UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS SOBRAL

CURSO DE GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO

EMANUEL DÊVID PAULINO FELIX

DETECÇÃO DE ÁUDIOS FALSOS ATRAVÉS DE WAVELETS E REDES NEURAIS CONVOLUCIONAIS

SOBRAL

2025

EMANUEL DÊVID PAULINO FELIX

DETECÇÃO DE ÁUDIOS FALSOS ATRAVÉS DE WAVELETS E REDES NEURAIS CONVOLUCIONAIS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia da Computação do Campus Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia da Computação.

Orientador: Prof. Dr. Marcelo Marques Simões de Souza

SOBRAL

2025

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

F36d Felix, Emanuel Dêvid Paulino.
Detecção de Áudios Falsos Através de Wavelets e Redes Neurais Convolucionais / Emanuel Dêvid Paulino Felix. – 2025.
64 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Sobral, Curso de Engenharia da Computação, Sobral, 2025.
Orientação: Prof. Dr. Marcelo Marques Simões de Souza.

1. Audio Deepfakes. 2. Detecção de áudio sintético. 3. Transformada Wavelet. 4. Redes Neurais Convolucionais. I. Título.

CDD 621.39

EMANUEL DÊVID PAULINO FELIX

DETECÇÃO DE ÁUDIOS FALSOS ATRAVÉS DE WAVELETS E REDES NEURAIS CONVOLUCIONAIS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia da Computação do Campus Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia da Computação.

Aprovada em: 25 de julho de 2025

BANCA EXAMINADORA

Prof. Dr. Marcelo Marques Simões de Souza (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Iális Cavalcante de Paula Júnior
Universidade Federal do Ceará (UFC)

Eng. Me. Lucas Pedrosa Valente
Chipus Microeletrônica

Ao Deus Triuno, meu Senhor e Rei. À minha família, meu conforto. À minha companheira, Sol que ilumina meus dias.

AGRADECIMENTOS

Agradeço primeiramente a Deus, Único que é Digno de toda honra, glória e louvor. Sem Ele eu não poderia ter concluído o presente trabalho nem feito coisa qualquer. Agradeço por seu sustento, graça e bênçãos ao longo de toda essa jornada.

À minha família. Meu Pai José Eumar e minha mãe Ana Cláudia, por amarem, ensinarem e cuidarem de mim. Sem eles eu não seria quem sou e não estaria onde estou. Aos meus irmãos Emile Dávila e Caio Italo, pela companhia, carinho e admiração mútua.

À minha companheira, Solane Ellen, por todo amor, carinho, atenção, apoio e motivação que foram fundamentais nessa caminhada.

Aos amigos que direta ou indiretamente contribuíram para que eu estivesse aqui e que sempre estiveram ao meu lado.

Aos irmãos da Igreja Bíblica de Morrinhos, por todo apoio e suas constantes orações, que com certeza muito me abençoaram.

Ao meu orientador, Prof. Dr. Marcelo Marques, por seus direcionamentos, conselhos, disposição e ensinamentos.

"Porque dele, e por meio dele, e para ele são todas as coisas. A ele, pois, a glória eternamente. Amém!"

(Romanos 11:36)

RESUMO

O avanço da Inteligência Artificial (IA) permitiu a criação de *deepfakes* de áudio — manipulações de voz sintética que imitam vozes de pessoas reais com grande precisão. Essa tecnologia, embora aplicada de forma útil e legítima, representa um risco significativo para a segurança da informação, sendo também utilizada em fraudes e desinformação. A sofisticação dos algoritmos de geração de áudios falsos torna a detecção um desafio, dificultando a capacidade de análise humana e de algoritmos convencionais, sendo necessário, portanto, o desenvolvimento de métodos de detecção mais eficientes. Para enfrentar este desafio, este trabalho propõe e avalia uma metodologia que combina processamento de sinais de áudio e aprendizado profundo. A abordagem transforma os sinais de áudio em escalogramas — imagens que representam a energia do sinal no tempo e na frequência — por meio da Transformada Wavelet Contínua (*Continuous Wavelet Transform* (CWT)). Para análise comparativa, o estudo utiliza escalogramas gerados a partir de três famílias de *wavelets* (Morlet, Complex Morlet e Mexican Hat) para treinar a arquitetura de Rede Neural Convolucional (*Convolutional Neural Network* (CNN)) MobileNet. Os experimentos, conduzidos sobre a base de dados *Fake or Real* (FoR), demonstram que a escolha da família de *wavelet* é um fator determinante para o desempenho. A configuração com a *wavelet* Morlet apresentou o melhor desempenho equilibrado, alcançando acurácia de 84,01% e F1-Score de 84,24%, o que demonstra a viabilidade da abordagem e estabelece esta combinação como a mais eficaz do estudo.

