Detecção de Emoções Utilizando Áudio com Deep Learning

Speech Emotion Recognition with Deep Learning

Gabriel de Carvalho Vasconcelos[1]

Anderson Tenório Sérgio[2]

Resumo: O reconhecimento de emoções, embora ainda represente um desafio no contexto social, tem se consolidado como uma tarefa cada vez mais acessível a algoritmos de aprendizado profundo. Esse avanço tem viabilizado sua integração progressiva em aplicações voltadas à interação humano-máquina. Este trabalho investiga a detecção de emoções humanas a partir de sinais de áudio, utilizando arquiteturas de deep learning. Para isso, foi empregada a base de dados EmoDB, composta por 535 gravações distribuídas em sete categorias emocionais, organizadas em conjuntos de treinamento e validação. Foram avaliados três modelos distintos — RNN, CNN+LSTM e BiLSTM — com os hiperparâmetros otimizados por meio da ferramenta Optuna. O desempenho de cada arquitetura foi analisado com base nas métricas de precisão, recall, F1-score e acurácia média. Os resultados revelaram limitações nas redes unidirecionais, com precisão média de 21% para a RNN e 44% para a CNN+LSTM. Em contraste, o modelo BiLSTM apresentou desempenho superior, alcançando precisão média de 64% e recall de 61%. Também foi considerada a relação entre desempenho e custo computacional, bem como a quantidade de parâmetros envolvidos em cada modelo. Os achados indicam que o processamento bidirecional de dependências temporais contribui significativamente para a extração de padrões emocionais em sinais acústicos. Como principais limitações, destacam-se o tamanho reduzido do conjunto de dados e a simplicidade das arquiteturas utilizadas. Para pesquisas futuras, recomenda-se a ampliação da base de dados e a exploração de modelos mais profundos, com vistas à melhoria dos resultados obtidos.

Palavras-chave: Detecção de Emoções; Aprendizado profundo; Espectrogramas; Cepstrum de frequência mel (MFCC).

Abstract: *Emotion recognition, while still a challenge in social contexts, has become increasingly accessible to deep learning algorithms. This progress has enabled their growing integration into human–machine interaction applications. In this study, we investigate the detection of human emotions from audio signals using deep learning architectures. The EmoDB dataset was employed, comprising 535 recordings distributed across seven emotional categories, split into training and validation sets. Three models were evaluated — RNN, CNN+LSTM, and BiLSTM — with hyperparameters optimized using the Optuna framework. Model performance was assessed through precision, recall, F1-score, and average accuracy. Results showed limited effectiveness in unidirectional networks, with average precision of 21% for the RNN and 44% for the CNN+LSTM. In contrast, the BiLSTM model achieved superior results, with 64% average precision and 61% recall. Additionally, we examined the trade-offs between model performance, computational cost, and*

[1] Graduando no curso de Ciência da Computação

[2] Orientador e Professor no curso de Ciência da Computação

parameter count. Findings indicate that bidirectional processing of temporal dependencies significantly enhances emotional pattern extraction from audio signals. The main limitations of this study include the relatively small dataset and the simplicity of the architectures used. For future work, we recommend expanding the dataset and exploring deeper models to improve predictive performance.

Keywords: *Speech Emotion Recognition; Deep Learning; Spectrogram; Mel-frequency cepstral coefficients (MFCC).*

1 Introdução

Com o aumento do uso de sistemas automatizados para usuários — como assistentes virtuais, serviços de atendimento e monitoramento de saúde — torna-se cada vez mais importante interpretar não apenas o conteúdo verbal, mas também o estado emocional dos indivíduos. Nos últimos anos, o reconhecimento de emoções a partir do áudio tem ganhado destaque por tornar as interações entre humanos e máquinas mais naturais e eficazes (Ahlam et. al., 2023).

O *Deep Learning*, um subcampo da inteligência artificial, utiliza redes neurais artificiais com múltiplas camadas para aprender representações de dados em alto nível de abstração. Essas redes, inspiradas no cérebro humano, são capazes de extrair automaticamente padrões de grandes volumes de dados, eliminando a necessidade de extração manual de características. Essa capacidade é especialmente eficaz em tarefas como reconhecimento de fala, visão computacional e processamento de linguagem natural (Lecun et. al., 2015).

Este estudo insere-se nesse cenário. Investigamos o reconhecimento da fala por meio de redes neurais profundas, avaliando sua viabilidade e analisando arquiteturas voltadas para a classificação de emoções. Dessa maneira, esta pesquisa contribui para o avanço do conhecimento em processamento de áudio via *Deep Learning* e destaca tecnologias promissoras na área (Singh, 2023).

1.1 Justificativa

A sociedade atual exige sistemas inteligentes capazes de interpretar e responder adequadamente aos seres humanos, o que transforma a nossa forma de comunicação e atuação em diversos setores. Com o avanço da interação

homem–máquina, tornou-se comum o desenvolvimento de tecnologias que identificam emoções, a computação afetiva é o exemplo disso, a área da computação que busca aprimorar a experiência homem-máquina a partir do desenvolvimento de sistemas que entendem as emoções humanas (Morais et al., 2017).

Este tema apresenta relevância prática significativa. Em interações com assistentes virtuais, robôs ou plataformas de atendimento ao cliente, a adaptação ao estado emocional dos usuários promove uma experiência mais humana e personalizada, resultando em maior empatia e efetividade na comunicação (Ho et al., 2022) . No contexto de marketing e vendas, a análise emocional vocal orienta o discurso e as estratégias promocionais segundo o perfil emocional do consumidor, aumentando a eficácia das campanhas e o engajamento (Delija, 2024) . Por fim, em aplicações de segurança e vigilância, sistemas que detectam emoções em tempo real podem monitorar comportamentos de risco ou identificar situações de perigo proativamente, permitindo intervenções preventivas (Delija, 2024)

Portanto, implementar e aprimorar métodos de detecção emocional em áudio por meio de *Deep Learning* atende a demandas reais de empresas e usuários. Essas tecnologias solucionam problemas de comunicação, fortalecendo a relação interpessoal entre organizações e seus públicos. Ademais, tornam sistemas automatizados mais seguros e confiáveis ao incorporar percepções afetivas, especialmente em contextos sensíveis como saúde, emergências e suporte emocional (Rouast et al., 2019).

1.2 Problema

O entendimento das emoções sempre foi um problema importante para a sociedade, especialmente em contextos de interação humana e ambientes empresariais, portanto, muitas organizações têm dificuldade em interpretar corretamente o estado emocional do cliente, o que gera insatisfação e prejudica o relacionamento a longo prazo. Um sistema capaz de identificar emoções em áudio poderia ajudar a diagnosticar problemas em tempo real e melhorar a resposta ao cliente.

Em 2023, a psicologia identificou uma nova condição, a Alexitimia (Martins, 2023), que se refere à dificuldade de expressar emoções, isso reforça a importância

de detectar emoções de maneira eficaz. Ferramentas como o imentiv.ai (Imentiv, 2025) analisam imagens e classificam emoções conforme traços do rosto. No entanto, esta pesquisa concentra-se na detecção de emoções via áudio, ampliando as possibilidades de aplicação onde a captura de vídeo não é viável ou ideal.

Então, com o avanço da tecnologia, e, em destaque o avanço da tecnologia de *Deep Learning*, podemos pensar na construção deste problema e vislumbrar soluções para a problemática. Embora o desenvolvimento de uma solução demande recursos e seja complexo, é possível com os recursos que temos no momento da construção desta pesquisa.

Diante disso, surgem questões como:

- Como um modelo de *Deep Learning* com a função de detectar emoções a partir do áudio, pode ter alto desempenho, considerando a diversidade de contexto dos áudios?
- Quais os principais desafios na implementação e construção dessa tecnologia?
- Qual o impacto e a necessidade de uma solução para esse problema na saúde humana?

Essas perguntas mostram a importância e complexidade do tema, conduzindo à questão central da pesquisa:

- “Como podemos detectar emoções humanas de forma precisa a partir da captura de áudio utilizando de *Deep Learning* com o intuito de entender as emoções humanas?”

