de características, fazendo com que mais dimensões espaciais dos dados sejam preservados durante o processo de convolução.
- *Strides*: É o parâmetro que consiste na quantidade de linhas e colunas que o núcleo convolucional vai estar se sobrepondo nos dados de entrada, da esquerda para a direita de cima para baixo, referente à matriz dos dados de entrada. Quanto maior o valor do *stride*, maior a perda de resolução dos dados, visto que o valor de stride define a quantidade de linhas e colunas que o núcleo vai pular, e essas informações serão perdidas.

2.8.2 Camada de *pooling*

Similarmente a camada de convolução, a camada de pooling possui um núcleo que é sobreposto aos dados repassados pela camada anterior e esses dados que são abrangidos pelo núcleo participam do processo de *pooling (Zhao et al., 2024)*. Dentre as diferentes operações pooling estão a *max pooling,* que seleciona, dentre dos dados que estão sob o núcleo, o maior valor entre eles, e a *average pooling,* que calcula a média aritmética entre as informações da região de dados sob o núcleo da camada (Taye, 2023).

A dimensão do núcleo mais comum é 2x2, enquanto o parâmetro do valor de strides, que também está presente nesta camada, geralmente é 2, podendo em certos casos ser utilizado o valor 1, quando não se quer diminuir o tamanho da amostra recebida pela camada anterior, o que não é tão comum (Zhao et al., 2024).

Assim, camadas de *pooling* reduzem a dimensão do mapa de características, e também diminuem a sensibilidade da rede às distorções e deslocamentos de imagem (Leite, 2022). Ou seja, essa camada subamostra o tamanho da imagem para que a carga computacional seja reduzida (Calheiro, 2021).

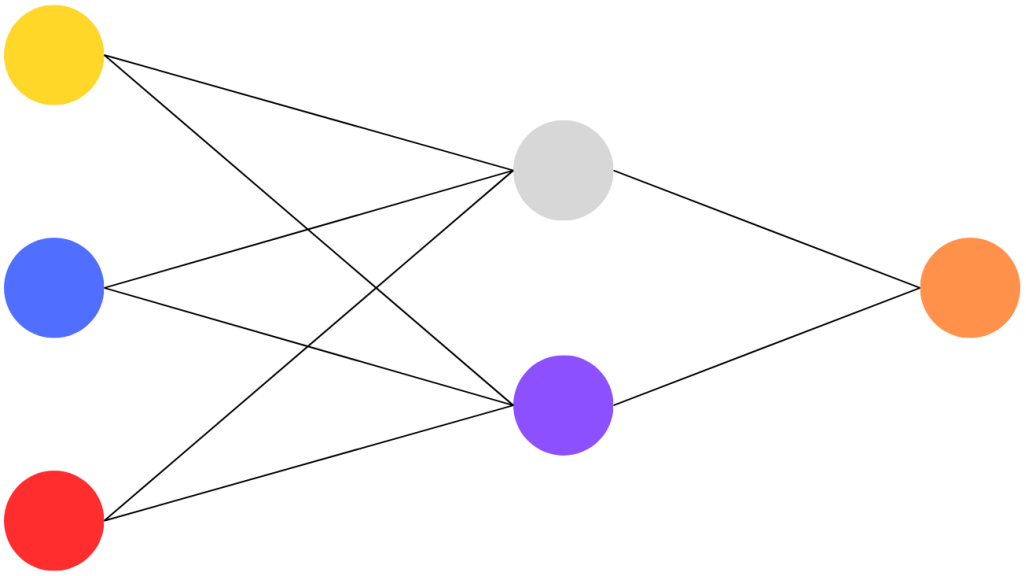
A camada de pooling resume o mapa de características da imagem em um único valor que será repassado para a camada posterior, uma camada totalmente conectada (Leite, 2022). Os neurônios dessa camada também, assim como a camada de convolução, se conectam apenas a algumas das saídas repassadas pela camada anterior, acrescenta (Calheiro, 2021).

2.8.3 Camada totalmente conectada

Por último, as camadas totalmente conectadas, similares às redes neurais de multicamadas tradicionais, ficam no final da rede e são as responsáveis por classificar os mapas de características obtidos e repassados pelas camadas anteriores (Leite, 2022). Essa camada é formada por uma estrutura de conexão entre os neurônios que realiza uma operação global nos dados provenientes da camada anterior.

Enquanto as camadas convolucionais e de *pooling* que realizam operações em porções dos dados, cada neurônio dessa camada está ligado a todos os neurônios da camada anterior e a todos os neurônios da camada posterior, como observado na Figura 5, e realizam a classificação dos dados usando pesos e vieses que compartilham (Zhao et al., 2024).

Figura 5 – Representação de uma camada totalmente conectada



Fonte: Elaborado pelo autor (2024).

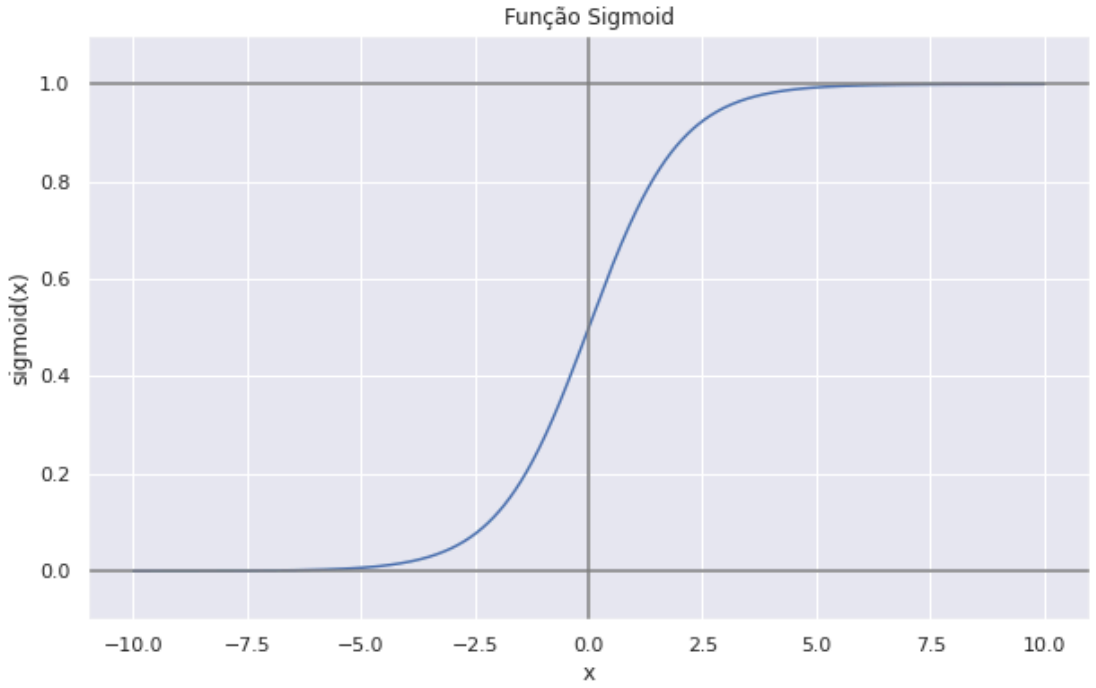
2.8.4 Funções de ativação

As funções são utilizadas para introduzir a não linearidade necessária para que os modelos de CNN sejam capazes de extrair padrões dos comportamentos de problemas complexos, que representam muitos dos problemas do mundo real, que não podem ser extraídos por redes rasas comuns com uma acurácia aceitável (Krichen, 2023). Estas funções estão presentes em camadas logo após as camadas de aprendizado (camadas não convolucionais e totalmente conectadas) (Taye, 2023).

A função *sigmoid* retorna como saída valores entre 0 e 1, porém podem receber qualquer valor real como entrada. Assim, essa função pode ser utilizada tanto para normalizar a saída de cada neurônio da rede, quanto como um modelo utilizado para predizer a probabilidade das saídas, utilizada principalmente em tarefas de classificação binária (Zhao et al., 2024). Essa função tem um formato em S, como pode ser percebido no Gráfico 1, e se dá pela fórmula 2.

$$f(x) = \frac{1}{1 + e^{-x}} \qquad (2)$$

Gráfico 1 – Função Sigmoid

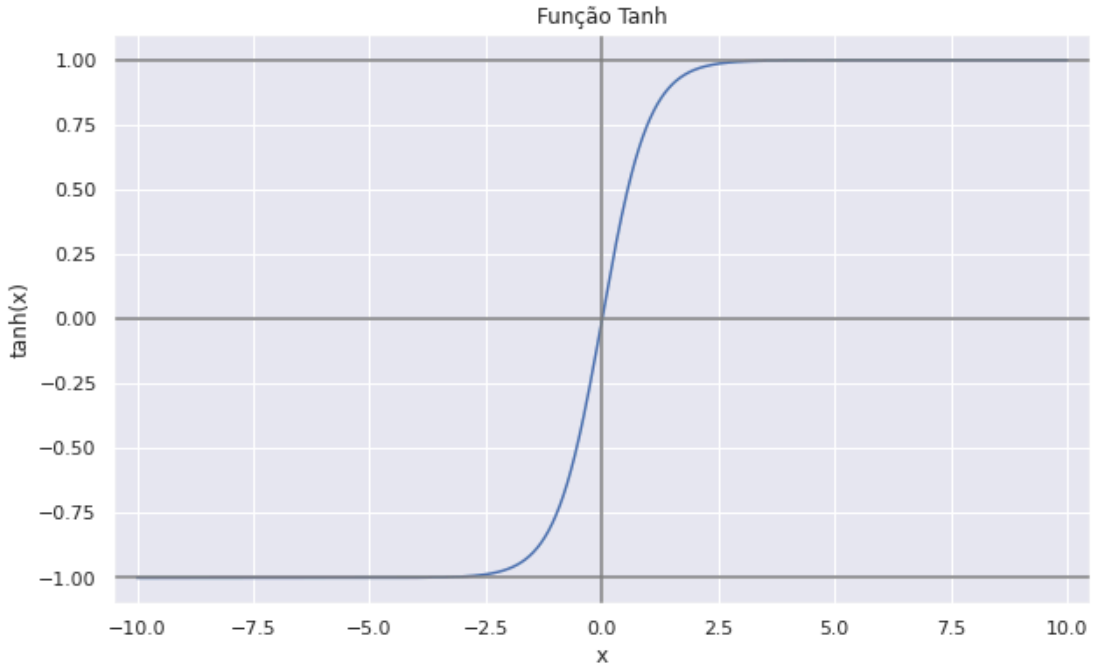


Fonte: Elaborado pelo autor (2024).

Podendo receber qualquer tipo de valor real, a função de ativação *Hyperbolic tangent* (Tanh) devolve saídas com valores entre -1 e 1. Assim como a função sigmoid, também apresenta um formato de S como observado no Gráfico 2, é definida pela fórmula 3.

$$f(x) = \frac{2}{1 + e^{-2x}} - 1 \qquad (3)$$

Gráfico 2 – Função Tanh

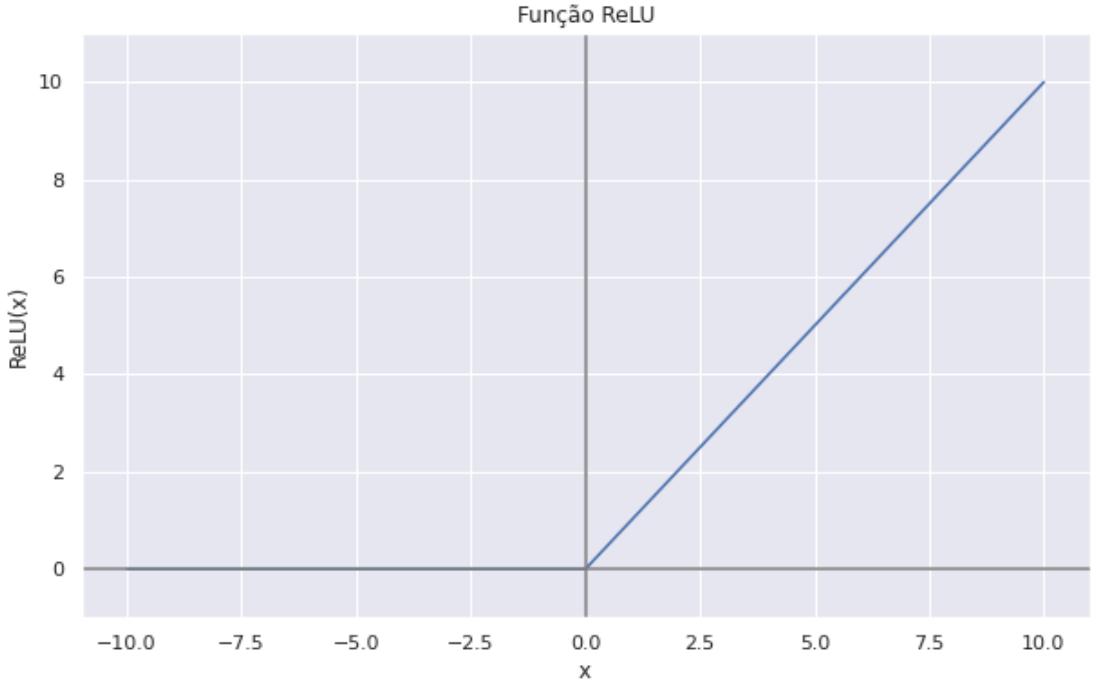


Fonte: Elaborado pelo autor (2024).