Palavras-chave: Audio Deepfakes; Detecção de áudio sintético; Transformada Wavelet; Redes Neurais Convolucionais.

ABSTRACT

The advancement of Artificial Intelligence (AI) has enabled the creation of audio *deepfakes*—synthetic voice manipulations that mimic real human voices with high precision. This technology, while having useful and legitimate applications, represents a significant risk to information security, being also used in fraud and disinformation. The sophistication of fake audio generation algorithms makes detection a challenge, hindering the analytical capabilities of both humans and conventional algorithms, thus necessitating the development of more effective detection methods. To address this challenge, this work proposes and evaluates a methodology that combines audio signal processing and deep learning. The approach transforms audio signals into scalograms — images representing the signal's energy in the time-frequency domain — by means of the Continuous Wavelet Transform (CWT). For a comparative analysis, the study uses scalograms generated from three *wavelet* families (Morlet, Complex Morlet, and Mexican Hat) to train the MobileNet Convolutional Neural Network (CNN) architecture. The experiments, conducted on the FoR dataset, demonstrate that the choice of *wavelet* family is a determining factor for performance. The configuration with the Morlet *wavelet* showed the best balanced performance, achieving an accuracy of 84.01% and an F1-Score of 84.24%, which demonstrates the feasibility of the approach and establishes this combination as the most effective in the study.

Keywords: Audio Deepfakes; Synthetic Audio Detection; Wavelet Transform; Convolutional Neural Networks.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AD	*Audio Deepfakes*
AMR	*Análise Multirresolução*
CNN	*Convolutional Neural Network*
CWT	*Continuous Wavelet Transform*
DL	*Deep Learning*
DWT	*Discrete Wavelet Transform*
EER	*Equal Error Rate*
FoR	*Fake or Real*
FT	*Fourier Transform*
GANs	*Generative Adversarial Networks*
GMM	*Gaussian Mixture Models*
IA	Inteligência Artificial
ML	*Machine Learning*
ReLU	*Rectified Linear Unit*
STFT	*Short-Time Fourier Transform*
TTS	*Text-to-Speech*
VAEs	*Variational Autoencoders*
VC	*Voice Conversion*
VITS	*Variational Inference Text-to-Speech*
WT	*Wavelet Transform*

SUMÁRIO

1 INTRODUÇÃO

Nos últimos anos, o termo Inteligência Artificial (IA) tornou-se bastante popular. O uso de tecnologias e algoritmos baseados em IA é cada vez mais comum na vida cotidiana. Isso se deve, ao menos, por dois motivos: a) o fácil acesso que se tem a essas tecnologia e b) o interesse crescente do público, seja por suas aplicações voltadas ao entretenimento, seja por seu potencial em auxiliar na resolução de problemas rotineiros. De fato, a evolução da IA trouxe inúmeros benefícios tanto em contextos do dia a dia quanto em ambientes industriais, acadêmicos e científicos. No entanto, esses avanços também contribuíram para o surgimento de novas vulnerabilidades, especialmente no campo da segurança da informação e da integridade das comunicações digitais.