Assim, esta pergunta foca no principal conteúdo da pesquisa e na identificação da viabilidade da construção de um experimento mapeando os conteúdos acadêmicos existentes da área.

1.3 Objetivos

1.3.1 Objetivo Geral

Essa pesquisa busca, com o principal objetivo, analisar a viabilidade de uso de modelos de *Deep Learning* para classificação automática de emoções em áudio.

1.3.2 Objetivos específicos

- Identificar modelos de deep learning para detecção de emoções em áudio mais comuns na literatura.
- Avaliar os modelos quanto às métricas de precisão, *recall*, F1-score e acurácia média.
- Identificar as limitações dos modelos aplicados.

1.4 Estrutura do Artigo

O artigo está dividido em 5 capítulos organizados da seguinte forma: capítulo 1 traz uma introdução sobre o contexto do problema relatado, o objetivo da pesquisa, e onde o mesmo se insere; o capítulo 2 aborda uma fundamentação teórica sobre os tópicos relevantes e principais para a construção da base da pesquisa e o entendimento das informações posteriores; o capítulo 3 aborda a metodologia da pesquisa, com informações sobre a base de dados, modelos e métricas que foram avaliadas no contexto do artigo; o capítulo 4 traz uma análise de desempenho entre os modelos escolhidos e avaliados para esta pesquisa; e no capítulo 5 é apresentado a conclusão da pesquisa, trazendo uma reflexão do que foi desenvolvido e do que pode vir a ser proposto.

2 Fundamentação Teórica

2.1 Conceitos e Fundamentos de Reconhecimento de Emoções na Fala

O conceito de reconhecer emoções a partir da fala remonta às primeiras raízes da computação afetiva e do processamento de fala. Com o tempo, a evolução do Reconhecimento de Emoções na Fala foi impulsionada por avanços no aprendizado de máquina, no processamento de sinais e pela crescente

disponibilidade de conjuntos de dados diversificados para treinar modelos de reconhecimento de emoções. Esses avanços resultaram em maior precisão e robustez na detecção e classificação de emoções a partir da fala, inaugurando uma nova era de aplicações de IA emocionalmente inteligentes. (Akçay; Ogun, 2020)

O Reconhecimento de Emoções na Fala, também conhecido como SER (*Speech Emotion Recognition*), é o processo de identificar e analisar as emoções transmitidas na fala. Esse processo envolve o uso de algoritmos avançados para detectar e interpretar o conteúdo emocional subjacente na linguagem falada, contribuindo para o desenvolvimento de sistemas de IA emocionalmente inteligentes. No contexto da inteligência artificial, o reconhecimento de emoções na fala é essencial para conferir uma compreensão emocional semelhante à humana às máquinas, permitindo que elas compreendam, interpretem e respondam melhor às emoções humanas. No âmbito da computação o Reconhecimento de Emoções na Fala é um subcampo da computação afetiva que se concentra em reconhecer emoções a partir de sinais vocais. Através da extração de características distintas, como tom, intensidade e taxa de fala, o SER utiliza técnicas de aprendizado de máquina para categorizar emoções — incluindo, mas não se limitando a, felicidade, tristeza, raiva e medo — a partir do conteúdo falado. (Kumar; Raj, 2022)

2.2 Pipeline de Processamento de Áudio para SER

2.2.1 Extração de Características

Para entendermos o funcionamento de uma aplicação de inteligência artificial que tem a capacidade de detectar emoções, primeiro, precisamos entender como os dados que vamos utilizar para treinar as IAs funcionam. Neste caso, utilizamos um dado não estruturado, o áudio. Inicialmente precisamos extrair as características de um áudio utilizando *Fourier Transform*. Técnicas como *Short-time Fourier Transform (STFT)*, *Fast Fourier Transform (FFT) e Discrete Fourier Transform (DFT)* são utilizadas para extrair essas características, mas o que seriam essas técnicas. A Transformada de Fourier (Fourier Transform) é reconhecida como uma das ferramentas mais tradicionais e importantes no processamento de sinais em diversas áreas científicas, especialmente eficaz para a análise de sinais periódicos. Essa técnica permite decompor um sinal, seja contínuo ou discreto, em uma combinação

de ondas senoidais com diferentes amplitudes e frequências, facilitando a análise em domínio de frequência (Zheng; Yue, 2021). A transformada de Fourier decompõe um sinal em uma combinação de ondas senoidais com diferentes amplitudes e frequências. Na Figura 1 temos um exemplo da transformada de Fourier discreta (Kumar; Raj, 2022).

Figura 1 - Discrete Fourier Transform

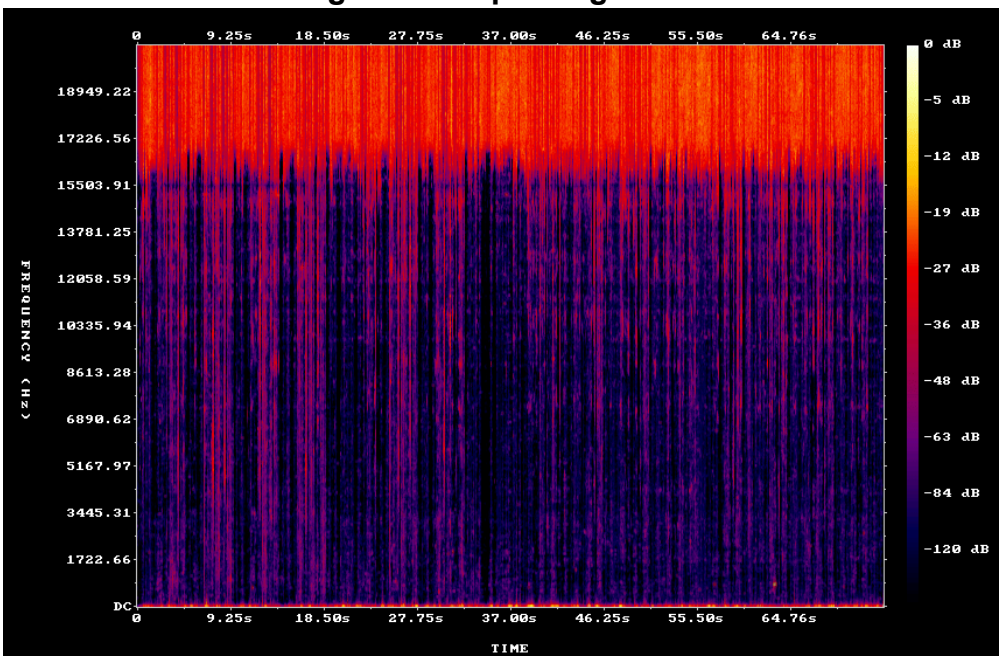
$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{2\pi jnK}{N}} \quad k = 1, 2, 3 \ldots N-1$$

FONTE: Zrar Kh, A; Abdulbasit, K, A (2022)

2.2.2 Espectrogramas

Um espectrograma é uma representação visual do espectro de frequências de um sinal conforme ele varia com o tempo (Figura 2). Quando aplicado a um sinal de áudio, os espectrogramas às vezes são chamados de sonogramas ou impressões vocais. Espectrogramas são amplamente usados nas áreas de música, linguística, sonar, radar, processamento de fala, sismologia e muito mais. Espectrogramas de áudio podem ser usados para identificar palavras faladas foneticamente e para analisar os vários chamados de animais (Olman, C et. al., 2022).