A função de ativação *Retified linear unit* (relu) é a mais simples e menos custosa, em termos de processamento, entre as três funções, visto que não apresenta cálculos exponenciais, retornando o maior valor entre o valor de entrada e 0, implicando em que para qualquer valor negativo a função retorna o valor 0, e é definida pela fórmula 4, como representa o Gráfico 3.

$$f(x) = max(0, x) \qquad (4)$$

Gráfico 3 – Função ReLU



Fonte: Elaborado pelo autor (2024).

2.8.5 *Batch Normalization*

O batch normalization é uma técnica de aplicada em redes neurais que realizar a normalização da saída da camada anterior, subtraindo a média dos valores e dividindo-os pelo valor de desvio padrão do *batch* (lote de informações que são passados à rede neural a cada época de treinamento) (Krichen, 2023).

Esse processo ajuda nos problemas que podem surgir no processo de *backpropagation*, durante o treinamento da rede (Zhao et al., 2024), como o *explosion gradient* (explosão de gradiente), quando os valores de gradiente são muito elevados e a rede não consegue convergir, além do problema de *vanishing gradient* (gradiente evanescente) que faz com que o sinal do gradiente diminua gradualmente à medida que a profundidade da rede aumenta, fazendo com que os pesos dos neurônios não sejam mais atualizados e a rede pare de aprender (Taye, 2023).

2.8.6 *Dropout*

A fim de impedir que o modelo não venha a se ater de forma mais intensa em apenas uma das características extraídas através dos dados, condição conhecida como *overfitting*, é aplicada a técnica de dropout que, de forma aleatória, ignora uma certa quantidade de neurônios da camada anterior, durante a fase de treinamento (Krichen, 2023), assim o modelo

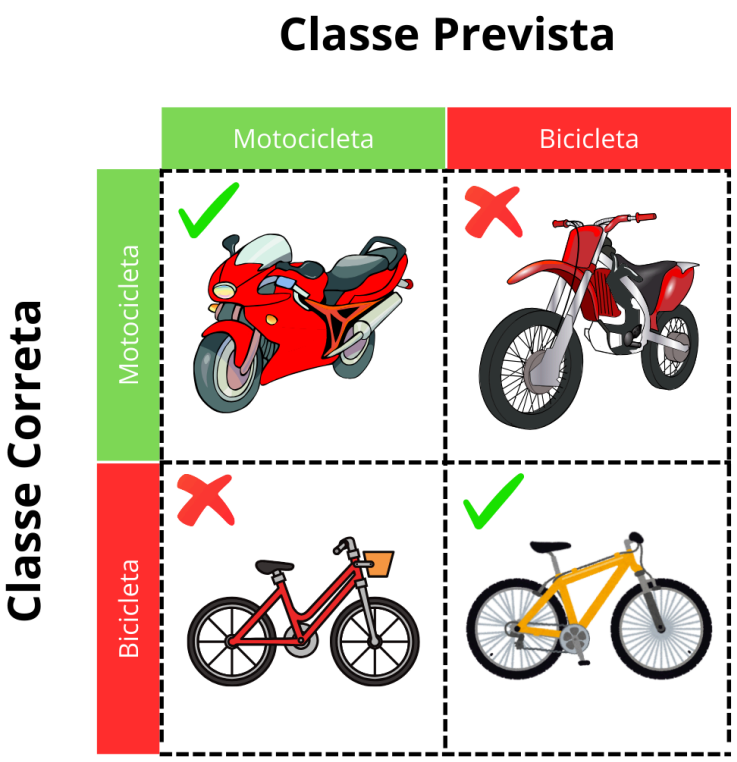
tende a aprender outras características adicionais de maneira mais igualitária, sem priorizar nenhuma caraterística em especial, atingindo uma capacidade de generalização do problema.

2.9 Métodos de avaliação

Para medir a performance do modelo podem ser utilizadas diferentes métricas, dependendo do objetivo que deseja alcançar, bem como a forma que se deseja comparar os resultados com aqueles obtidos por diferentes abordagens do mesmo problema.

A matriz de confusão é uma forma de visualizar os resultados obtidos pelo modelo de maneira mais clara, visto que é possível visualizar a quantidade de classificações que foram feitas para cada classe tanto as corretas quanto as equivocadas. Onde, nesta matriz, as linhas representam as classes reais, que o modelo deveria classificar, enquanto as colunas representam as classes que o modelo, de fato, classificou (FIGURA 6).

Figura 6 – Exemplo de matriz de confusão



Fonte: Elaborado pelo autor (2024).

A partir da matriz de confusão é possível obter certas variáveis que são utilizadas para realizar o cálculo de algumas métricas que ajudam a avaliar melhor o desempenho do modelo de classificação. As variáveis obtidas são:

- VP: quantidade de classificações verdadeiras positivas;
- VN: quantidade de classificações verdadeiras negativas;
- FP: quantidade de classificações falsas positivas;
- FN: quantidade de classificações falsas negativas.

A acurácia é utilizada para medir a quantidade de previsões feitas pelo modelo que foram corretas, indicando a performance geral do modelo. E se dá pela fórmula 5:

$$\text{acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (5)$$

A precisão é a métrica de avaliação que diz respeito a quantidade de classificações positivas que foram feitas de forma correta. Calculado de acordo com a fórmula 6:

$$\text{precisão} = \frac{VP}{VP + FP} \quad (6)$$

A sensibilidade mede a quantidade de classificações que foram realizadas de forma correta, dentre todas as classificações positivas que eram esperadas. Obtida a partir da fórmula 7:

$$\text{sensibilidade} = \frac{VP}{VP + FN} \quad (7)$$

Enquanto a métrica F1-*Score* consiste na média harmônica entre a sensibilidade e a precisão do modelo. Definida pela fórmula 8:

$$\text{F1-Score} = \frac{2 * precisão * sensibilidade}{precisão + sensibilidade} \quad (8)$$

2.10 Trabalhos Relacionados

Nesta seção serão apresentados alguns dos trabalhos relacionados presentes na literatura, que abordam o tema de classificação de emoções de forma automática através da fala.

Campos e Moutinho (2020) buscaram desenvolver e validar uma arquitetura composta por um conjunto de modelos de CNNs especialistas e modelos de *Deep Learning* chamada de Detection of voice Emotion in Portuguese language (DEEP). Os modelos foram treinados em língua portuguesa utilizando a base de dados *Voice Emotion Recognition dataBase in Portuguese language* (VERBO), criada criada no Instituto de Matemática e Ciências da Computação da Universidade de São Paulo (ICMC-USP).

A base de dados é composta por 1176 arquivos de áudio balanceados foneticamente que variam de 2 a 5 segundos, gravados por doze atores brasileiros de diferentes idades e regiões, sendo seis do sexo masculino e seis do sexo feminino, expressasndo sete emoções distintas: alegria, nojo, medo, raiva, surpresa, tristeza e neutra.

Assim, Campos e Moutinho (2020) coletaram e categorizaram os dados de voz para serem aplicados no método de aprendizagem de máquina supervisionado, sendo utilizadas as *features* MFCC, Prosódicas e Cromáticas. Os classificadores foram desenvolvidos com o *framework Keras* em *Python* e tiveram o seus hiper parâmetros otimizados a fim de obter melhores resultados do que os modelos desenvolvidos e apresentados na literatura. O modelo apresentou uma acurácia média de 76% para a classificação das sete emoções.

Santos (2022) desenvolve e avalia o desempenho de um classificador implementado utilizando uma rede neural convolucional com o intuito de classificar emoções através da fala, uma vez que o autor reconhece que esse tipo de rede neural apresenta um bom desempenho em tarefas de classificação de áudio e reconhecimento de fala.

Ele usa a base de dados RAVDESS, um banco de dados audiovisuais de fala e música emocional da Universidade de Ryerson, composto por 1440 arquivos de áudios, de menos de 4 segundos em média, gravados por atores profissionais com sotaque norte-americano neutro, divididos igualmente entre homens e mulheres. Os áudios do dataset representam oito sentimentos: raiva, nojo, calma, medo, tristeza, felicidade, surpresa e neutro.

O modelo desenvolvido por Santos (2022) foi utilizado para realizar a classificação a partir de três técnicas de análise de áudio. A primeira técnica foi a transformada de *Fourier* de tempo curto (STFT), com essa técnica o modelo conseguiu uma acurácia de 8% na classificação do conjunto de teste, após um treinamento de 50 épocas, e 23% após o treinamento com 75 épocas.

Analisando os resultados, Santos (2022) concluiu que o modelo apresentou o problema de *underfitting*, que é o termo utilizado para se referir ao comportamento indesejado apresentado pelos modelos de aprendizagem de máquina que não são capazes de extrair

relações e padrões a partir das amostras utilizadas no processo de treinamento do modelo, ou seja, não conseguem aprender a partir dos dados de treinamento.

A segunda técnica de análise utilizada por Santos (2022) foi a de Espectrograma Mel, o modelo submetido a 50 épocas de treinamento apresentou uma acurácia de 28% a partir da classificação do conjunto de teste, e o modelo submetido a 75 épocas de treinamento obteve uma acurácia de 38%. Enquanto a terceira técnica de análise de áudio, as *features* de classificação *Mel Frequency Cepstral Coefficients* (MFCC), apresentou, para o conjunto de teste, uma acurácia de 62% quando utilizado o modelo que foi submetido a 50 épocas de treinamento e 65% com modelo treinado por 75 épocas.

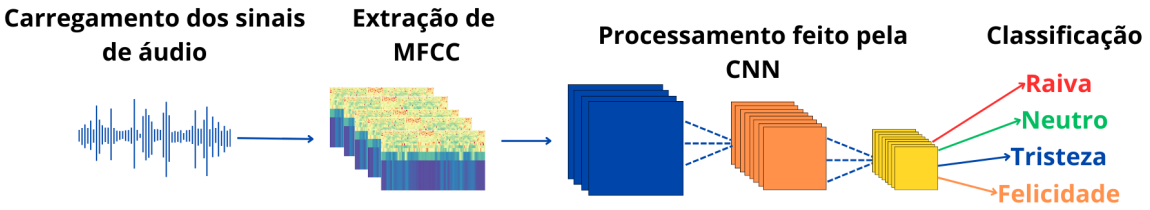
Peixoto e Linhares (2023) apresentam o desenvolvimento de uma rede neural convolucional treinada para reconhecer emoções através da fala. A rede foi treinada com arquivos de áudios que compõem a base de dados emoUERJ, base criada pela Universidade Estadual do Rio de Janeiro, a fim de proporcionar material em português do Brasil para o desenvolvimento de sistemas de classificação de emoções através da fala. Para treinar a rede utilizaram as *features* MFCC.

Peixoto e Linhares (2023) obtiveram uma acurácia de 90% na fase de treinamento, porém em testes feitos para determinar se as variações linguísticas do português afetariam o desempenho do classificador, e obtiveram resultados considerados insatisfatórios nestes testes, em média uma acurácia de 24% para as quatro variações linguísticas. Assim, concluíram que a falta de bases de dados em português que abranjam as diferentes variações linguísticas presentes no Brasil dificultam o processo de treinamento e desenvolvimento de classificadores eficientes.