Um exemplo bastante sensível está na área de síntese de voz, onde técnicas de IA permitem a geração de áudios que imitam fielmente a voz humana. Embora sejam úteis em aplicações de acessibilidade e assistência virtual, os algoritmos de clonagem de voz têm sido utilizados para fins maliciosos, como fraudes, clonagem de identidade vocal, manipulação de provas e desinformação. Essas últimas formas de uso provocaram a disseminação das deepfakes de áudio (do inglês, *Audio Deepfakes* (AD)). De modo geral, *deepfakes* são conteúdos ou materiais, gerados ou manipulados sinteticamente, através de IA, com o intuito de aparentar autenticidade, englobando síntese de áudio, vídeo, imagem e texto (KHANJANI *et al.*, 2023).

A sofisticação das técnicas de síntese e clonagem de voz traz um novo desafio para a segurança da informação. À medida que os métodos de geração de áudio sintético se aprimoram, impulsionados por algoritmos de IA, torna-se quase impossível distinguir as *deepfakes* de áudio de gravações autênticas, não apenas para ouvintes humanos, mas também para algoritmos de detecção menos sofisticados. Esse desafio é intensificado pela necessidade de os sistemas de detecção serem robustos e capazes de generalizar para diferentes idiomas e novas técnicas de síntese, um dos problemas em aberto na área (MAWALIM *et al.*, 2025). Portanto, é relevante o desenvolvimento de novas abordagens de detecção capazes de responder a esses avanços.

Os métodos de detecção podem ser agrupados em duas categorias: os de *Machine Learning* (ML) e os de *Deep Learning* (DL) (ALMUTAIRI; ELGIBREEN, 2022). Os métodos baseados em ML baseiam-se em modelos estatísticos para estabelecer modelos a partir de características de gravações de áudio genuínas e sintéticas. Já os métodos de DL recorrem às redes neurais profundas para identificar padrões complexos diretamente nos dados, dispensando extração de características manuais (DIXIT *et al.*, 2023). Nesse segundo grupo, as Redes Neurais

Convolucionais (do inglês, *Convolutional Neural Network* (CNN)) são as mais utilizadas por sua capacidade de processar dados multidimensionais, como imagens e representações visuais de sinais de áudio. As CNN são especialmente eficazes na detecção de padrões espaciais e temporais sutis, sendo amplamente aplicadas em tarefas de reconhecimento de fala, classificação de sons e na detecção de áudios falsificados.

A Transformada *Wavelet* (do inglês, *Wavelet Transform* (WT)) é também uma ferramenta matematicamente importante no processamento de sinais de áudio, pois permite decompor sinais em diferentes níveis de detalhe. Esse último aspecto favorece sua aplicação na extração de características globais e locais dos sinais já que os essas transformadas operam em diferentes escalas e resoluções, oferecendo uma visão tanto abrangente quanto detalhada do sinal (GRAPS, 1995).

Nesse cenário, este trabalho traz uma proposta para a detecção de áudios falsos que combina a Transformada *Wavelet* com Redes Neurais Convolucionais. A proposta consiste em gerar escalogramas (mapas tempo-frequência) dos sinais de áudio a partir da WT. Assim, é possível treinar modelos de CNNs com os escalogramas, representados como imagens, para identificar padrões que indiquem manipulação artificial.

Serão utilizadas amostras de áudio da base de dados *Fake or Real* (FoR), amplamente reconhecido na área. Na época de sua publicação, a base se destacou por conter amostras sintéticas produzidas por algoritmos de síntese de fala considerados de ponta, com naturalidade similar à fala humana real (REIMAO; TZERPOS, 2019). Três famílias de *wavelet* serão consideradas para a geração dos escalogramas: Morlet, Mexican Hat e Complex Morlet, originando três conjuntos distintos de imagens. Cada conjunto será utilizado para treinar uma arquitetura de CNN, a MobileNet, cujo desempenho será avaliado com base em métricas como acurácia, revocação, precisão e *F1-score*. Com isso, busca-se comparar os resultados entre as diferentes famílias de *wavelet* para estabelecer qual delas gera a representação mais eficaz para a detecção de áudios falsificados.