Figura 2 - Espectrogramas

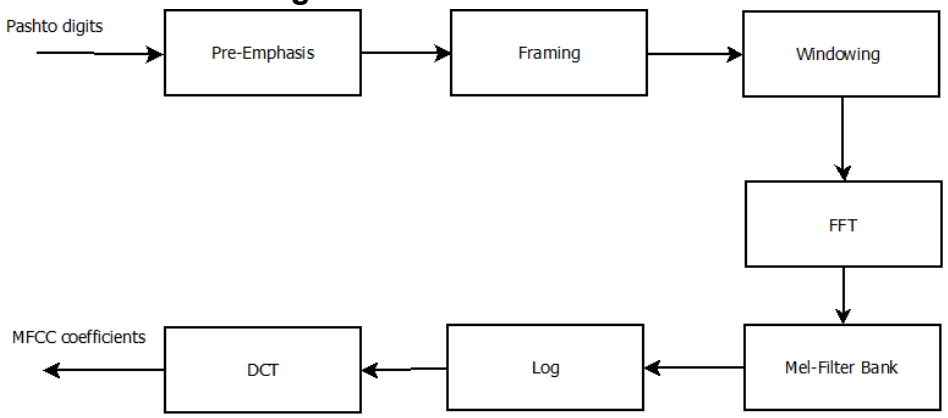


FONTE: Blog AI Mastering explicando sobre 'Limiter error spectrogram' (2024)

2.2.3 *Mel Spectrogram - Mel Frequency Cepstral Coefficients* (MFCCs)

O MFCC (*Mel Frequency Cepstral Coefficients*) é uma das características comumente utilizadas em várias aplicações, especialmente no processamento de sinais de voz e extração de *features*, como reconhecimento de locutor, reconhecimento de voz e identificação de gênero. O cálculo do MFCC envolve cinco processos consecutivos (Figura 3): a segmentação do sinal em quadros, o cálculo do espectro de potência, a aplicação de um banco de filtros Mel aos espectros de potência obtidos, o cálculo dos valores logarítmicos de todos os filtros e, por fim, a aplicação da Transformada Discreta de Cosseno (DCT) (Abdul et. al., 2022).

Figura 3 - *MFCC Framework*



FONTE: Nisar, et. al. (2017)

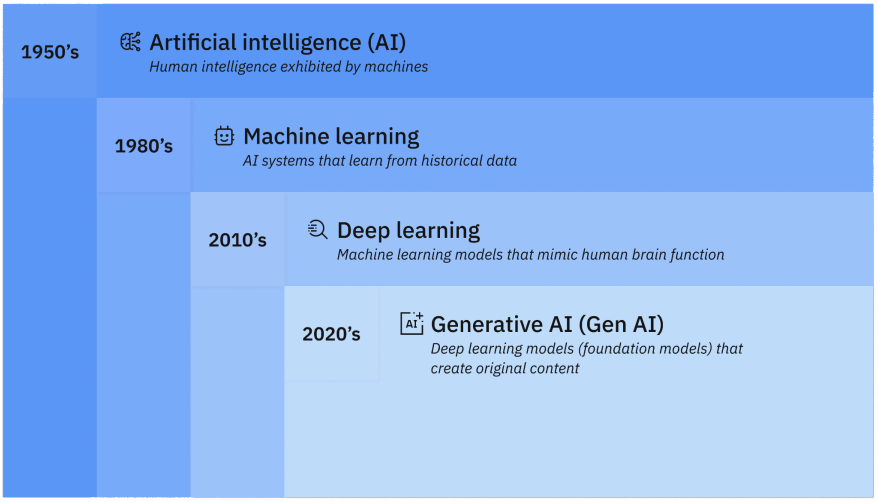
2.3 Deep Learning

Deep Learning, ou aprendizado profundo, é uma subárea do aprendizado de máquina que se baseia em redes neurais artificiais com muitas camadas (Figura 4). As redes neurais são inspiradas no funcionamento do cérebro humano, onde cada neurônio é uma unidade de processamento que realiza operações matemáticas sobre os dados de entrada. No contexto de redes neurais, o Deep Learning permite que os modelos aprendam representações complexas de dados (LeCun et al., 2015).

O campo do *Deep Learning* evoluiu significativamente desde as primeiras redes neurais simples. Com o aumento da capacidade computacional e a disponibilidade de grandes volumes de dados, tornou-se possível treinar redes neurais muito profundas e complexas. Inicialmente, técnicas como redes neurais

convolucionais (CNNs) foram aplicadas principalmente ao reconhecimento de imagens (Krizhevsky et al., 2012).

Figura 4 - Inteligência Artificial



FONTE: *What is Artificial Intelligence* escrito no site da IBM (2024)

2.4 Arquiteturas de Redes Neurais para Reconhecimento de Emoções na Fala

Atualmente, os classificadores de Reconhecimento de Emoções na Fala (SER) podem ser divididos em dois tipos: classificadores tradicionais e classificadores de aprendizado profundo, que será visto nos tópicos posteriores. Classificadores tradicionais incluem *Gaussian Mixture Models* (GMMs), *Hidden Markov Models* (HMMs) e *Support Vector Machine* (SVM) (Khalil et. al., 2019), que dependem de um extenso pré-processamento e de *feature engineering* precisa. Com o avanço da tecnologia de aprendizado profundo, o desempenho dos sistemas SER melhorou significativamente, portanto, a maioria dos frameworks de reconhecimento baseados em aprendizado profundo utiliza *Convolutional Neural Networks* (CNNs), *Long-Short Term Memory* (LSTM) e suas combinações (Khalil et. al., 2019).

2.4.1 *Convolutional Neural Network* (CNN)

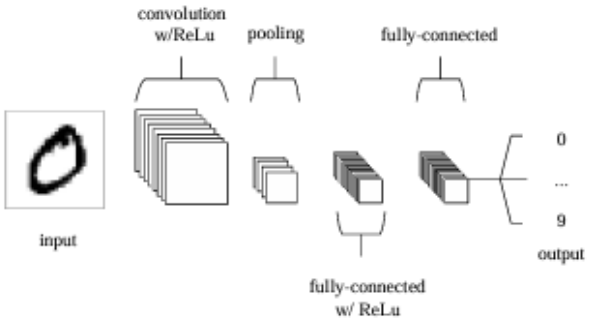
As Redes Neurais Convolucionais **(**CNN**)** são baseadas na combinação hierárquica de três tipos principais de camadas: *convolucionais*, de *pooling* e *fully-connected.* Quando empilhadas, essas camadas permitem a extração gradual de recursos visuais, culminando em decisões de classificação complexas (Yamashita et. al., 2018).

As camadas convolucionais utilizam filtros (ou kernels) que deslizam sobre o volume de entrada — tipicamente uma imagem ou espectrograma — aplicando operações de convolução para gerar mapas de ativação. Esses filtros aprendem automaticamente características locais, como bordas ou texturas, conferindo à rede capacidade de percepção translacional invariável (Yamashita et. al., 2018).

As camadas de *pooling* reduzem a dimensão espacial dos mapas de ativação, preservando apenas as características mais salientes. Essa operação reduz o número de parâmetros, diminui custos computacionais e atenua o risco de overfitting, além de consolidar a invariância a pequenas variações posicionais (Yamashita et. al., 2018).

Já as camadas *fully-connected* recebem o mapa de ativação achatado (flattened) e atuam como uma rede densa que realiza a composição de alto nível dos recursos extraídos. Essas camadas finais são responsáveis por gerar resultados das classes, geralmente através de uma ativação softmax, consolidando o processo de classificação (Yamashita et. al., 2018). Uma arquitetura simplificada de CNN está ilustrada na Figura 5.

Figura 5 - *CNN Architecture on MNIST Dataset*



FONTE: Keiron, O; Ryan, N (2015)

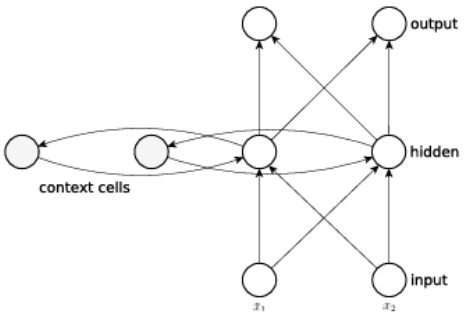
2.4.2 *Recurrent Neural Network (RNN)*

As Redes Neurais Recorrentes (RNN) possuem um estado interno em cada etapa da classificação. Isso ocorre devido às conexões recursivas entre os neurônios das camadas superiores e inferiores, além de conexões de auto-feedback. Essas conexões de feedback permitem que as RNNs propaguem dados de eventos

anteriores para as etapas de processamento atuais. Assim, as RNNs constroem uma memória de eventos em séries temporais (Geron, 2019).