3 METODOLOGIA

Neste trabalho será realizada uma pesquisa bibliográfica a fim de obter o embasamento teórico necessário para melhor entender o problema do reconhecimento de emoções através da fala, as suas aplicações, bem como conhecer quais são as principais técnicas e tecnologias utilizadas no desenvolvimento de uma solução passível de implementação e de aferição dos resultados. O fluxo de classificação das emoções pode ser observado na FIGURA 7.

Figura 7 – Representação do fluxo do processo de classificação das emoções.



Fonte: Elaborado pelo autor (2024).

3.1 Ambiente de desenvolvimento

Para o desenvolvimento desse projeto apresentado neste trabalho, será utilizado uma máquina com processador Intel Core I5, 8GB de RAM, com o sistema operacional Linux Mint. Será utilizada a linguagem de programação Python na versão 3.6.9 para a realizar o processamento dos áudios, extração dos MFCCs bem como a construção do modelo de CNN.

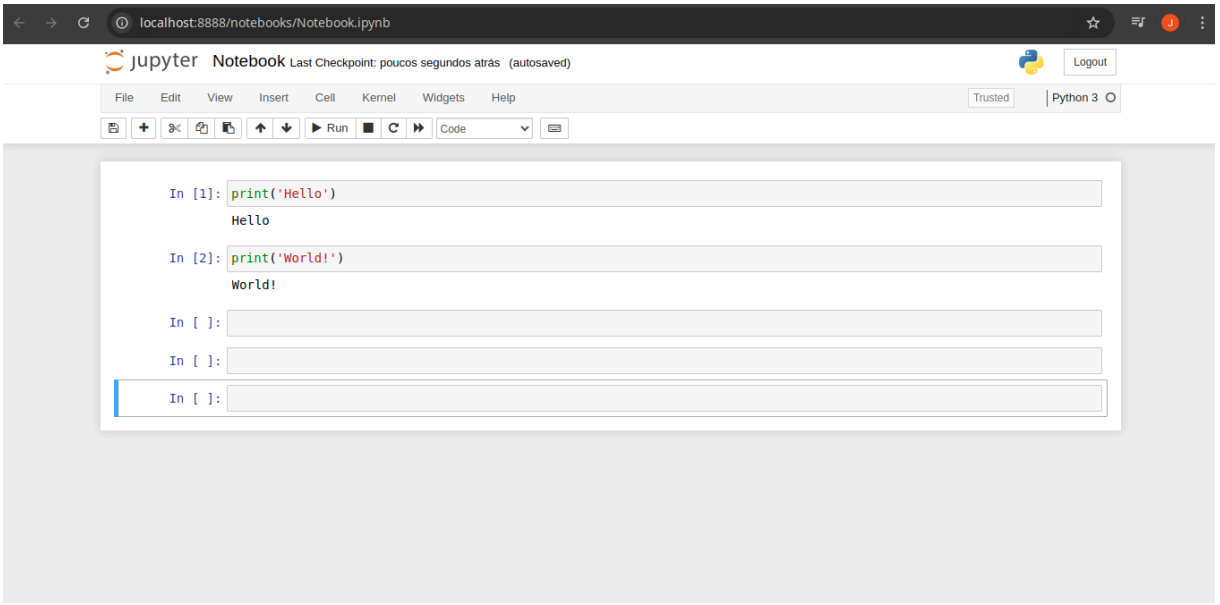
Para prosseguir com o processamento dos áudios e a extração dos MFCCs será utilizada a biblioteca Librosa[1]. Além disso, para a construção da rede neural convolucional bem como o seu treinamento e teste, será utilizada a biblioteca *TensorFlow*[2], uma interface utilizada para o desenvolvimento de algoritmos de aprendizado de máquina que podem ser executados em diferentes plataformas, como dispositivos móveis, sistemas distribuídos de larga escala, sistemas embarcados, entre outros, com poucas ou nenhuma alteração no código desenvolvido.

Também será utilizada a biblioteca *Keras*[3], uma API desenvolvida em Python que funciona como uma abstração de alto nível para diferentes frameworks de Deep Learning como o TensorFlow, que proporciona um desenvolvimento simples, flexível e poderoso capaz de oferecer performance e escalabilidade em nível industrial.

[1] Biblioteca librosa: https://doi.org/10.5281/zenodo.11192913.
[2] Biblioteca TensorFlow: https://www.tensorflow.org/.
[3] Biblioteca Keras: https://keras.io/.

E para a execução desse desenvolvimento será usada a ferramenta Jupyter Notebook, que possibilita executar código python em células de forma dinâmica diretamente de um navegador web (FIGURA 8), ajudando no desenvolvimento pois permite que seja possível testar o código em pequenas partes, facilitando a identificação de possíveis erros de forma mais fácil, além de proporcionar uma visualização interativa de dados, sendo possível plotar gráficos e imagens de forma fácil e rápida.

Figura 8 – Jupyter notebook com código Hello World!



Fonte: Elaborado pelo autor (2024).

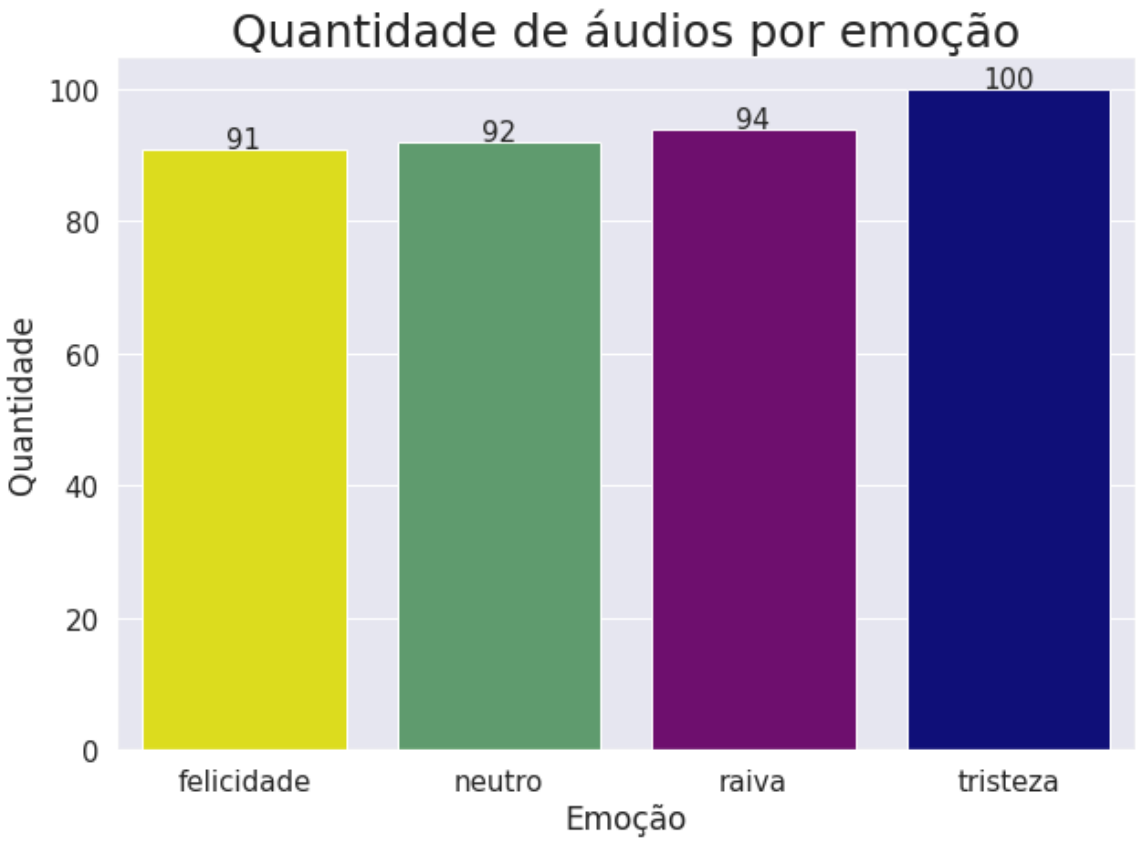
3.2 Base de dados utilizada

Será utilizada como a base de dados de áudios, estes que contém as vozes representando as emoções que servirão para treinar a rede neural convolucional, a base de dados emoUERJ[4], produzida pela Universidade do Rio de Janeiro devido ao fato de haverem poucas bases de dados na língua portuguesa do Brasil para ser usada no desenvolvimento de sistemas de reconhecimento de emoções através da fala.

Esta base possui 377 áudios que abordam quatro emoções, sendo gravados por oito atores, quatro homens e quatro mulheres, os quais gravaram dez sentenças, as quais foram escolhidas pelos atores, para pronunciarem essas sentenças em quatro emoções distintas: alegria, raiva, tristeza e neutra (Gráfico 4).

[4] Base de dados emoUERJ: https://zenodo.org/doi/10.5281/zenodo.5427548.

Gráfico 4 – Quantidade de áudios representando cada emoção



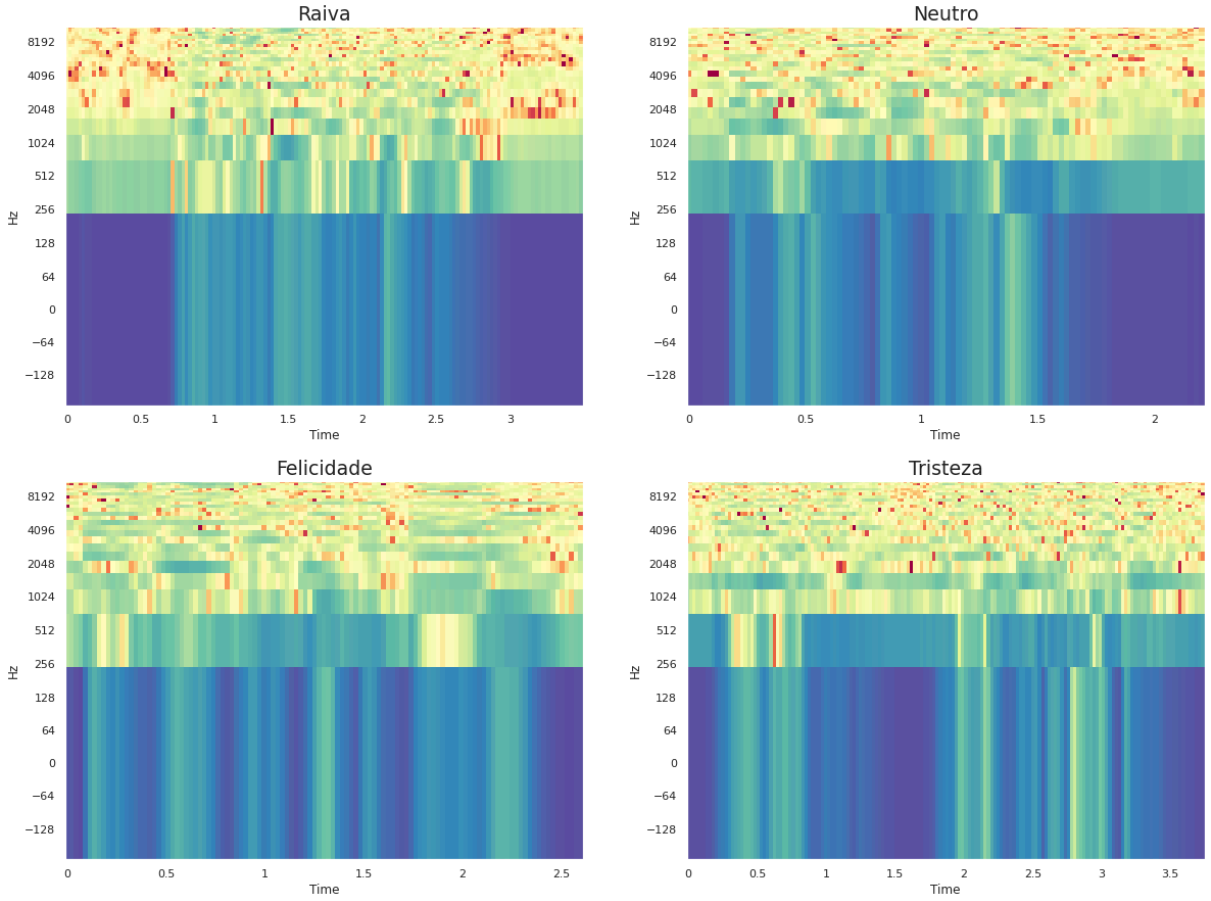
Fonte: Elaborado pelo autor (2024).