2 OBJETIVOS

2.1 Objetivo Geral

Estabelecer um modelo para detecção de áudios falsos que integre a Transformada Wavelet (WT) e Redes Neurais Convolucionais (CNN).

2.2 Objetivos Específicos

- Extrair atributos em múltiplas resoluções dos sinais de áudio utilizando escalogramas produzidos a partir de famílias de wavelet distintas;
- Construir três conjuntos de dados compostos por escalogramas, cada um gerado com uma família distinta de wavelet, a partir de amostras de áudio de vozes da base pública FoR;
- Avaliar a capacidade discriminativa da arquitetura de CNN MobileNet na tarefa de classificação de áudios reais ou falsos, a partir dos conjuntos de dados;
- Comparar o desempenho obtido com cada um dos três conjuntos, a fim de analisar a influência de cada família de *wavelet* na extração de características e identificar a representação mais eficaz para a detecção.

3 DEEPFAKES DE ÁUDIO

Audio Deepfakes (AD) são áudios gerados ou modificados por IA com o objetivo de imitar a fala humana de forma realista (KHANJANI *et al.*, 2023). Essa técnica utiliza modelos treinados para reproduzir características como timbre, entonação e ritmo de uma voz, podendo gerar falas inéditas como se fossem ditas por uma pessoa real.

O processo de clonagem de voz começa com a coleta de áudios de uma pessoa. Em seguida, esses dados alimentam modelos, geralmente redes neurais profundas, que aprendem as características da voz. Após o treinamento, é possível gerar novos áudios com alto grau de fidelidade, mesmo a partir de um simples texto.

Essas tecnologias vêm sendo aplicadas em áreas como dublagem, assistentes virtuais e acessibilidade. No entanto, também representam riscos, como fraudes, golpes e manipulação de informações, já que permitem simular falas de qualquer pessoa com aparência legítima.

Casos recentes evidenciam os riscos associados as AD. Em 2025, celebridades brasileiras como Neymar e Paolla Oliveira foram alvos de vídeos manipulados que se espalharam nas redes sociais, gerando preocupação com a disseminação de desinformação digital (O GLOBO, 2025). No cenário político, durante as eleições presidenciais de 2022, circulou um vídeo falso em que a apresentadora Renata Vasconcellos, do Jornal Nacional, anunciava uma pesquisa eleitoral adulterada, invertendo os resultados reais entre os candidatos Jair Bolsonaro e Luiz Inácio Lula da Silva (DAUER, 2024).

Neste capítulo, serão abordados os principais tipos de *deepfakes* de áudio, as tecnologias utilizadas em sua geração e os métodos atuais e mais eficientes para sua detecção. A Seção 3.1 trata das diferentes formas de manipulação e a Seção 3.2 das principais ferramentas de geração. Por fim, a Seção 3.3 discute os métodos de detecção e compara o desempenho entre eles.

3.1 Tipos

Os principais tipos de *deepfakes* de áudio podem ser resumidos em *Text-to-Speech* (TTS), *Voice Conversion* (VC), Emotion Fake, Scene Fake e Partially Fake (YI *et al.*, 2023).

3.1.1 Text-to-Speech

A síntese de texto em fala, do inglês *Text-to-Speech* (TTS), é uma tecnologia para conversão de texto em áudio com características naturais da fala humana. Essa conversão envolve três etapas: análise textual, modelagem acústica e vocodificação. A primeira etapa é responsável por transformar o texto em representações fonéticas e prosódicas. Na modelagem acústica são previstos os parâmetros da voz do locutor que se pretende simular e, por fim, a vocodificação é a etapa que produz a forma de onda de saída. As soluções modernas utilizam redes neurais profundas para realizar esse mapeamento de forma direta, alcançando níveis elevados de fluência e realismo na fala sintetizada (KHANJANI *et al.*, 2023).