As RNNs variam de parcialmente a totalmente conectadas, um dos modelos simples de RNNs é a rede de Elman, que é semelhante a uma rede neural de três camadas, mas, adicionalmente, as saídas da camada oculta são salvas em células de contexto (*context cells*). A saída de uma célula de contexto é circularmente realimentada para o neurônio oculto, juntamente com o sinal original. Cada neurônio oculto possui sua própria célula de contexto e recebe entrada tanto da camada de entrada quanto das células de contexto (Geron, 2019). A Figura 6 mostra uma RNN como exemplo.

Figura 6 - *RNN Architecture - Elman Network*

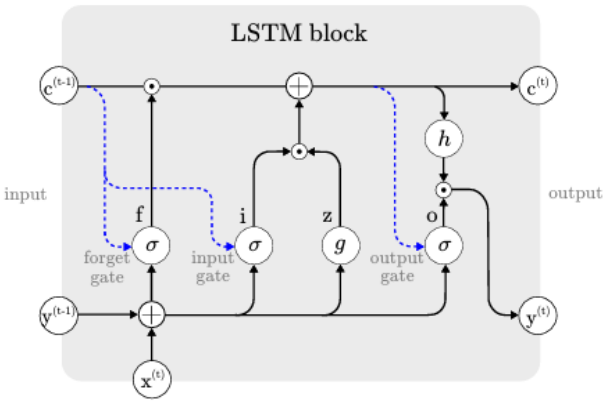


FONTE: Understanding LSTM, a tutorial into LSTM Recurrent Neural Networks (2024)

2.4.3 *Long-Short Term Memory* (LSTM)

O LSTM (Long Short-Term Memory) é um modelo avançado de redes neurais recorrentes, desenvolvido para superar os problemas do *exploding gradient* e do *vanishing gradient*, que são comuns no aprendizado de dependências temporais de longo prazo. Uma rede LSTM básica é composta por uma célula de memória e três “portões” (gates), o input gate, o output gate e o forget gate, que regulam o fluxo de informações na rede e permitem que ela mantenha, atualize ou esqueça informações relevantes. A célula é responsável por armazenar valores por períodos de tempo arbitrários, enquanto os gates controlam o ingresso, a saída e a retenção dessas informações. Essencialmente, a arquitetura do LSTM é formada por blocos de memória, que são sub-redes conectadas de forma recorrente, projetadas para preservar estados ao longo do tempo. (Hochreiter; Schmidhuber, 1997).

Figura 7 - *LSTM Block Architecture*



FONTE: LSTM Block Architecture (2019)

2.5 Classificação de Emoções

Para classificar emoções corretamente, é necessário uma base de dados que contém informações atreladas a uma emoção, por exemplo, caso seja preciso categorizar corretamente uma emoção, sendo, raiva ou felicidade, no contexto de áudio, é necessário uma base de dados que possua áudios que estejam relacionados com cada uma dessas emoções, para que, o modelo determinado consiga classificar o áudio que está recebendo corretamente, então, a partir desta base, é possível classificar qualquer emoção, desde que, exista uma base para ser injetada no modelo, para que a classificação seja possível. (Kumar et al., 2023)

3 Metodologia

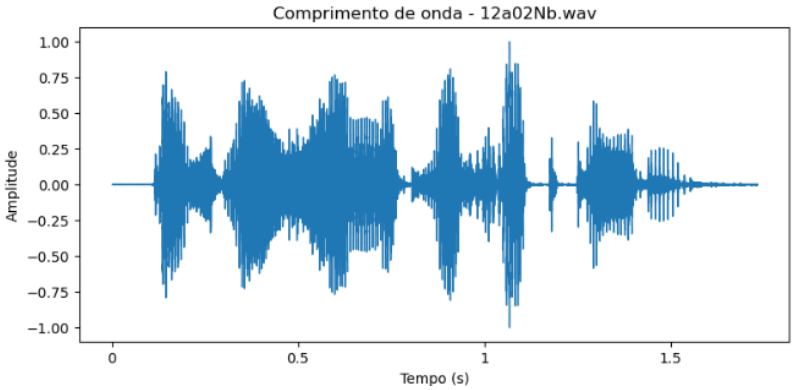
Este estudo centra-se nas técnicas de aprendizagem por máquina com aprendizado profundo para detecção de emoções a partir do áudio, selecionadas com base em sua relevância ao tema, *benchmarks* recentes e na inovação nos métodos de treinamento e detecção. A natureza da pesquisa é aplicada, pois o foco é solucionar problemas práticos de detecção de emoções com dados limitados utilizando exemplos das tecnologias mais recentes. A abordagem é quantitativa, utilizando dados numéricos para avaliar o desempenho dos modelos. Os objetivos são exploratórios e explicativos, visando descobrir e avaliar percepções em detecção de emoções a partir do áudio e explicar eficácia dessas técnicas.

3.1 Base de Dados

A base de dados central para este estudo consiste em áudios na língua alemã e que foram gravados em de 48-kHz e *down-sampled* para 16-kHz. Esses áudios fazem parte do *dataset* EmoDB, um dataset público amplamente utilizado para aplicações de reconhecimento de emoções a partir do áudio (Khalil et al., 2019). Essa coleção é caracterizada por uma diversidade significativa de áudios nas mais diversas emoções. Os áudios foram coletados pelo Instituto de Ciências da Comunicação da Universidade Técnica da cidade de Berlin, Alemanha, onde, 10 palestrantes profissionais, 5 são do sexo masculino e 5 são do sexo feminino. Os palestrantes não são identificados, porém suas gravações são públicas, podendo ser utilizadas para o fim desta pesquisa (Burkhardt, et al., 2005).

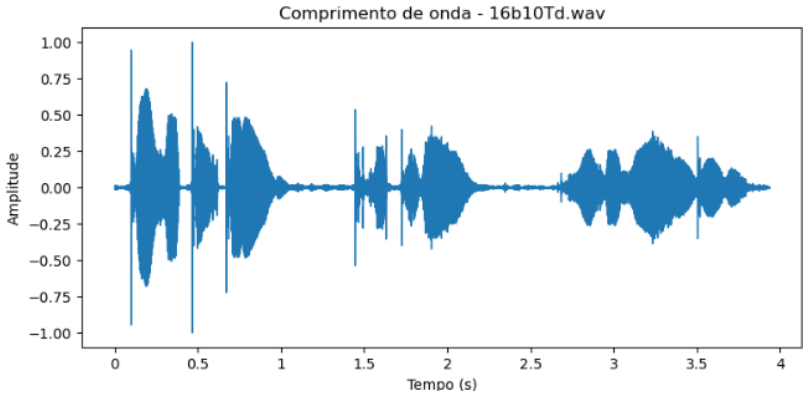
A base de dados é composta de 535 áudios que são divididos em 7 classes, onde cada classe corresponde a uma emoção, são as seguintes: Felicidade, Neutro, Raiva, Tristeza, Ansiedade, Tédio, Desgosto. Então, os dados da base foram divididos para treino e para validação, onde a base de treinamento é composta por um total de 454 áudios e a base de validação é composta por 81 áudios. Como podemos ver na Figura 8, temos uma diferença em como esses áudios são construídos a partir da variância na amplitude durante o comprimento de onda. A diferença no número de áudios na base de treino em comparação com a base de teste, visa proporcionar um equilíbrio para os modelos considerando o número de dados que possuímos na base. (Burkhardt, et al., 2005)