3.2 Extração de MFCC

Para realizar a extração dos coeficientes MFCC (FIGURA 9) será utilizada a biblioteca Librosa, que é utilizada para realizar análises de áudios e músicas. De cada áudio foram extraídos 24 coeficientes MFCC para a realização da análise e classificação de emoções. Após a extração, os valores dos coeficientes serão atualizados para a média de cada ndarray com esses valores, visto que os arrays apresentam diferentes tamanhos na dimensão do tempo.

Para que fosse possível manter os áudios originais, deveria ser realizado algum tratamento a fim de deixar os áudios com a mesma duração, porém, isso poderia ocasionar a perda de algumas informações importantes que estivessem presentes nos áudios, e assim afetar o processo de classificação.

Figura 9 – MFCCs de diferentes emoções



Fonte: Elaborado pelo autor (2024).

3.3 Modelo de Rede Neural Convolucional (CNN)

3.3.1 Arquitetura

O modelo da CNN será desenvolvido utilizando as bibliotecas TensorFlow e Keras. Como pode-se observar na Figura 10 e na Figura 11, a rede neural iniciará com a camada de entrada, com o formato de entrada (24, 1), logo após haverá uma camada convolucional com 128 filtros com o núcleo convolucional de valor 3, e a função de ativação relu, que será utilizada em todas as camadas convolucionais e totalmente conectadas deste modelo, exceto pela última camada totalmente conectada, que terá como função de ativação a função *softmax*.

Em seguida, será adicionada uma camada de *dropout* que desativa 40% dos neurônios da camada de convolução anterior a esta. A próxima camada tratará de outra camada de convolução, só que esta apresentando apenas 64 filtros e com o mesmo tamanho do núcleo da camada de convolução anterior. Posterior a esta, outra camada de *dropout* será adicionada ignorando 30% dos neurônios da camada convolucional logo anterior.

Em seguida, será inserida uma camada de flatten, para achatar as saídas da camada anterior e passá-las para a próxima camada, a próxima camada será uma camada totalmente conectada, sendo chamada de dense no framework Keras, esta camada apresenta 128 neurônios.

Logo após, outra camada de dropout com uma taxa de desativação de neurônios de 20%. Em seguida, outra camada totalmente conectada, com 64 neurônios, seguida de outra camada totalmente conectada com 32 neurônios, e por fim uma última camada totalmente conectada com 4 neurônios, que é a quantidade de classes do problema de classificação em questão, as quatro emoções, com a função de ativação softmax.

Não será utilizada nenhuma camada de *Pooling* nem de *Batch Normalization* na composição da CNN. Nos testes previamente realizados com arquiteturas que possuíam essas duas camadas, ou, que contavam com apenas uma dessas duas camadas na sua composição, nenhuma dessas apresentou uma performance razoavelmente robusta. Talvez pela quantidade reduzida dos dados de treinamento, que impossibilita uma melhor performance da rede quando essas camadas são adicionadas.

3.3.2 Otimização (Fine Tuning)

Para que a rede possa atualizar os pesos e vieses da rede neural convolucional durante o processo de treinamento, serão utilizados algoritmos de otimização que funcionam de forma a buscar a diminuição da função de perda (Krichen, 2023).

Para isso, o algoritmo de otimização utilizado será o *Adaptive Moment Estimation* (Adam), que combina os pontos fortes dos métodos de gradientes adaptativos e da raiz quadrada média da propagação, percorrendo todo o espaço dos parâmetros enquanto corrige os vieses, além de conservar os recursos computacionais proporcionando diferentes taxas de aprendizado para os diferentes parâmetros (Kang et al., 2024).

Também será utilizada, no processo de treinamento, uma técnica de agendamento que age diminuindo a taxa de aprendizado do modelo quando este, depois de uma quantidade definida de épocas, para de apresentar melhora nos resultados, sendo definida uma variável a qual deseja monitorar, como acurácia ou perda do modelo.

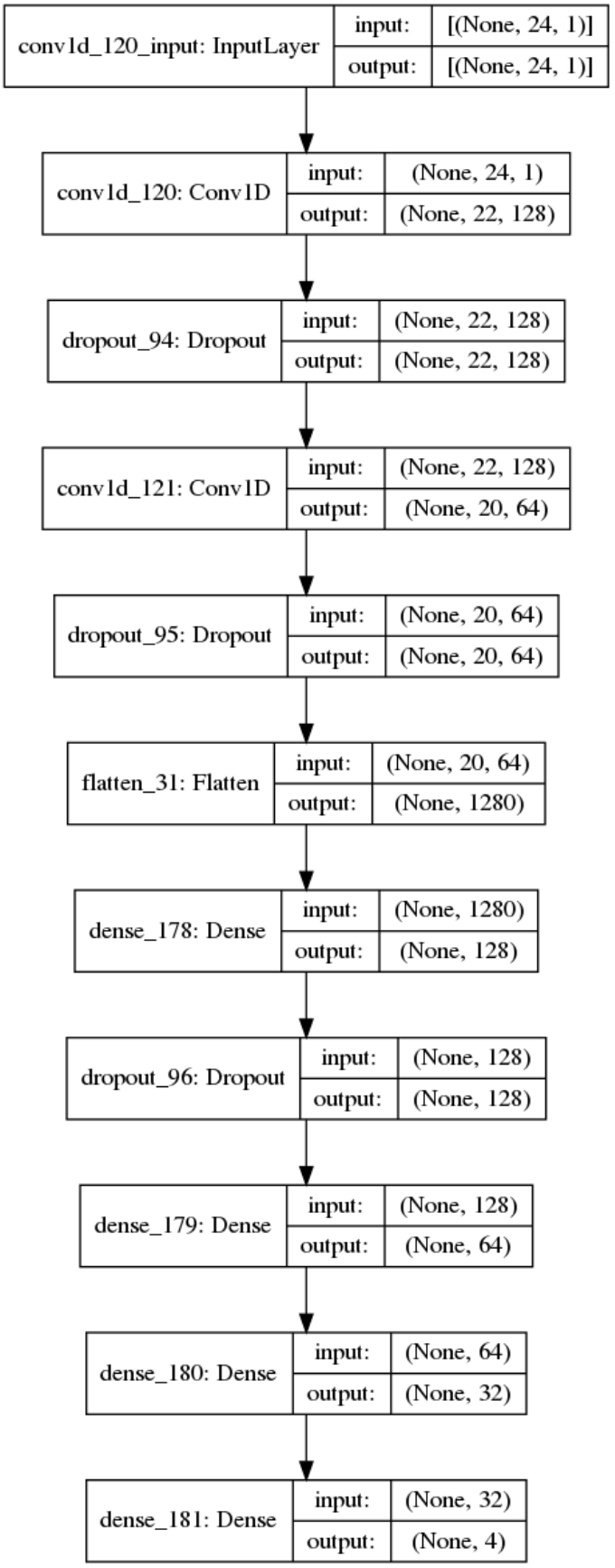
Assim, quando o modelo, durante o treinamento, apresenta uma estagnação ou mesmo piora da variável monitorada, após a quantidade de épocas definida, a taxa de aprendizado do modelo é diminuída na proporção definida, assim o modelo passa a avaliar os mapas de características de forma mais lenta, podendo apresentar uma melhora na sua performance. Esta técnica de agendamento é chamada de ReduceLROnPlateau.

Figura 10 – Código da CNN

```
modelo4 = keras.Sequential([
    keras.layers.Conv1D(128, activation=tensorflow.nn.relu, kernel_size=3,input_shape=(24, 1)),
    keras.layers.Dropout(0.4),
    keras.layers.Conv1D(64, activation=tensorflow.nn.relu, kernel_size=3),
    keras.layers.Dropout(0.3),
    keras.layers.Flatten(),
    keras.layers.Dense(128, activation=tensorflow.nn.relu),
    keras.layers.Dropout(0.2),
    keras.layers.Dense(64, activation=tensorflow.nn.relu),
    keras.layers.Dense(32, activation=tensorflow.nn.relu),
    keras.layers.Dense(4, activation=tensorflow.nn.softmax),
    ]
)

# Otimizador
adam1 = keras.optimizers.Adam(learning_rate=0.001)

# Técnica de agendamento que diminui a taxa de aprendizagem
reduce_lr1 = keras.callbacks.ReduceLROnPlateau(monitor='val_loss',
                                               factor=0.1,
                                               patience=10,
                                               min_lr=0.00001)

# Compilando o modelo
modelo4.compile(optimizer=adam1,
                loss='categorical_crossentropy',
                metrics=['accuracy'])
```

Fonte: Elaborado pelo autor (2024).

Figura 11 – Arquitetura da CNN



Fonte: Elaborado pelo autor (2024).

3.3.3 Treinamento

Para realizar o treinamento da rede neural convolucional, os parâmetros MFCCs que serão extraídos dos áudios, serão divididos nas proporções 80% e 20% dos dados, de forma aleatória, para serem utilizados no processo de treinamento e teste respectivamente, resultando em 301 dos dados de MFCC para treinamento, e os 76 restantes para a fase de teste. Além disso, 15% dos dados de treinamento serão utilizados para realizar a validação da rede durante a fase de treinamento. O treinamento foi realizado em 100 épocas.

4 RESULTADOS

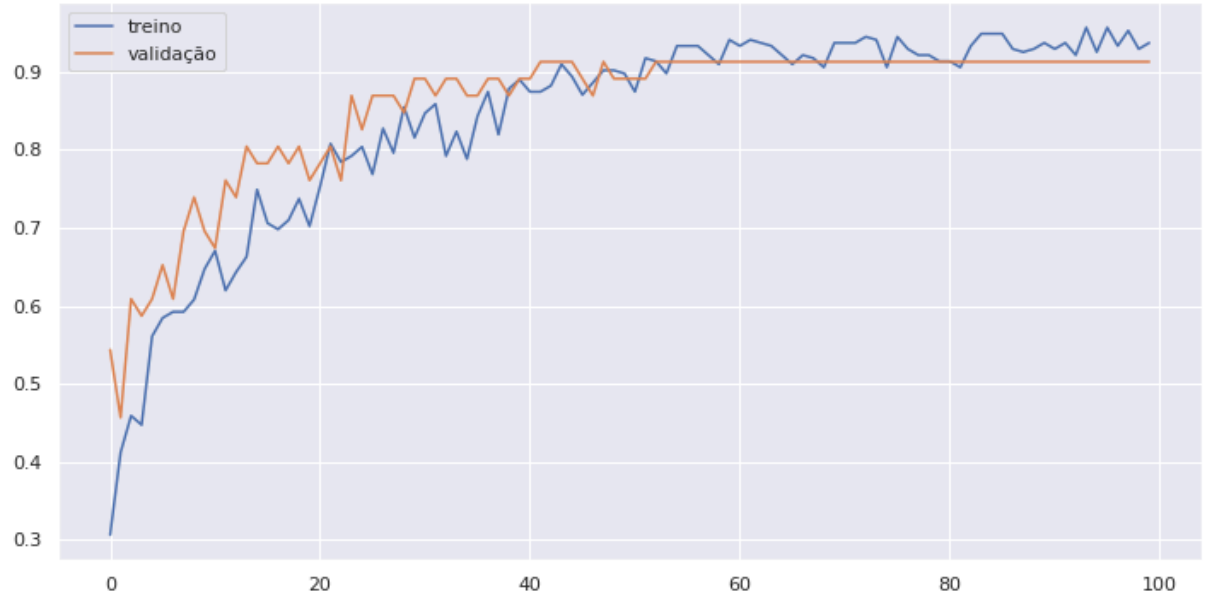
Primeiramente, foi realizada uma análise dos arquivos de áudios, que como está descrito na documentação da base de dados emoUERJ, os nomes dos arquivos contém a descrição da emoção representada no áudio, bem como se foi gravado por uma pessoa do gênero masculino ou femino, entre outras informações. Para tal tarefa, foi desenvolvida uma função que extrai as classes das emoções dos áudios, bem como o caminho do arquivo de áudio e salva essas informações em um dataframe, da biblioteca pandas.