Dentre os modelos propostos, destacam-se o Tacotron 2, FastSpeech, Glow-TTS e *Variational Inference Text-to-Speech* (VITS). Tais modelos representam diferentes abordagens para geração de espectrogramas Mel, e na conversão desses espectrogramas em áudio, vocoders como WaveNet, MelGAN e Parallel WaveGAN são amplamente utilizados. Isso porque tais vocoders são capazes de reconstruir formas de onda com alta qualidade (SHAABAN *et al.*, 2023; YI *et al.*, 2023; DIXIT *et al.*, 2023).

Com a evolução desses modelos, o TTS passou a ser aplicado em diversas áreas, como leitores automáticos, assistentes virtuais, sistemas de navegação e ferramentas de acessibilidade. No entanto, a mesma tecnologia também possibilita a criação de AD convincentes, especialmente quando treinada com amostras de voz de pessoas reais.

3.1.2 Voice Conversion

Voice Conversion (VC) é uma técnica que modifica a voz de um locutor para que soe como a de outra pessoa, sem alterar o conteúdo linguístico da fala. O objetivo é preservar a mensagem transmitida, mas alterando as características vocais como timbre, entonação e altura da voz, mantendo a naturalidade e a inteligibilidade do áudio.

Os primeiros VCs foram construídos através de métodos estatísticos, como o *dynamic time warping* e modelos baseados em *Gaussian Mixture Models* (GMM), que modelam estatisticamente a correspondência entre as características espectrais das vozes de origem e de destino. O advento da aprendizagem profunda trouxe então arquiteturas baseadas em redes neurais, que dominam atualmente esse campo. Modelos *end-to-end* baseados em *variational autoencoders*, *Generative Adversarial Networks* (GANs) e *Transformers*, permitem realizar a conversão de voz

com alta fidelidade, mesmo a partir de poucas amostras da voz-alvo (SHAABAN *et al.*, 2023).

A VC consolidou-se como a abordagem mais comum para a criação de AD pela flexibilidade que propicia para replicar a identidade vocal de qualquer pessoa em falas inéditas. Essa técnica tem sido aplicada tanto em contextos legítimos, como tradução simultânea e personalização de assistentes virtuais, quanto em cenários maliciosos, como fraudes e falsificação de identidade.

3.1.3 Emotion Fake

As falsificações emocionais (do inglês, *emotion fakes*), consistem na modificação do estado emocional expresso em uma fala, sem alterar seu conteúdo semântico nem a identidade do locutor. O objetivo é produzir a fala parecendo transmitir emoções diferentes daquelas originalmente expressas, como transformar uma fala neutra em uma fala triste, raivosa ou entusiasmada.

Esse processo envolve a manipulação de características prosódicas, como o ritmo, a entonação, a intensidade e a altura da voz, o que pode ser feito por modelos treinados para alterar essas propriedades, seja a partir de representações acústicas intermediárias, como espectrogramas, ou diretamente na forma de onda. Técnicas baseadas em redes neurais, incluindo *variational autoencoders* e GANs, são frequentemente utilizadas para esse propósito.

Em contextos legítimos, a modulação emocional pode ser aplicada em dublagens expressivas, síntese de fala empática e personalização de assistentes virtuais. No entanto, quando usada maliciosamente, essa técnica visa alterar a intenção original da mensagem, distorcendo depoimentos, declarações públicas ou falas jornalísticas, com potencial para desinformação e manipulação de opinião.

3.1.4 Scene Fake

As falsificações de cenário (do inglês, *scene fakes*) referem-se à manipulação do ambiente sonoro em que a voz está inserida, com o objetivo de simular que a fala ocorreu em um contexto acústico diferente do original. Por exemplo, é possível alterar uma gravação feita em um ambiente silencioso para que pareça ter sido realizada em um local público, como um aeroporto ou restaurante.

Essa modificação pode ser feita pela inserção ou substituição de elementos do fundo sonoro, utilizando técnicas de mistura e filtragem de áudio, ou pela aplicação de convoluções

com perfis acústicos típicos de determinados ambientes. Também é possível empregar modelos treinados para simular reverberações e ruídos específicos de um local-alvo.