Figura 8 - Exemplo de áudio emoção neutra



FONTE: Elaborado pelo autor

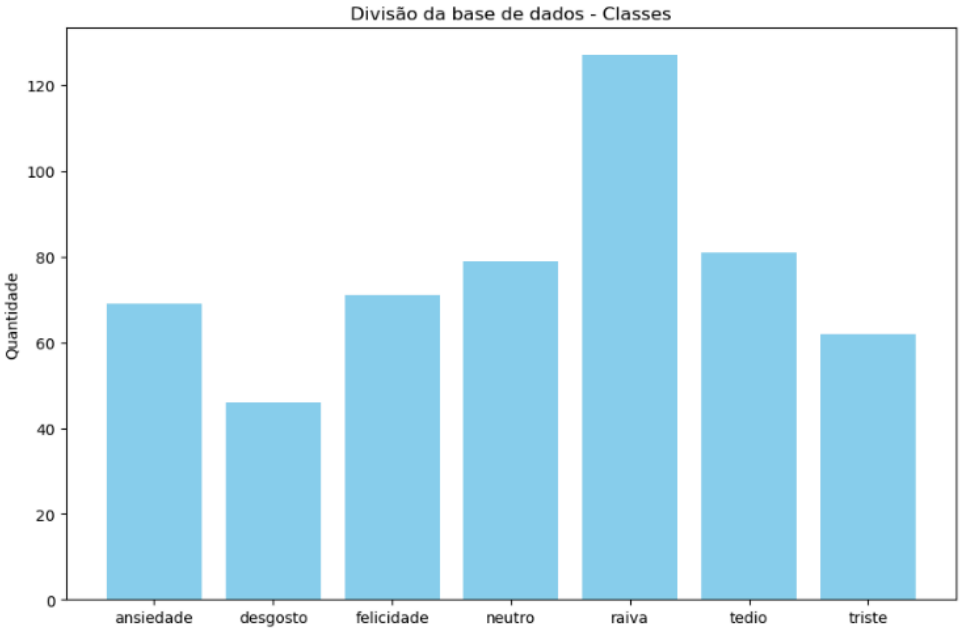
Figura 9 - Exemplo de áudio emoção tristeza



FONTE: Elaborado pelo autor

Como o banco de dados deste estudo contém uma gama diversificada de categorias, como podemos ver na Figura 10, a estratégia adotada para este estudo e para o treinamento foi conseguir diferenciar da melhor maneira cada áudio, visto o problema multi classificatório enfrentado. Sendo assim, como vemos na Figura, as classes estão desbalanceadas, possuindo mais dados na classe raiva se comparado com as outras classes, esse desbalanceamento pode alterar os resultados obtidos de forma enviesada. Portanto, a pesquisa concentra-se em desenvolver e otimizar os modelos, com o objetivo de classificar com precisão cada áudio.

Figura 10 - Divisão da base de dados em cada classe



FONTE: Elaborado pelo autor

3.2 Modelos

A escolha dos modelos listados abaixo para a detecção de emoções utilizando áudio é justificada por diversos motivos. Primeiramente, esses modelos demonstraram resultados positivos em *benchmarks* relevantes, como descrito em Meng et al., (2019) e Khalil et al., (2019), indicando uma eficácia notável em diferentes cenários e quantidades de dados. Além disso, por serem modelos com o adjetivo de estado da arte para esse contexto, fazem sua utilização ser algo natural. Outra questão envolvendo esses modelos são sua documentação bem organizada, facilitando o entendimento e replicação dos resultados, um aspecto crucial para a pesquisa e desenvolvimento se tratando da área de *Deep Learning* e derivados.

Por outro lado, a decisão de limitar a escolha a apenas três modelos foi influenciada pela necessidade de otimizar o tempo e os recursos disponíveis para a pesquisa. Os modelos selecionados parecem ser os mais otimizados em termos de desempenho e eficiência, oferecendo um equilíbrio ideal entre a precisão e a viabilidade prática. A escolha de modelos com uma combinação de alta performance e documentação acessível permite focar em abordagens que prometem resultados tangíveis, maximizando o retorno sobre o investimento em termos de tempo e recursos dedicados à pesquisa (Islam et al, 2018; Khalil et al, 2019). Com base nas necessidades da pesquisa, os seguintes modelos foram escolhidos para serem avaliados:

3.2.1 Informações do Treinamento

Para este treinamento, foi definido a padronização das características de treinamento, ou seja, foi usado os mesmos parâmetros de treinamento para todos os modelos, tendo em vista que seria realizado uma comparação de resultados, se é necessário a padronização de parâmetros para que ocorra a padronização dos resultados.

Então, foi definido que todos os modelos fossem treinados com os parâmetros de 40 épocas em todos os algoritmos, também, com a utilização do otimizador Optuna (Akiba et al., 2019) que é uma ferramenta para otimização de treinamento de modelos, com a escolha de 30 iterações de estudo, para otimização dos hiperparâmetros de todos os modelos, essas escolhas se deram no artigo aqui apresentado com o objetivo de encontrar um equilíbrio entre ter uma quantidade

representativa de iterações entre os modelos e otimizar o seu desempenho. Sendo assim, buscamos evitar o risco de sobrecarregar os modelos com os mesmos dados que já foram passados para treinamento, o que poderia levar ao *overfitting*, especialmente no nosso cenário já que não possuímos uma quantidade tão extensa de dados. Portanto, o objetivo é que não seja gerado problemas em nosso treinamento a ponto de alterar os resultados.

3.2.2 *Recurrent Neural Network* (RNN)

Para avaliar a capacidade de captura de dependências temporais em sinais de áudio, foi implementada uma RNN simples utilizando *Keras-TensorFlow*. Todos os áudios foram amostrados a 16 kHz, normalizados e transformados em coeficientes MFCC. A arquitetura consiste em: camada de entrada que recebe tensores (max_len, n_mfcc); camada SimpleRNN com U unidades, camada *Dropout* com taxa D e camada *Dense* final com ativação *softmax* e sete neurônios. O modelo foi treinado por até 40 épocas, *batch size* de 32 e *Early Stopping* monitorando acurácia de validação com critério de paciência de valor 3, restaurando os melhores pesos. A função de perda foi a *categorical cross entropy* com o otimizador *Adam*.

Para otimização de parâmetros foi utilizado o *Optuna* (Akiba et al., 2019) em 30 iterações, otimizando: número de unidades da RNN (U entre 32 e 256), taxa de dropout (D entre 0,1 e 0,5) e taxa de aprendizado (lr entre 10^{-4} e 10^{-2}, distribuição logarítmica). Cada combinação era avaliada pela acurácia máxima em validação, retornada ao final de cada iteração.

3.2.3 *Convolutional Neural Network + Long Short-Term Memory Layer* (CNN + LSTM)

Para explorar a capacidade de extração de características espaciais do aprendizado temporal, implementou-se um modelo híbrido de *Convolutional Neural Network + Long Short-Term Memory* com *Keras-Tensorflow*. Os dados foram padronizados da mesma forma descrita para a RNN. A sequência de camadas é: camada de entrada (n_mfcc, max_len, 1); Conv2D com f1 filtros 3×3, ReLU e *padding* “*same*”; MaxPooling2D 2×2; Conv2D com f2 filtros 3×3, ReLU e padding “*same*”; MaxPooling2D 2×2; *Reshape* para (timesteps, f2), onde *timesteps* = (altura × largura) pós-pooling; camada LSTM com u_lstm unidades; camada *Dropout* com

taxa D; camada *Dense* final com ativação *softmax* e sete neurônios. O treinamento foi executado com 40 épocas, *batch size* 32 e *Early Stopping* monitorando a acurácia de validação com critério de paciência de valor 3.

Na busca de hiperparâmetros pelo *Optuna* (Akiba et al., 2019), em 30 iterações foram ajustados: número de filtros na primeira convolução (f1 entre 16 e 64), número de filtros na segunda convolução (f2 entre 32 e 128), unidades LSTM (u_lstm entre 32 e 256), taxa de *Dropout* (D entre 0,1 e 0,5) e taxa de aprendizado (lr entre 10^{-4} e 10^{-2}, logarítmico). Cada iteração compilava o modelo com *categorical cross entropy* e otimizador Adam. Cada combinação era avaliada pela acurácia máxima em validação, retornada ao final de cada iteração.

3.2.4 *Bidirectional Long Short-Term Memory* (BiLSTM)

Para capturar contexto em ambas as direções temporais, utilizou-se uma arquitetura BiLSTM construida com *Keras-Tensorflow*. Os coeficientes MFCC foram preparados da mesma forma que na RNN. A arquitetura inclui: camada de entrada (max_len, n_mfcc); camada *Bidirectional* composta por duas subcamadas LSTM (cada uma com U unidades), cujas saídas ocultas são concatenadas em vetor de dimensão 2U; camada *Dropout* com taxa D; camada *Dense* final com *softmax* e sete neurônios. O treinamento seguiu os mesmos parâmetros fixos, 40 épocas, *batch size* 32 e *Early Stopping* em acurácia de validação com critério de paciência de valor 3. A função de perda utilizada foi a *categorical cross entropy*, juntamente com o otimizador *Adam*.