Em posse das classes de emoções de cada áudios e seus respectivos caminhos, foram extraídos os parâmetros MFCCs de cada áudio e foram armazenados no mesmo dataframe com as demais informações. Após ter extraído os parâmetros de todos os áudios, os dados foram divididos em duas partes, uma parte dos dados, 80%, foram separados para o processo de treinamento, e o restante dos dados para a fase de teste.

4.1 Fase de treinamento

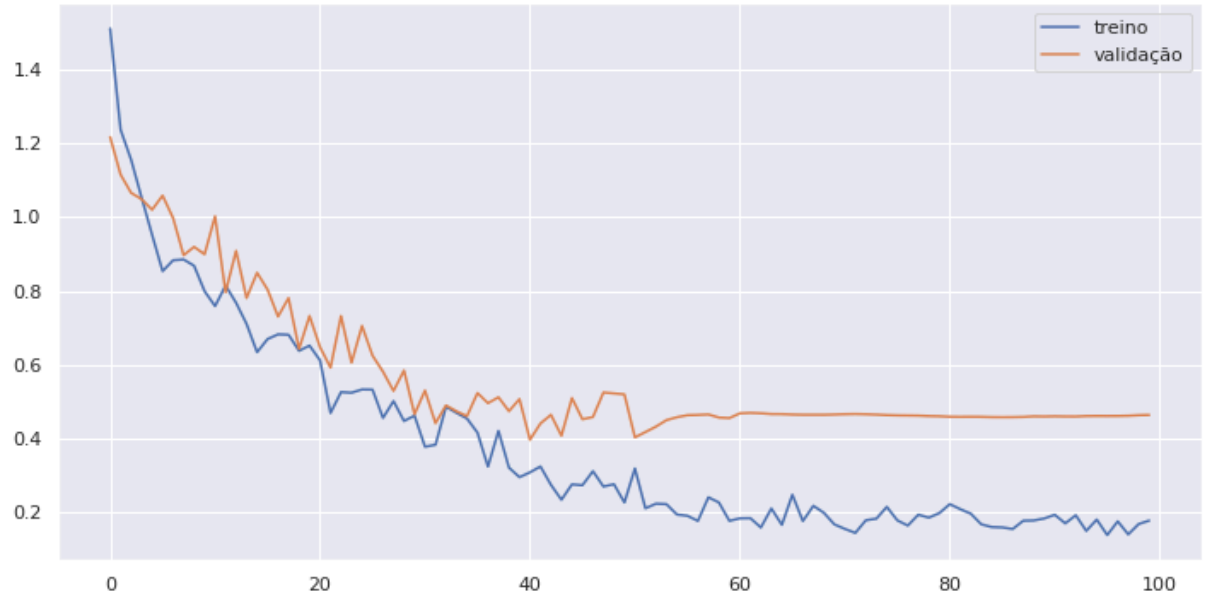
Então, seguiu-se a fase de treinamento com os dados separados para a esta fase, resultando em uma acurácia, no final do processo, de 93,73% com conjunto de treinamento, e uma acurácia final de 91,30% no conjunto de validação, conjunto de 15% dos dados de treinamento (GRÁFICO 5). Além disso, o modelo apresentou uma perda de 0,1783 com o conjunto de treinamento no final do processo e para o conjunto de validação o valor de perda foi 0,4642, como pode ser observado no Gráfico 6.

Gráfico 5 – Acurácia durante o treinamento



Fonte: Elaborado pelo autor (2024).

Gráfico 6 – Perda durante o treinamento

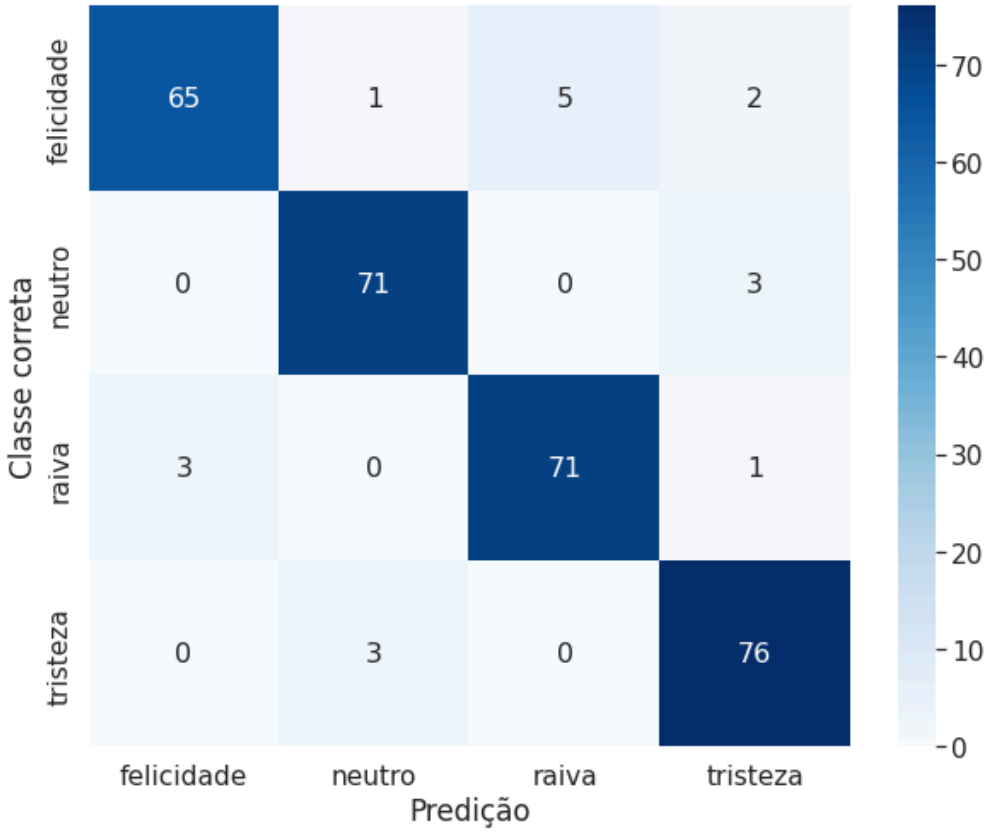


Fonte: Elaborado pelo autor (2024).

Analisando o desempenho do classificador, na fase treinamento, por cada classe de emoção, como observa-se na matriz de confusão (FIGURA 12), tem-se que a emoção com o maior número de acertos pela quantidade de amostras foi a classe felicidade, com 76 acertos do total de 79 amostras, tendo 3 amostras da classe tristeza sendo classificadas como pertencentes a classe neutro.

E a classe com o menor número de acertos por número de amostras, foi a classe felicidade, que teve 65 acertos dentre o total de 73 amostras, com 2 amostras dessa classe sendo classificadas como pertencentes a classe tristeza, outras 5 classificadas como sendo da classe raiva, e apenas uma foi classificada, erroneamente, como sendo da classe neutro.

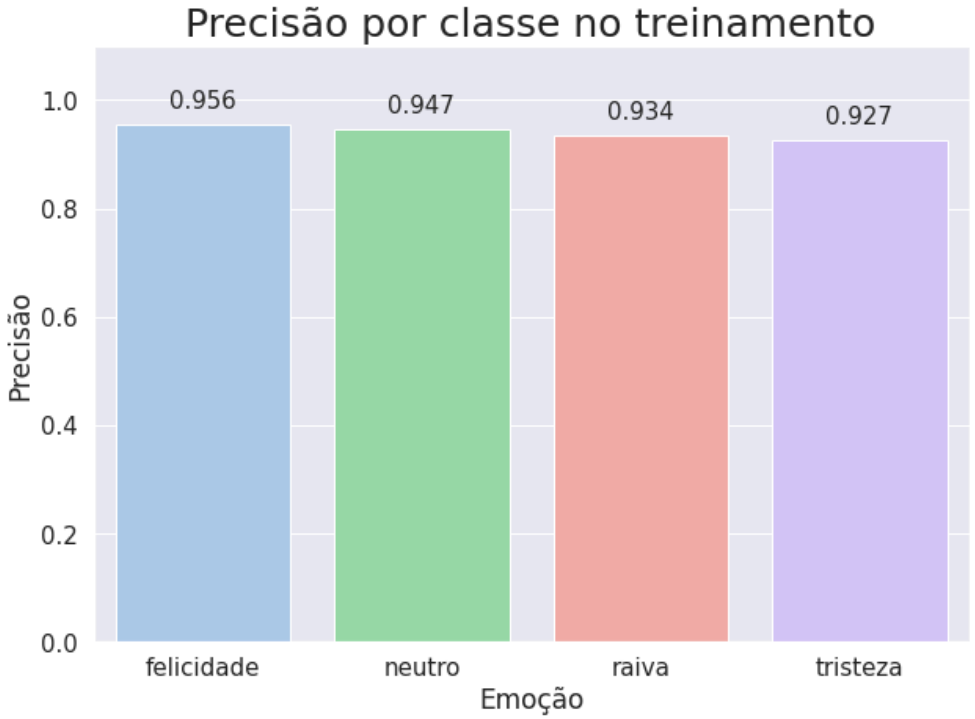
Figura 12 – Matriz de confusão fase treinamento



Fonte: Elaborado pelo autor (2024).

A partir dos valores de precisão calculados individualmente para cada classe, observa-se que a classe que apresentou a melhor taxa de precisão foi a felicidade, tendo em vista que foi a classe com o menor número de falsos positivos em relação à quantidade total de amostras da classe, com o resultado 0,95 de precisão. Já a classe que apresentou o maior número de falsos positivos pela quantidade total de amostras da classe, apresentando assim o menor desempenho na métrica de precisão, foi a tristeza, com 0,92 de precisão (GRÁFICO 7).

Gráfico 7 – Precisão fase treinamento



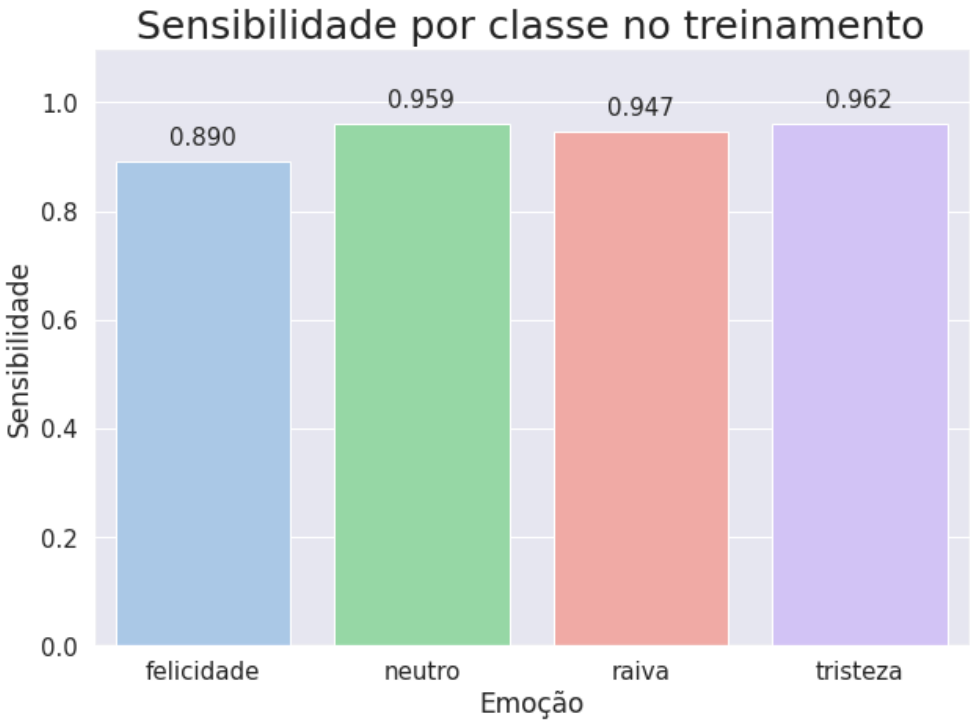
Fonte: Elaborado pelo autor (2024).

Para os valores de sensibilidade, a classe com o melhor desempenho, foi a classe tristeza, a qual apresentou o menor número de falsos negativos em relação à quantidade total de amostras da classe, com 0.96 para o cálculo da sensibilidade. Enquanto a classe felicidade apresentou o pior desempenho, visto que teve o maior número de falsos negativos quando comparado com o total de amostras da classe, com o resultado de 0,89 (GRÁFICO 8).