Embora existam aplicações legítimas, como em design sonoro, produção audiovisual e reconstituições simuladas, esse tipo de falsificação levanta sérias preocupações quando utilizado para alterar a percepção do contexto de uma fala. Em ambientes jurídicos, jornalísticos ou forenses, a adulteração do ambiente pode comprometer a autenticidade do registro e influenciar a interpretação dos fatos.

3.1.5 Partially Fake

As falsificações parciais (do inglês, *partially fakes*), ocorrem quando apenas segmentos específicos de uma gravação original são manipulados ou substituídos por trechos sintéticos. Ao contrário dos casos em que todo o áudio é gerado artificialmente, aqui a maior parte do conteúdo é genuína, tornando a alteração mais difícil de ser detectada.

Esse tipo de manipulação é normalmente realizado por meio da substituição de palavras ou frases isoladas, utilizando sistemas de *text-to-speech* treinados para replicar a voz original com alta precisão. Em alguns casos, a edição pode ocorrer diretamente em representações acústicas, com técnicas de corte, inserção e recomposição do sinal, preservando as transições naturais entre os trechos reais e falsificados.

O uso de *partially fakes* representa um risco elevado, especialmente em contextos onde a credibilidade do áudio é essencial, como investigações, reportagens ou julgamentos. A substituição de uma única palavra pode alterar totalmente o significado de uma declaração, sem levantar suspeitas evidentes nem deixar vestígios facilmente detectáveis por sistemas automáticos.

3.2 Técnicas de Geração

Os avanços na geração de AD têm sido impulsionados por diferentes arquiteturas de redes neurais profundas. As abordagens podem ser divididas em categorias principais, como os modelos baseados na geração de espectrogramas, os *vocoders* neurais responsáveis pela síntese da forma de onda, as redes adversariais para conversão de voz e os modelos *end-to-end* que unificam o processo. A seguir, são apresentadas as principais tecnologias empregadas.

3.2.1 Modelos baseados em espectrogramas

Grande parte dos sistemas modernos de síntese de fala segue uma abordagem em duas etapas: primeiro, um modelo converte o texto de entrada em uma representação acústica intermediária, como o espectrograma mel; em seguida, essa representação é usada por um *vocoder* para sintetizar a forma de onda final. Modelos como o Tacotron 2 popularizaram o uso de arquiteturas *sequence-to-sequence* com mecanismos de atenção para realizar a primeira etapa. Para aumentar a velocidade e a estabilidade da inferência, abordagens mais recentes como FastSpeech e Glow-TTS otimizaram este processo, empregando arquiteturas baseadas em *Transformer* e fluxos normalizadores, respectivamente (SHAABAN *et al.*, 2023; YI *et al.*, 2023).

3.2.2 Vocoders Neurais

Após a geração do espectrograma, um *vocoder* neural é responsável por converter essa representação em uma forma de onda audível. O WaveNet foi um dos primeiros modelos autoregressivo a alcançar uma qualidade de áudio próxima à da fala humana, embora com um alto custo computacional. Para solucionar a lentidão, modelos mais rápidos como o MelGAN e o Parallel WaveGAN foram propostos posteriormente, utilizando arquiteturas não autorregressivas e adversariais para acelerar a síntese sem uma perda significativa de qualidade sonora (DIXIT *et al.*, 2023).

3.2.3 Modelos baseados em GANs

Para a tarefa de VC, que transforma a voz de um locutor para que soe como a de outro, as GANs têm sido amplamente exploradas. Diferentemente do TTS, o foco é alterar características vocais como timbre e entonação, preservando o conteúdo linguístico. Modelos como o CycleGAN-VC e o StarGAN-VC se destacam por permitirem essa conversão mesmo sem dados paralelos — sem a necessidade de o mesmo texto ser falado por ambos os locutores — o que torna as GANs particularmente eficazes para cenários de clonagem de voz (KHANJANI *et al.*, 2023).