O *Optuna* (Akiba et al., 2019) otimizou em 30 iterações: número de unidades LSTM em cada direção (U entre 32 e 256), taxa de dropout (D entre 0,1 e 0,5) e taxa de aprendizado (LR entre 10^{-4} e 10^{-2}, logarítmico). Em cada iteração, o modelo era compilado, treinado até 40 épocas ou até o *Early Stopping*, com critério de paciência de valor 3, e avaliado pela acurácia de validação.

3.3 Métricas para Avaliação dos Modelos

Para atingir o objetivo proposto da pesquisa o principal critério avaliativo dos modelos é desempenhar de forma similar ao que é demonstrado no estado da arte

(Kumar et. al., 2023). Como temos na Tabela 1, os resultados de um modelo CNN+LSTM na base de dados EmoDB, devemos propor que o objetivo da pesquisa seja atingir resultados próximos com os três modelos escolhidos para comparação.

No contexto da pesquisa, as métricas *precision* (precisão), *recall* (revocação) e *F1-score* são fundamentais para avaliar o desempenho dos modelos. A *precision* quantifica a proporção de previsões positivas corretas em relação ao total de previsões positivas realizadas pelo modelo, ou seja, mede a exatidão das classificações positivas. Já o *recall* avalia a capacidade do modelo em identificar corretamente todas as instâncias positivas reais, indicando a abrangência na detecção de casos positivos. O *F1-score* é a média harmônica entre *precision* e *recall*, proporcionando uma medida única que equilibra ambas as métricas, sendo especialmente útil quando há desequilíbrio entre as classes no conjunto de dados (Powers, 2020).

TABELA 1 - Classification Report *CNN+LSTM*

	Precision	Recall	F1-Score	Support
Anger	0.92	0.98	0.95	124
Happiness	0.74	0.95	0.83	73
Sadness	0.92	0.84	0.88	93
Neutral	0.95	0.65	0.77	65
accuracy			**0.88**	355
macro avg	0.88	0.85	0.86	355
weighted avg	0.89	0.88	0.87	355

FONTE: Kumar et al. (2023)

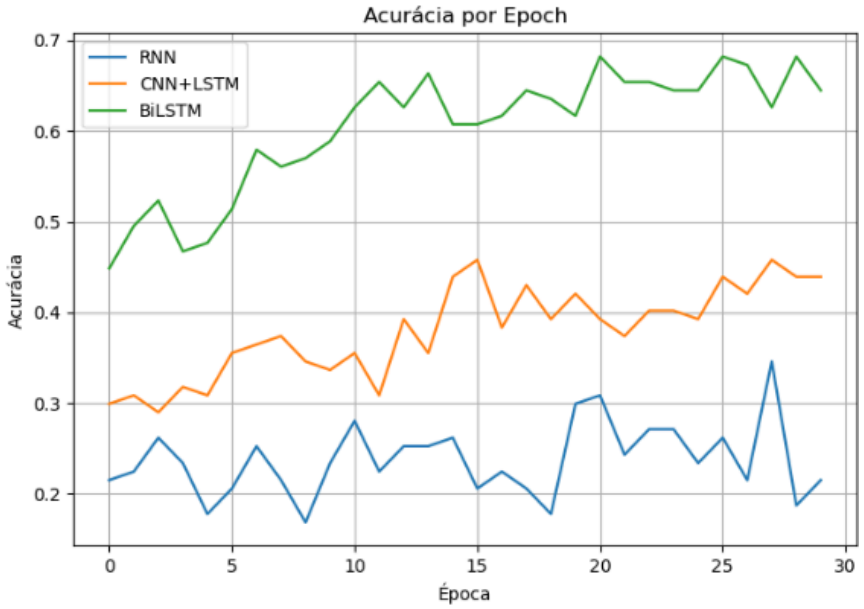
4 Análise dos Resultados

4.1 Resultado dos Treinamentos

Para avaliar quão bons foram os resultados obtidos pelos modelos, foram avaliadas as métricas descritas no capítulo anterior. Estas métricas, fornecem uma compreensão abrangente do desempenho dos modelos em tarefas de predição, portanto, são cruciais para a aplicação desta pesquisa.

O gráfico da Figura 11, as Tabelas 2, 3 e 4, e as Figuras 12, 13 e 14 são representações visuais das métricas propostas para os modelos: RNN; CNN+LSTM e BiLSTM, obtidas após treinar os modelos por 40 épocas. Estes gráficos fornecem uma visão intuitiva do comportamento e desempenho de cada modelo ao longo das iterações, destacando tendências, consistências e variações nos resultados. A visualização dessas métricas em gráficos e tabelas facilita a comparação direta entre os modelos e ajuda a identificar áreas onde cada modelo se destaca ou necessita de otimizações.

Figura 11 - Acurácia entre os modelos por Época na base de treino



FONTE: Elaborado pelo autor

Analisando as métricas de acurácia para os três modelos mostrados na Figura 11, observamos que os modelos não possuem uma acurácia tão alta, somente se avaliarmos os modelos entre si, com médias próximas a 70%, 50% e 30% para os modelos BiLSTM, CNN+LSTM e RNN respectivamente. Os resultados apresentados sugerem desafios enfrentados pelos modelos, possivelmente relacionados às características dos dados ou as técnicas de modelagem utilizadas, com ênfase neste último que possui um grande espaço para otimização com foco na expansão da arquitetura dos modelos. Como previsto, é identificado pelo gráfico uma tendência clara de melhoria na precisão ao longo das 40 iterações, indicando que a cada iteração os modelos melhoram seu desempenho e que podem a vir

melhorar com o aumento no número de iterações.

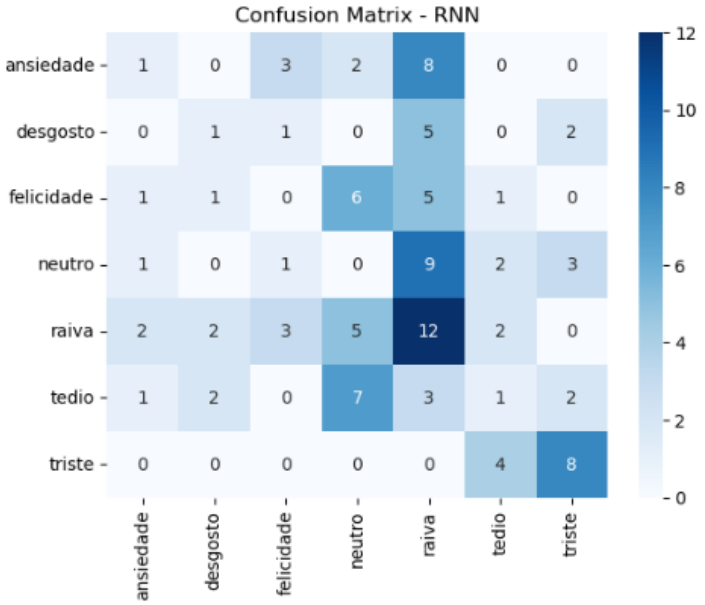
TABELA 2 - Classification Report RNN

Classe	Precision	Recall	F1-Score	Suporte
ansiedade	0.17	0.07	0.1	14
desgosto	0.17	0.11	0.13	9
felicidade	0.0	0.0	0.0	14
neutro	0.0	0.0	0.0	16
raiva	0.29	0.46	0.35	26
tedio	0.1	0.06	0.08	16
triste	0.53	0.67	0.59	12
accuracy			0.21	107
macro avg	0.18	0.2	0.18	107
weighted avg	0.18	0.21	0.19	107

FONTE: Elaborado pelo autor (2025)

Com relação às métricas de precisão, recall e f1-score para o modelo RNN, a Tabela 2 mostra esses dados relacionados para cada classe da base de dados de emoções, demonstrando principalmente valores baixos de precisão relacionados a algumas classes em específico, como as emoções de neutro e felicidade que tiveram como resultado a precisão de 0%, um resultado bastante negativo. Observamos que, possivelmente o modelo pode estar classificando essas emoções como outra classe, visto a discrepância entre as emoções de tristeza e tédio se comparado com as outras. Na Figura 12, temos um gráfico da matriz de confusão com mais informações sobre as predições do modelo e as predições verdadeiras em cima das classes da base de dados EmoDB.