Em relação aos valores obtidos a partir do cálculo da métrica *f1-score* para o conjunto de treino, observou-se que a classe felicidade apresentou o menor resultado, uma vez que foi a classe que apresentou o maior número de falsos negativos pela quantidade total de amostras da classe, com um valor de 0,92 (GRÁFICO 9).

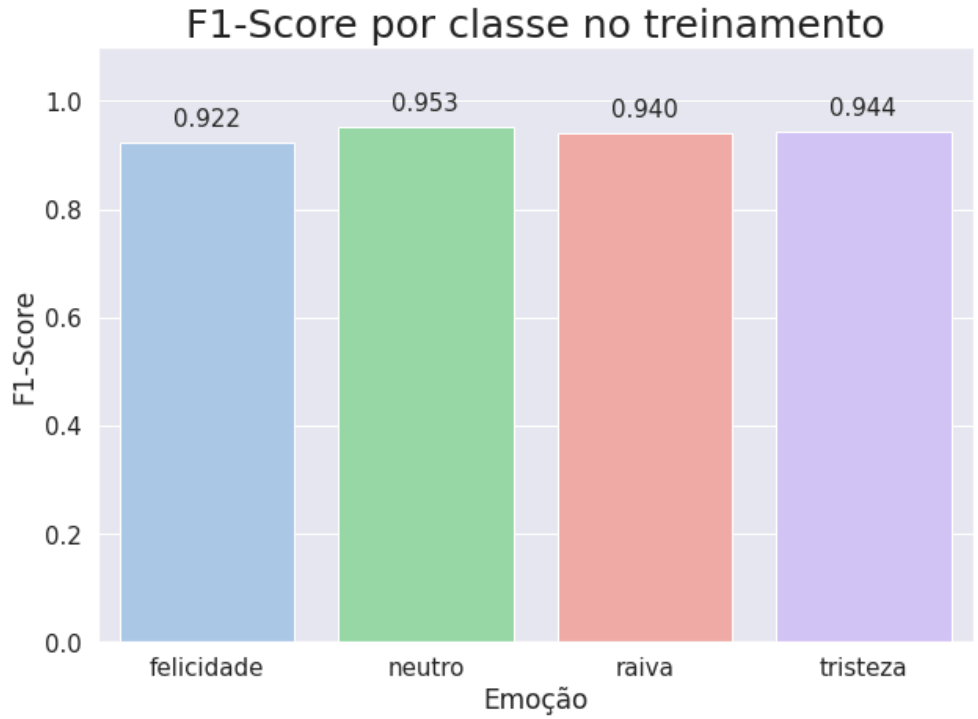
Já a classe que apresentou o melhor valor para a métrica, foi a classe neutro, que apesar da classe ter apresentado a mesma quantidade de falsos negativos que a classe tristeza, apresentou uma menor quantidade de falsos positivos em relação à classe tristeza, com o resultado de 0,95 para a métrica *f1-score* (GRÁFICO 9).

Gráfico 8 – Sensibilidade fase treinamento



Fonte: Elaborado pelo autor (2024).

Gráfico 9 – F1-Score fase treinamento



Fonte: Elaborado pelo autor (2024).

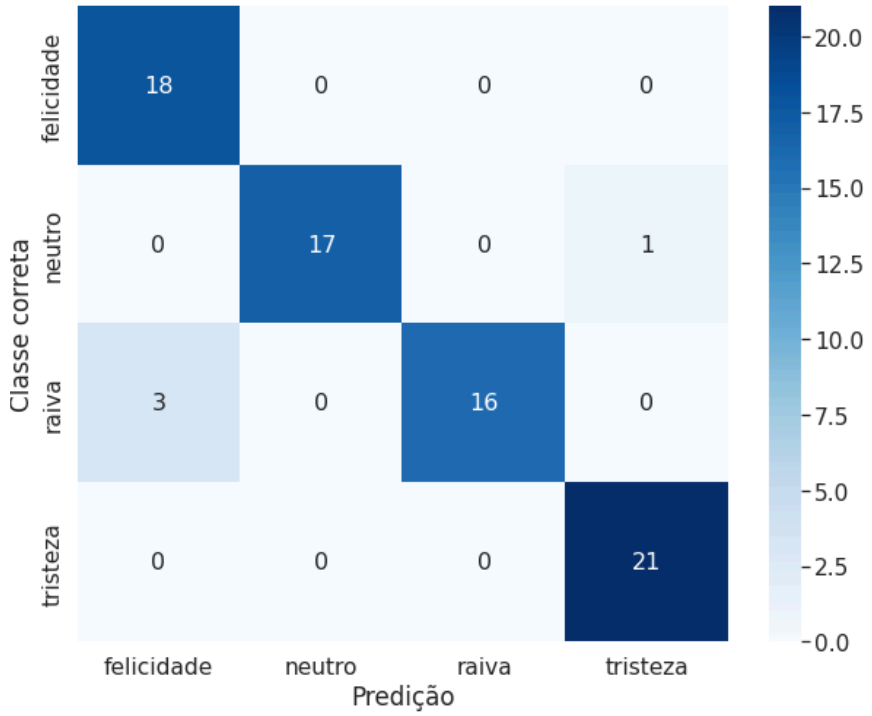
4.2 Fase de teste

Utilizando a porção dos dados separados para o processo de teste do modelo CNN classificador. O modelo conseguiu obter uma acurácia de 95%, maior que as acurácias observadas na fase treinamento tanto do conjunto de treinamento quanto com a porção de validação, e apresentou ainda uma perda de 0.306, que foi maior que a observada na fase de treinamento, porém menor que a perda obtida na validação. Com esses resultados pode-se concluir que o modelo não apresentou overfitting aos dados usados na fase de treinamento.

Com a análise do desempenho do classificador em relação a cada classe de emoção, observa-se na matriz de confusão (FIGURA 13) que, dentre as amostras utilizadas no treinamento, duas classes apresentaram 100% de acerto, foram elas as classes tristeza e felicidade.

Enquanto a classe raiva apresentou o menor número de acertos pela quantidade total de amostras, com 16 acertos do total de 19 amostras, tendo as outras três amostras sendo classificadas como pertencentes à classe felicidade. Já a classe de emoção neutro, teve 17 acertos do total de 18 amostras, enquanto uma das amostras foi classificada erroneamente como sendo da classe tristeza.

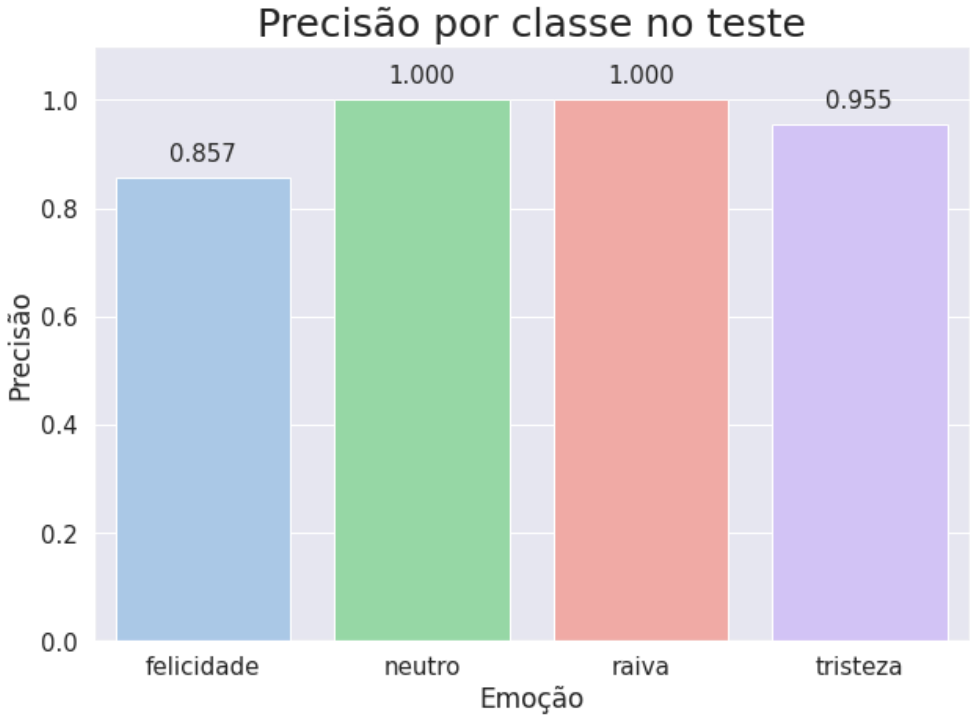
Figura 13 – Matriz de confusão teste



Fonte: Elaborado pelo autor (2024).

Duas classes apresentaram valor máximo para a precisão, sendo elas a classe neutro e a classe raiva, uma vez que estas não apresentaram nenhum falso positivo na classificação das amostras. Já a classe com a pior precisão foi a classe felicidade, apresentando a maior quantidade de falsos positivos em relação à quantidade total de amostras da classe com uma precisão de 0,857. Enquanto a classe tristeza só apresentou um falso positivo e ficou com uma precisão de 0,955 (GRÁFICO 10).

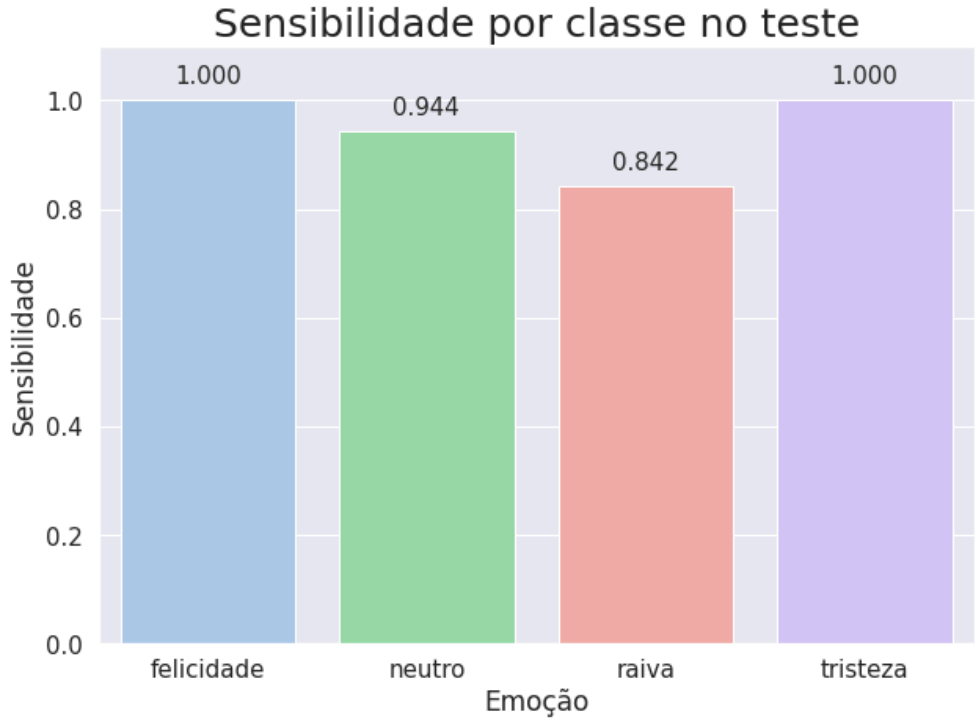
Gráfico 10 – Precisão fase teste



Fonte: Elaborado pelo autor (2024).

Em relação à sensibilidade, duas classes apresentaram valor máximo, as classes felicidade e tristeza. Enquanto a classe que apresentou o menor valor de sensibilidade, e portanto, a maior quantidade de falsos negativos em relação ao total de amostras para a classe, foi a classe raiva, com a sensibilidade de 0,84. Já a classe neutro apresentou uma sensibilidade de 0,94 (GRÁFICO 11).