3.2.4 Autoencoders Variacionais

Os *Variational Autoencoders* (VAEs) são outra classe de modelos generativos utilizados na síntese de fala, especialmente por sua capacidade de aprender representações latentes

controláveis da voz. Essa característica permite manipular atributos específicos do áudio, como a identidade de um locutor ou sua emoção, com maior flexibilidade. Por essa razão, os VAEs são frequentemente aplicados em tarefas de personalização da fala e geração de emoções artificiais, tanto em sistemas TTS quanto VC (SHAABAN *et al.*, 2023).

3.2.5 Modelos end-to-end

Modelos *end-to-end* integram todas as etapas do processo em uma única arquitetura unificada. O VITS é um exemplo desse tipo de modelo, combinando aprendizado variacional, adversarial e autoregressivo para realizar a síntese de voz de forma direta e com qualidade comparável à fala humana. Esse tipo de arquitetura tende a ser mais eficiente e menos dependente de ajustes manuais em cada estágio (YI *et al.*, 2023).

Embora tenham sido inicialmente desenvolvidos para fins legítimos como acessibilidade, personalização e tradução automatizada, os modelos *end-to-end* têm sido cada vez mais explorados na criação de deepfakes realistas, o que evidencia a importância de compreender seu funcionamento e seus riscos potenciais.

3.2.6 Síntese das Tecnologias de Geração

As diversas tecnologias para a geração de áudio sintético discutidas nesta seção são resumidas na Tabela 1.

Tabela 1 – Principais tecnologias utilizadas na geração de *deepfakes* de áudio.

Tecnologia	**Descrição**
Tacotron 2	Modelo *sequence-to-sequence* que converte texto em espectrogramas mel.
FastSpeech / Glow-TTS	Arquiteturas não autorregressivas baseadas em *Transformers* e *normalizing flows* para geração paralela de espectrogramas.
WaveNet	*Vocoder* autorregressivo que gera sinais de áudio de alta qualidade a partir de espectrogramas.
MelGAN / Parallel WaveGAN	*Vocoders* baseados em GANs que sintetizam áudio de forma não autorregressiva, com maior velocidade.
CycleGAN-VC / StarGAN-VC	Modelos de conversão de voz (VC) que transferem características vocais entre locutores sem a necessidade de dados paralelos.
Autoencoders Variacionais (VAEs)	Arquiteturas que aprendem representações latentes para manipulação controlada de atributos da voz, como identidade ou emoção.
VITS (Variational Inference Text-to-Speech)	Modelo *end-to-end* que combina VAEs e GANs para sintetizar fala de alta qualidade diretamente do texto.

Fonte: Adaptado de (SHAABAN *et al.*, 2023; YI *et al.*, 2023; DIXIT *et al.*, 2023).

3.3 Técnicas de Detecção

A detecção de AD evolui a partir de duas abordagens complementares: a tradicional, baseada na extração manual de atributos acústicos combinada com classificadores clássicos; e a moderna, que adota modelos *end-to-end* com aprendizado profundo capazes de identificar padrões diretamente, a partir da forma de onda ou de representações espectrais. As seções a seguir detalham aspectos de cada uma dessas abordagens.

3.3.1 Abordagens Clássicas

Os métodos clássicos de detecção baseiam-se na extração de atributos acústicos projetados manualmente, conhecidos como *handcrafted features*, que buscam representar informações relevantes da estrutura espectral e prosódica da fala. Coeficientes como MFCC (*Mel-Frequency Cepstral Coefficients*) e CQCC (*Constant-Q Cepstral Coefficients*) são amplamente utilizados para esta tarefa, sendo subsequentemente processados por classificadores estatísticos, como o GMM ou sua variante GMM-UBM (*Universal Background Model*). Essa abordagem se destacou nas competições ASVspoof de 2015 e 2017, apresentando resultados competitivos em ambientes controlados (KHANJANI *et al.*, 2023).