Figura 12 - Confusion Matrix RNN



FONTE: Elaborado pelo autor (2025)

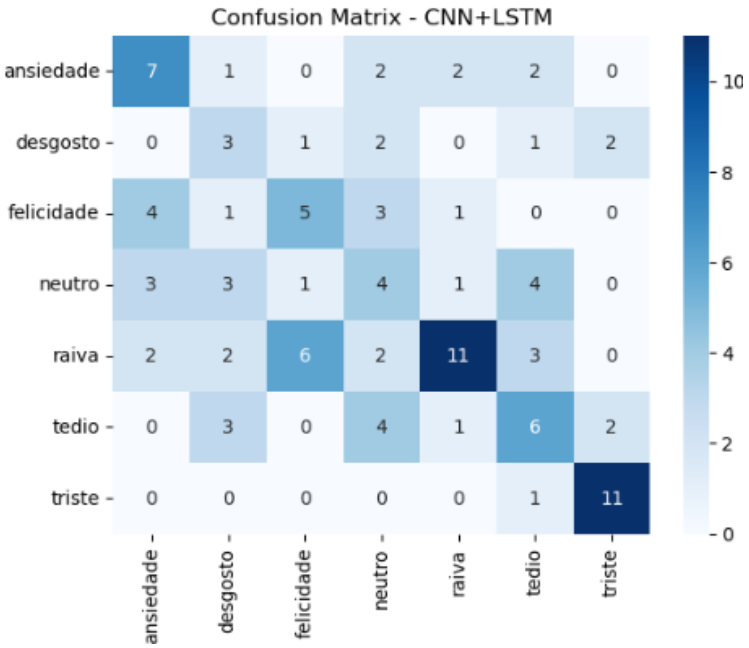
TABELA 3 - Classification Report CNN-LSTM

Classe	Precision	Recall	F1-Score	Suporte
ansiedade	0.44	0.5	0.47	14
desgosto	0.23	0.33	0.27	9
felicidade	0.38	0.36	0.37	14
neutro	0.24	0.25	0.24	16
raiva	0.69	0.42	0.52	26
tedio	0.35	0.38	0.36	16
triste	0.73	0.92	0.81	12
accuracy			0.44	107
macro avg	0.44	0.45	0.44	107
weighted avg	0.46	0.44	0.44	107

FONTE: Elaborado pelo autor (2025)

Com relação às métricas de precisão, recall e f1-score para o modelo CNN-LSTM, a Figura 13 mostra esses dados relacionados para cada classe da base de dados de emoções, demonstrando principalmente valores baixos de precisão relacionados a algumas classes em específico novamente, como as emoções de desgosto e neutro que também demonstraram resultados baixos no modelo RNN, demonstrado anteriormente, o modelo CNN-LSTM também obteve precisão baixa na classe de desgosto, com 23%, porém, algumas classes obtiveram resultados consideráveis como as emoções raiva e tristeza. Podemos considerar que o mesmo problema apontado para o modelo anterior esteja ocorrendo com este.

Figura 13 - Confusion Matrix CNN-LSTM



FONTE: Elaborado pelo autor (2025)

TABELA 4 - Classification Report BiLSTM

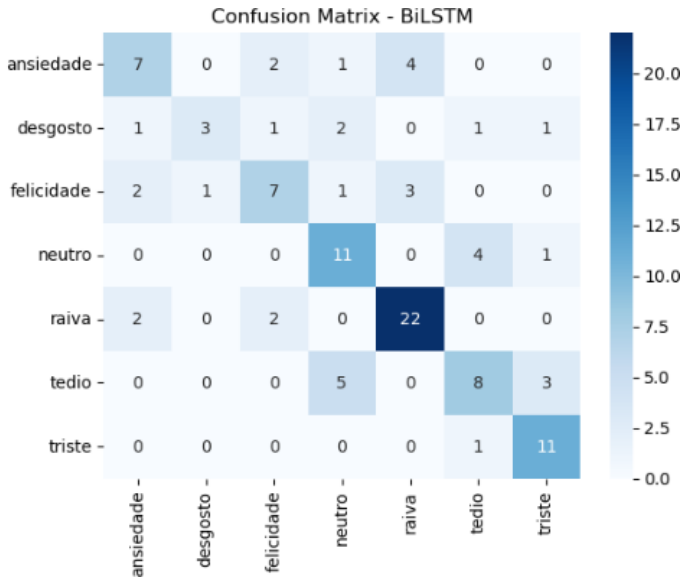
Classe	Precision	Recall	F1-Score	Suporte
ansiedade	0.58	0.5	0.54	14
desgosto	0.75	0.33	0.46	9
felicidade	0.58	0.5	0.54	14
neutro	0.55	0.69	0.61	16
raiva	0.76	0.85	0.8	26
tedio	0.57	0.5	0.53	16
triste	0.69	0.92	0.79	12
accuracy			0.64	107
macro avg	0.64	0.61	0.61	107
weighted avg	0.64	0.64	0.63	107

FONTE: Elaborado pelo autor (2025)

Com relação às métricas de precisão, recall e f1-score para o modelo BiLSTM, a Tabela 4 mostra esses dados relacionados para cada classe da base de dados de emoções, demonstrando uma clara evolução se comparar com os resultados obtidos anteriormente, e também, valores próximos aos apresentados na Tabela 1, valorizando o desempenho apresentado pelo modelo BiLSTM, com resultados de 75% na classe desgosto e 76% na classe raiva, e com melhoras em classes que demonstraram resultados baixos nos modelos anteriores. Observamos que, alguns problemas podem estar ocorrendo com a classificação do modelo visto o valor baixo de recall da classe desgosto, o que demonstra falhas na classificação,

já que dos valores classificados como correto, somente 33% foram classificadas corretamente.

Figura 14 - Confusion Matrix BiLSTM



FONTE: Elaborado pelo autor (2025)

TABELA 5 - Comparativo de métricas entre os modelos

Modelo	Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-Score	Weighted Avg F1-Score
RNN	0.21	0.18	0.2	0.18	0.19
CNN+LSTM	0.44	0.44	0.45	0.44	0.44
BiLSTM	0.64	0.64	0.61	0.61	0.63

FONTE: Elaborado pelo autor (2025)

Como mostrado nos resultados anteriores e na Tabela 5 é evidente que todos os modelos analisando apresentam resultados que variam consideravelmente, obtendo resultados num geral insatisfatórios em termos de precisão, recall e f1-score, com métricas abaixo do critério pré estabelecido anteriormente para comparação, com exceção do modelo BiLSTM, que obteve resultados consideráveis. Uma das principais causas dessa baixa performance pode ser atribuída a quantidade de dados existentes em nossa base para treinamento e validação, visto que, para a obtenção de resultados mais expressivos e com mais qualidade, é necessário uma base de dados mais extensa, com mais dados para serem utilizados como dados de validação para comparativo de métricas. Outro quesito importante a ser considerado também é o tamanho das redes utilizadas para treinamento, mesmo com a utilização de aplicações como o *Optuna* para otimização de treinamento, as redes utilizadas possuem uma arquitetura consideravelmente

simples e padrão, podendo ser estendida com o objetivo de atingir resultados ainda melhores, apesar do custo computacional maior.

Uma possível explicação para o desempenho superior do BiLSTM é a sua capacidade de capturar simultaneamente dependências temporais passadas e futuras no sinal de áudio, o que enriquece a representação dos padrões emocionais. Enquanto modelos unidirecionais observam apenas o contexto anterior, o BiLSTM integra informações de toda a sequência, permitindo decisões mais robustas mesmo diante de variações sutis na entonação e no ritmo da fala. Dessa forma, esse duplo fluxo de informação contribui para uma detecção de emoções mais precisa, especialmente em classes menos frequentes ou com características acústicas menos evidentes.