Gráfico 11 – Sensibilidade fase teste

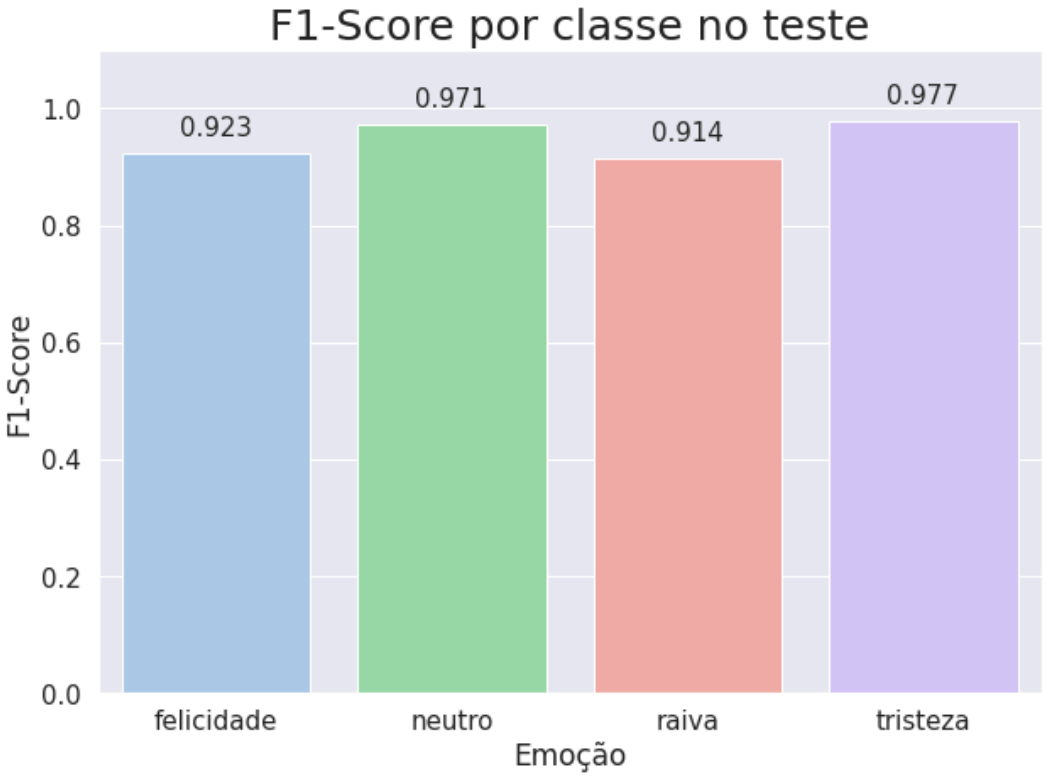


Fonte: Elaborado pelo autor (2024).

Para a métrica f1-score, as classes neutro e tristeza apresentaram apenas um falso negativo e um falso positivo respectivamente, porém, como a quantidade de amostras da classe tristeza foi maior que a da classe neutro, o maior valor de f1-score foi 0,977 para a classe tristeza. A classe neutro apresentou a segunda melhor performance para a métrica com o valor 0,971 (GRÁFICO 12).

Enquanto a classe com o menor valor para o resultado da f1-score foi a classe raiva, que apresentou três falsos negativos que foi a mesma quantidade de falsos positivos da classe felicidade, porém, visto que a quantidade de amostras da classe felicidade foi maior que a da classe raiva, esta apresentou um menor valor para a métrica f1-score com o resultado de 0,91, já a classe felicidade com o valor 0,92 para a métrica (GRÁFICO 12).

Gráfico 12 – F1-Score fase teste



Fonte: Elaborado pelo autor (2024).

4.3 Experimentos

A fim de analisar a performance do classificador com a entrada de parâmetros MFCC extraídos de outros arquivos de áudio fora do contexto do dataset emoUERJ que foi utilizado para realizar o treinamento da rede neural convolucional, foram extraídos segmentos de áudios de vídeos do YouTube, todos eles de um canal de humor brasileiro com vídeos de tratando de diferentes temas interpretados por bons atores.

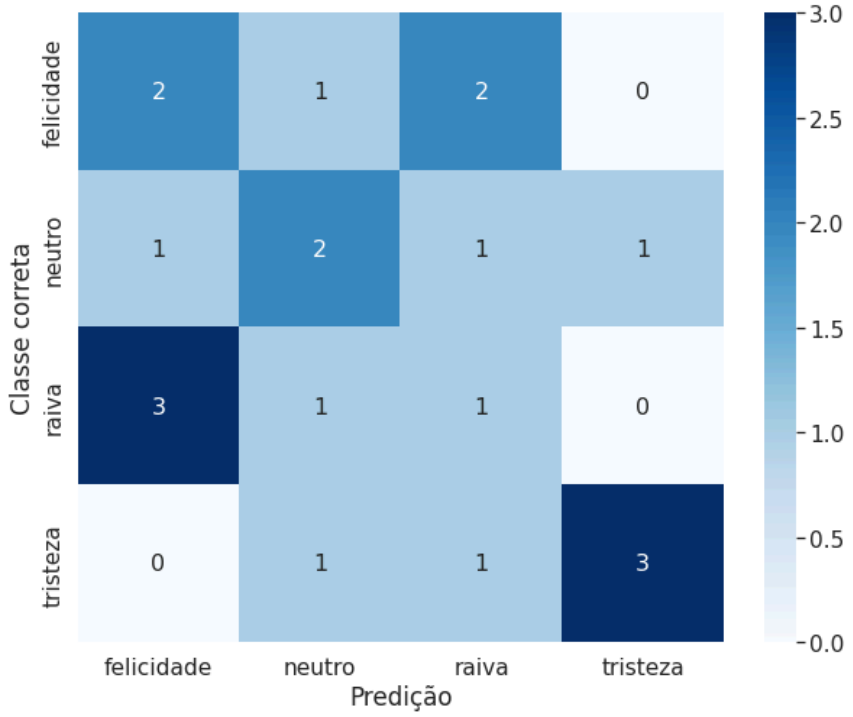
Para cada emoção que foi utilizada para realizar o treinamento da rede neural convolucional, foram coletados 5 trechos de áudios que as representavam em certo grau, avaliado pelo autor, as emoções avaliadas pela rede, no total 20 trechos de áudios com duração média de 11 segundos foram coletados.

Após realizar extração dos MFCCs, e a definição das classes de cada amostra, os parâmetros e as amostras foram submetidos à avaliação da rede, que resultou em uma acurácia de apenas 40%. O resultado mostra que a rede não conseguiu apresentar uma performance boa com os dados de um escopo fora dos dados de treinamento e teste provenientes da base de dados utilizada, caracterizando o overfitting.

Analisando a performance da CNN individualmente por classe, observando a matriz de confusão (FIGURA 14), temos que a classe que apresentou o resultado mais razoável, foi a classe tristeza, com 60% de classificação correta, ou seja, três das cinco amostras de áudios da classe tristeza foram classificadas como tal, enquanto as outras duas amostras foram classificadas como raiva e neutro, respectivamente.

Enquanto a classe que apresentou o pior resultado foi a classe raiva, com apenas 20% de acerto, ou seja, apenas uma das cinco amostras foi classificada corretamente. Já as outras duas classes apresentaram 40% de acerto com 2 classificações corretas cada.

Figura 14 – Matriz de confusão experimento antes do treino



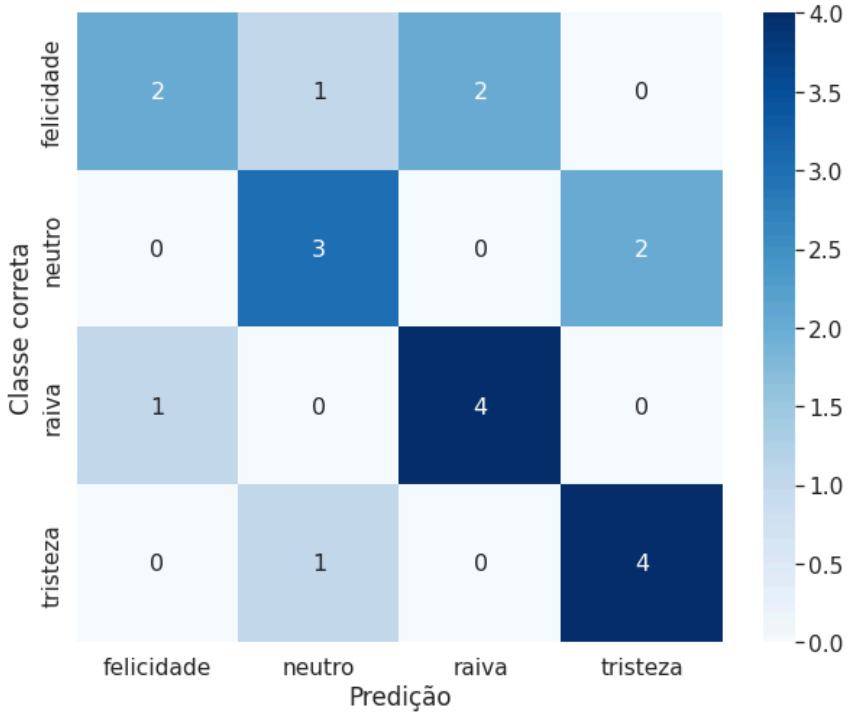
Fonte: Elaborado pelo autor (2024).

A fim de verificar se caso a CNN fosse agora treinada com os novos dados, a rede seria capaz de melhorar a acurácia na classificação desses dados, bem como, se faria com que a rede perdesse muita performance na classificação dos dados da base utilizada no treinamento, assim, prosseguiu-se com o procedimento do treino com os dados novos. Após o novo treinamento da rede neural, obteve-se uma nova acurácia de 65% na classificação dos áudios novos, já a acurácia obtida com a classificação dos dados de treinamento foi de 89,4%, uma piora na acurácia face a obtida na fase de treinamento inicial que foi de 93,73%.

Porém, o resultado pode ser considerado positivo, visto que aumentou a acurácia do classificador em 15% para na classificação dos dados novos, contra menos de 5% de perca na classificação dos dados de treinamento inicial, o que mostra que o modelo foi capaz de generalizar melhor o problema, podendo assim, melhorar sua performance na classificação de dados externos ao contexto de treinamento.

Avaliando a nova matriz de confusão (FIGURA 15) obtida com a classificação dos dados de áudios coletados através dos vídeos, temos uma melhora significativa na classe raiva, na nova classificação quatro das cinco amostras foram classificadas de forma correta, 80% das amostras da classe. As classes tristeza e neutro, também aumentaram as suas performances classificando uma amostra a mais de forma correta, com 80% e 60% de acerto, respectivamente. Já a classe felicidade apresentou os mesmos resultados na classificação.

Figura 15 – Matriz de confusão experimento depois do treino



Fonte: Elaborado pelo autor (2024).

5 CONCLUSÃO

Neste trabalho foi realizada uma pesquisa bibliográfica sobre as redes neurais convolucionais, bem como sobre os coeficientes cepstrais de frequência mel além das técnicas de classificação de emoções, que possibilitou a proposição de uma arquitetura de CNN assim como o seu desenvolvimento e teste.

Foi possível observar nos testes realizados com a rede, que devido ao fato dela ter sido treinada com uma quantidade de dados considerada pequena para a realização de treinamento de redes neurais convolucionais que são capazes de aprender e chegar a uma boa generalização a partir de treinamento com grandes quantidades de dados.

Assim, no processo de treinamento obteve-se uma acurácia de quase 94%, enquanto na fase de teste a acurácia de 95%. Porém, com o experimento realizado utilizando áudios extraídos de vídeos do YouTube a acurácia foi de apenas 40%. Portanto, a rede apresentou overfitting aos dados da base emoUERJ.

Após o novo treinamento do classificador com os novos dados, o modelo foi capaz de apresentar uma acurácia de 65% para os segmentos de áudios extraídos dos vídeos, e para os áudios utilizados no treinamento observou-se uma diminuição na acurácia, que foi 89%, apresentando uma melhora na generalização do modelo.

Apesar da sua simplicidade, o classificador pôde apresentar uma acurácia de 65%, como citado anteriormente, na classificação de áudios brutos extraídos de vídeos que apresentavam um contexto complexo, muito próximo ao que se teria em situações comuns do mundo real nas quais se aplicaria o classificador de emoções a fim de inferir a emoção predominante expressa nessas situações.

Porém, considerando que a base de dados utilizada para a realização do treinamento da rede neural convolucional foi pequena, bem como, não foi utilizada nenhuma técnica de tratamento para os áudios, como aplicação de ruído, deslocamento ou afinação, que poderiam melhorar o desempenho do classificador, nem foram utilizadas técnicas de aumento artificial da base de dados, pode-se concluir que a rede apresentou uma performance consideravelmente positiva.