5 Conclusão

Este trabalho teve como objetivo realizar um estudo comparativo entre diferentes algoritmos e arquiteturas de modelos para detecção de emoções a partir do áudio, com o foco em redes *Deep Learning.* Foram explorados três modelos de *Speech Emotion Recognition,* com ênfase na avaliação de sua precisão e seus resultados gerados com análise das métricas escolhidas. Cada algoritmo foi meticulosamente analisado com base em exemplos de desempenho na literatura e nas métricas descritas anteriormente como precisão, *recall* e f1-score. Além disso, foi feita uma comparação direta entre os algoritmos para determinar qual é mais eficiente neste contexto específico, considerando os resultados obtidos a partir da matriz de confusão. Diante dos problemas descritos, foram identificados desafios e possíveis soluções para aprimorar a detecção e a qualidade dos modelos utilizando de técnicas existentes.

O principal problema identificado em nosso estudo, foi a eficácia limitada dos modelos de reconhecimento de emoções pela voz na classificação certeira de uma emoção que possui características próximas a outra, particularmente nas métricas de precisão e *recall*. Os modelos RNN e CNN+LSTM apresentaram um resultado baixo nessas métricas, com ênfase no modelo RNN que obteve uma precisão média de 18%, um resultado extremamente baixo, já o modelo CNN+LSTM obteve uma

precisão média de 44%, já nas métricas de média de *recall* foram 20% e 45% respectivamente, já o BiLSTM, que obteve os melhores resultados com uma precisão média de 64% e um *recall* médio de 61%, métricas com resultados consideráveis se comparados com os resultados anteriores. Estes resultados indicam uma dificuldade significativa dos modelos em detectar corretamente casos realmente positivos (*recall*) e determinar com precisão (*precision*) seus resultados.

Com base no cenário visto, e visando a possibilidade de resolver os problemas gerados pelos resultados insatisfatórios das métricas observadas, algumas estratégias podem ser consideradas. Primeiramente, um melhor refinamento da base de dados poderia ser crucial, com o objetivo primário de ampliar o conjunto de dados, e também, a execução de um pré-processamento mais eficiente dos dados com o objetivo de aumentar a diversidade, representatividade e padronização no tamanho das amostras. Além disso, a experimentação com arquiteturas mais extensivas de modelos pode ser benéfica, expandindo a quantidade de camadas dos modelos existentes, além de mais iterações tanto no otimizador do Optuna quanto das iterações por época. Considerando a variedade de abordagens disponíveis, a escolha de algoritmos alternativos ou a combinação de múltiplas técnicas pode levar a uma melhoria do desempenho.

Este trabalho destaca diversos aspectos do uso de algoritmos de detecção de emoções pela voz com ênfase em modelos *Deep Learning,* com o objetivo de oferecer informações valiosas para pesquisas futuras a partir da comparação entre modelos.

REFERÊNCIAS

AHLAM, H.; ARIF, M.; ALGHAMDI, M. **Speech Emotion Recognition approaches: A systematic review.** Speech communication, Volume 154, 2023. Disponível em: Speech emotion recognition approaches: A systematic review - ScienceDirect. Acesso em: 6 jun. 2025.

SINGH, J; SAHEER, L; FAUST, O. **Speech Emotion Recognition Using Attention Model.** Int J Environ Res Public Health, 2023.

ZHENG, D.; YUE, J. **Fourier transform is the most popular and the oldest of the signal processing tools in different science fields that is typically convenient for analysis of periodic signals.** [S.l.]: Academic Press, 2021.

MARTINS, C. Alexitimia: a dificuldade em identificar e descrever as próprias emoções. Revista Galileu, 2023

KHALIL, R.; JONES, E.; BABAR, M. **Speech Emotion Recognition Using Deep Learning Techniques: A Review.** IEEE Access, 2019.

MENG, H.; YAN, T.; YUAN, F.; WEI, H. **Speech Emotion Recognition From 3D Log-Mel Spectrograms With Deep Learning Network.** IEEE Access, 2019.

LECUN, Y.; BENGIO, Y.; HINTON, G. **Deep Learning.** Nature, maio. 2015.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. **ImageNet Classification with Deep Convolutional Neural Networks.** Communications of the ACM. maio 2012.

WHAT is Artificial Intelligence (AI)?. [*S. l.*], ago. 2024. Disponível em: https://www.ibm.com/topics/artificial-intelligence. Acesso em: 3 nov. 2024.

ZRAR KH, A; ABDULBASIT, K, A. **Mel Frequency Cepstral Coefficient and its Applications: A Review.** IEEE Access, 18 nov. 2022. 1 Figura.

KEIRON, O; RYAN, N. **An introduction to Convolutional Neural Networks**. arXiv, 2015.

KUMAR, C.S.A; MAHARANA, A.D; KRISHNAN, S.M; HANUMA, S.S.S; et al. **Speech Emotion Recognition using CNN-LSTM and Vision Transformer**. Innovations in Bio-Inspired Computing and Applications. mar, 2023. 1 figura

AKIBA, Takuya; SANO, Shotaro; YANASE, Toshihiko; OHTA, Takeru; KOYAMA, Masanori. **Optuna: a next-generation hyperparameter optimization framework.** *Preprint*, [s.l.], jul. 2019. Disponível em: *https://arxiv.org/abs/1907.10902*.

POWERS, David M. W. **Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.** *Journal of Machine Learning Technologies*, 2020

MORAIS, Felipe de; SILVA, Juarez da; REIS, Helena; ISOTANI, Seiji; JAQUES, Patricia. Computação afetiva aplicada à educação: uma revisão sistemática das pesquisas publicadas no Brasil. SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE), 2017.

NISAR, Shibli; SHAHZAD, Ibrahim; KHAN, Muhammad Adnan; TARIQ, Muhammad. **Pashto spoken digits recognition using spectral and prosodic based feature extraction.** In: INTERNATIONAL CONFERENCE ON ADVANCED COMPUTATIONAL INTELLIGENCE (ICACI), 2017. 1 Figura.

AKÇAY, M. B.; OGUN, V. U. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, jan. 2020. Disponível em: https://doi.org/10.1016/j.specom.2019.12.001. Acesso em: 6 jun. 2025.

KUMAR, M.; RAJ, B. **Speech Emotion Recognition Approaches: A Systematic Review.** ***Speech Communication***, mar. 2022. Disponível em: https://doi.org/10.1016/j.specom.2022.01.002. Acesso em: 6 jun. 2025.

GERON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 2. ed. O'Reilly, 2019.

BURKHARDT, F.; PAESCHKE, A.; RWEISER, T.; SCHRÖDER, M.; ENGELER, F.; POLZEK, T. **A database of German emotional speech**. *INTERSPEECH*, 2005. Disponível em: https://www.researchgate.net/publication/221491017_A_database_of_German_emotional_speech. Acesso em: 6 jun. 2025.

DELIJA, Milaim. **Emotion Recognition through AI: Applications, Challenges and Future Trends**. 2024. Disponível em: https://www.researchgate.net/publication/383022041_Emotion_Recognition_through_AI_Emotion_Recognition_through_AI_Applications_Challenges_and_Future_Trends. Acesso em: 18 jun. 2025.

HO, T. G. D. K. Sumanathilaka et al. **Hybrid Emotion Recognition: Enhancing Customer Interactions Through Acoustic and Textual Analysis.** arXiv, 2025. Disponível em: https://arxiv.org/abs/2503.21927. Acesso em: 18 jun. 2025.

ROUAST, Philipp V.; ADAM, Marc T. P.; CHIONG, Raymond. **Deep Learning for Human Affect Recognition: Insights and New Developments.** arXiv, 2019. Disponível em: https://arxiv.org/abs/1901.02884. Acesso em: 18 jun. 2025.

YAMASHITA, R; NISHIO, M; DO, G. K. R; TOGASHI, K. **Convolutional neural networks: an overview and application in radiology.** *Insights into imaging,* 2018.