Vista a importância do reconhecimento de emoções através da fala, além da variedade de suas possíveis aplicações, mesmo que tenha sido possível obter bons resultados neste trabalho, é possível de se conseguir resultados ainda mais robustos, dado que as redes neurais convolucionais tendem a apresentá-los quando treinadas a partir de grandes quantidades de dados, o que foi uma limitação na execução deste trabalho dada a restrição da capacidade de

processamento, bem como da disponibilidade de bases de áudio, principalmente de língua portuguesa.

As principais limitações encontradas no desenvolvimento desta aplicação foram justamente a quantidade limitada de bases de áudios para a realização do treinamento do modelo de aprendizagem de máquina, bem como a necessidade de um ambiente de desenvolvimento robusto, com uma boa capacidade de processamento e uma boa quantidade de memória para que seja possível de se realizar o treinamento das redes neurais utilizando grandes quantidades de dados em um espaço de tempo razoável.

Futuros trabalhos podem fazer a aplicação de técnicas de pré-processamento de áudio mais refinadas, fazer ainda a aplicação da técnica de data augmentation, a fim de incrementar a quantidade de dados com variações criadas artificialmente a partir dos dados existentes, bem como fazer o uso de outras bases de dados, inclusive em diferentes idiomas, para realizar o treinamento do modelo, o que pode culminar em uma generalização do problema ainda melhor e vir a apresentar uma performance ainda mais precisa e robusta.

Assim, conclui-se que a rede CNN desenvolvida, apesar da sua arquitetura simples, foi capaz de apresentar resultados consideravelmente positivos na classificação de emoções utilizando a técnica MFCC, mesmo tendo sido treinada a partir de uma resumida quantidade de dados. Portanto, as CNNs são propícias a apresentar resultados exitosos na classificação de emoções a partir da fala, reforçando a importância de pesquisas mais extensivas aplicando outras técnicas e ajustes e a relevância do tema.

REFERÊNCIAS

ALZUBI, J.; NAYYAR, A.; KUMAR, A. **Machine learning from theory to algorithms**: An overview. Journal of Physics: Conference Series, 2018. Disponível em: https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012/pdf. Acesso em: 21 fev. 2024.

BADR., Y.; MUKHERJEE., P.; THUMATI., S. Speech emotion recognition using mfcc and hybrid neural networks. In: INSTICC. **Proceedings of the 13th International Joint Conference on Computational Intelligence (IJCCI 2021) - NCTA**. [S.l.]: SciTePress, 2021. Disponível em: https://www.researchgate.net/publication/355880503_Speech_Emotion_Recognition_using_MFCC_and_Hybrid_Neural_Networks. Acesso em: 18 maio 2024.

CALHEIRO, M. S. **Classificação de anomalias em sons respiratórios utilizando processamento digital de sinais de áudio e redes neurais artificiais**. Monografia (Trabalho de conclusão de curso) — Escola Superior de Tecnologia da Universidade do Estado do Amazonas, 2021. Disponível em: https://repositorioinstitucional.uea.edu.br/bitstream/riuea/3730/1/Classifica%C3%A7%C3%A3o%20de%20anomalias%20em%20sons%20respirat%C3%B3rios%20utilizando%20processamento%20digital%20de%20sinais%20de%20%C3%A1udio%20e%20redes%20neurais%20artificiais.pdf. Acesso em: 21 maio 2024.

CAMPOS, G. A.; MOUTINHO, L. d. S. **DEEP: Uma arquitetura para reconhecer emoção com base no espectro sonoro da voz de falantes da língua portuguesa**. Monografia (Trabalho de conclusão de curso) - Departamento de Ciência da Computação, Universidade de Brasília, 2020. Disponível em: https://bdm.unb.br/bitstream/10483/27583/1/2020_GabrielCampos_LucasMoutinho_tcc.pdf. Acesso em: 27 abr. 2024.

FERREIRA, I. G.; ANDRADE, J. O.; KOMATI, K. S. Reconhecimento de emoção da fala: um estudo comparativo de classificadores. **Open Science Research IV**, 2022. Disponível em: https://downloads.editoracientifica.com.br/articles/220609157.pdf. Acesso em: 20 mar. 2024.

HELALI, W.; HAJAIEJ, ; CHERIF, A. Real time speech recognition based on pwp thresholding and mfcc using svm. **Engineering, Technology and Applied Science Research**, Greece, v. 10, 2020. Disponível em: https://etasr.com/index.php/ETASR/article/view/3759. Acesso em: 17 fev. 2024.

HILLESHEIN, H. **DESENVOLVIMENTO DE UM SISTEMA DE RECONHECIMENTO DE LOCUTOR UTILIZANDO APRENDIZADO DE MÁQUINA**. Monografia (Trabalho de conclusão de curso) — Curso de Engenharia de Telecomunicações, Instituto Federal de Santa Catarina, 2018. Disponível em: https://wiki.sj.ifsc.edu.br/images/1/17/TCC290_Henrique_Hillesheim.pdf. Acesso em: 19 mar. 2024.

GOMES JUNIOR, S. P. **Reconhecimento de Emoções em Sinais de Fala Usando Transferência de Aprendizado**. Dissertação (Mestrado em Sinais e Sistemas de Comunicações) — Programa de Pós-Graduação em Engenharia Eletrônica, da Universidade do Rio de Janeiro, 2019. Disponível em:

https://www.bdtd.uerj.br:8443/bitstream/1/11760/1/Sergio%20Pinto%20Gomes%20Junior_BDTD.pdf. Acesso em: 21 mar. 2024.

KANG, J. et al. Fault diagnosis of a wave energy converter gearbox based on an adam optimized cnn-lstm algorithm. **Renewable Energy**, 2024. Disponível em: https://www.sciencedirect.com/science/article/pii/S0960148124010905. Acesso em: 23 fev. 2024.

KHALIL, R. A. et al. Speech emotion recognition using deep learning techniques: A review. **IEEE Access**, 2019. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8805181. Acesso em: 29 abr. 2024.

KRICHEN, M. Convolutional neural networks: A survey. **Computers**, 2023. Disponível em: https://www.mdpi.com/2073-431X/12/8/151. Acesso em: 20 fev. 2024.

LEITE, D. R. A. **DESENVOLVIMENTO DE UM MODELO DE CLASSIFICAÇÃO DA TIPOLOGIA DOS SINAIS VOCAIS COM BASE NO DEEP LEARNING**. Tese (Pós-graduação em Modelos de Decisão e Saúde – Nível Doutorado) — Universidade Federal da Paraíba – UFPB, 2022. Disponível em: https://repositorio.ufpb.br/jspui/handle/123456789/25176. Acesso em: 21 mai. 2024.

LIBRALON, G. L. **Modelagem computacional para reconhecimento de emoções baseada na análise facial**. Tese (Doutorado em Ciências de Computação e Matemática Computacional) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2014. Disponível em: https://teses.usp.br/teses/disponiveis/55/55134/tde-10042015-104538/publico/teseDoutorado_GiampaoloLibralon _VersaoRevisada _bkp.pdf. Acesso em: 19 mar. 2024.

MENDOZA, L. A. F. **Redes Neurais e Máquinas de Vetores de Suporte no reconhecimento de locutor usando coeficientes MFC e características do sinal glotal**. Dissertação (Mestrado em Engenharia de Telecomunicações) — Universidade Federal Fluminense, 2009. Disponível em: http://app.uff.br/riuff/handle/1/31094. Acesso em: 17 mai. 2024.

MEZENCIO, R. **Aferição de Usabilidade de Interfaces Assistivas sob o Prismada Computação Afetiva**. Tese (Doutorado em Engenharia e de Computação) — Engenharia Elétrica e de Computação da Universidade Federal de Goiás, 2022. Disponível em: https://repositorio.bc.ufg.br/tedeserver/api/core/bitstreams/807f61e1-8338-4178-abdb-16de63625346/content. Acesso em: 29 abr. 2024.

MORAES, L. V. d. **Detecção de Depressão pela Fala Empregando Rede Neurais Profundas**. Dissertação (Mestre no Programa de Pós-Graduação em Ciência da Computação) — Instituto de Informática da Universidade Federal de Goiás, 2020. Disponível em: https://repositorio.bc.ufg.br/tedeserver/api/core/bitstreams/3095f224-b75e-429f-aeaf-c0444c4517fc/content. Acesso em: 21 mar. 2024.

NUNES, G. A. P. **Avaliação de sinais acústicos para o pré-diagnóstico de covid- 19**. Monografia (Graduação em Ciência da Computação) — UNESP, 2021. Disponível em:

https://repositorio.unesp.br/server/api/core/bitstreams/5b8f77e7-5ae8-4062-bc33-c10390efeb8e/content. Acesso em: 20 mai. 2024.

OTTONI, L. T. C.; OTTONI, A. L. C.; CERQUEIRA, J. d. J. F. A deep learning approach for speech emotion recognition optimization using meta-learning. **Electronics**, 2023. Disponível em: https://www.mdpi.com/2079-9292/12/23/4859. Acesso em: 19 fev. 2024.

PAN, W.; LI, H.; ZHOU, X.; JIAO, J.; ZHU, C.; ZHANG, Q. Research on Pig Sound Recognition Based on Deep Neural Network and Hidden Markov Models. **Sensors**, 2024. Disponível em: https://www.mdpi.com/1424-8220/24/4/1269. Acesso em: 10 nov. 2024.

PEIXOTO, G.; LINHARES, J. Reconhecimento de emoções através da fala utilizando rede neural convolucional. In: **Anais do L Seminário Integrado de Software e Hardware**. [s.n.], 2023. Disponível em: https://sol.sbc.org.br/index.php/semish/article/view/25067. Acesso em: 21 mar. 2024.

SANTOS, A.; REIS, V. Extração de características da fala para reconhecimento de emoções utilizando aprendizado de máquina. In: **V Seminário Gaúcho de Acústica e Vibrações**. [s.n.], 2020. Disponível em: https://www.researchgate.net/publication/356189665_Extracao_de_caracteristicas_da_fala_para_reconhecimento_de_emocoes_utilizando_a_e_Seminario_Gaucho_de_Acustica_e_Vibracoes. Acesso em: 21 mar. 2024.

SANTOS, V. M. **Reconhecimento de emoções atravás da fala utilizando redes neurais**. Monografia (Bacharel em Engenharia de Controle e Automação) — Universidade Estadual Paulista (Unesp), 2022. Disponível em: http://hdl.handle.net/11449/236555. Acesso em: 14 maio 2024.

SILVA, D. D. C. d. **Desenvolvimento de um IP core de pré-processamento digital de sinais de voz para aplicação em sistemas embutidos**. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Campina Grande, 2006. Disponível em: http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/1293. Acesso em: 10 mai. 2024.

SILVA, V. G. R. d. **Análise do sinal de fala para reconhecimento de emoções utilizando representação semântica**. Dissertação (Mestrado em Engenharia Elétrica) — Universidade Federal de Sergipe, 2022. Disponível em: https://ri.ufs.br/jspui/handle/riufs/17002. Acesso em: 13 mai. 2024.

SUGUINO, P. et al. Datasets reais para reconhecimento de emoção em voz. In: **Jornada de Iniciação Científica do CTI Renato Archer**. [s.n.], 2022. Disponível em: https://www.gov.br/cti/pt-br/publicacoes/producao-cientifica/jicc/xxiv-jicc-2022/pdf/jicc-2022-paper-22.pdf. Acesso em: 19 mar. 2024.

TAYE, M. M. Theoretical understanding of convolutional neural network: Concepts, architectures, applications, future directions. **Computation**, 2023. Disponível em: https://doi.org/10.3390/computation11030052. Acesso em: 21 fev. 2024.

VRECA, J.; PILIPOVIC, R.; BIASIZZO, A. Hardware–software co-design of an audio feature extraction pipeline for machine learning applications. **Electronics**, 2024. Disponível em: https://doi.org/10.3390/electronics13050875. Acesso em: 17 jul. 2024.

ZHAO, X. et al. A review of convolutional neural networks in computer vision. **Artificial Intelligence Review**, 2024. Disponível em: https://doi.org/10.1007/s10462-024-10721-6. Acesso em: 22 fev. 2024.