Além dos coeficientes cepstrais, outras representações como espectrogramas log-mel ou gerados por FFT também são empregadas como entrada para algoritmos como as Máquinas de Vetores de Suporte (SVM), permitindo a classificação em tarefas de detecção de imitação vocal com altas taxas de acerto. Complementarmente, características prosódicas, que descrevem o ritmo e a melodia da fala — como o *pitch* (frequência fundamental), a duração das sílabas e a intensidade do sinal —, também são aplicadas em conjunto com classificadores como *Random Forest* (RF) e Regressão Logística (LR), atingindo um desempenho de detecção considerável em diversos cenários (DIXIT *et al.*, 2023).

3.3.2 Redes Neurais Convolucionais

A estratégia de detecção baseada em Redes Neurais Convolucionais (CNNs) representa uma das abordagens mais proeminentes no combate aos AD (PHAM *et al.*, 2024). O princípio fundamental dessa abordagem consiste em transformar o problema de análise de um sinal de áudio unidimensional em uma tarefa de classificação de imagens bidimensionais (TSALERA *et al.*, 2021). Para isso, o sinal de áudio é primeiramente convertido em uma representação no domínio tempo-frequência, como um espectrograma, um mel-espectrograma ou um escalograma.

Essas representações visuais são então utilizadas como entrada para uma CNN, que é treinada para identificar padrões e texturas que diferenciam a fala humana autêntica da fala gerada sinteticamente. A CNN é capaz de aprender automaticamente uma hierarquia de características (KISKIN *et al.*, 2020). As camadas iniciais da rede podem detectar artefatos de baixo nível, como inconsistências espectrais ou bordas de frequência não naturais, enquanto as camadas mais profundas aprendem a reconhecer combinações complexas desses padrões que formam uma "assinatura"da síntese artificial. Dessa forma, a CNN automatiza a extração de características discriminativas, superando muitas vezes a eficácia de métodos que dependem de atributos acústicos pré-selecionados manualmente (*handcrafted features*) (TSALERA *et al.*, 2021; KISKIN *et al.*, 2020).

3.3.3 Métodos End-to-End e Arquiteturas Avançadas

Além das abordagens que utilizam representações visuais, uma outra frente de pesquisa foca em modelos *end-to-end*, que aprendem características diretamente da forma de onda do áudio. A principal vantagem dessa estratégia é evitar a perda de informação que pode

ocorrer na conversão do sinal para um espectrograma, permitindo que a rede neural otimize a extração de atributos desde o primeiro estágio. Um exemplo proeminente é a arquitetura AASIST (sigla para *Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks*), que emprega Redes de Atenção de Grafo (do inglês, *Graph Attention Networks* (GAT)) para modelar as complexas relações espectro-temporais diretamente do áudio bruto, mostrando-se uma técnica de ponta em desafios da área (JUNG *et al.*, 2021; ALMUTAIRI; ELGIBREEN, 2022).

Outra estratégia avançada é o uso de aprendizado autossupervisionado (do inglês, *self-supervised learning* (SSL)). Nesse paradigma, os modelos são pré-treinados em vastas quantidades de dados de áudio não rotulados para aprender representações ricas e generalizáveis da fala. Modelos como o SSAD (*Self-Supervised Audio Deepfake Detection*) utilizam essa abordagem com o objetivo de melhorar a capacidade de generalização do detector, especialmente contra ataques de *deepfake* que não foram vistos durante a fase de treinamento supervisionado (YI *et al.*, 2023).

Finalmente, uma terceira linha de pesquisa explora a otimização automática de arquiteturas (do inglês, *Neural Architecture Search* (NAS)). Em vez de projetar manualmente a rede neural, essa abordagem utiliza algoritmos para encontrar a estrutura mais eficaz para a tarefa. Métodos como o PC-DARTS, baseados em busca de arquitetura diferenciável, demonstram que é possível gerar automaticamente modelos que são competitivos com aqueles desenhados por especialistas, oferecendo um bom equilíbrio entre desempenho e eficiência computacional (SHAABAN *et al.*, 2023).

3.3.4 Desempenho Comparativo das Abordagens

A fim de avaliar a eficácia dos métodos de detecção de AD, a comunidade científica se apoia em desafios e bases de dados padronizadas, como o ASVspoof e o ADD. A análise dos resultados publicados nesses desafios revela uma clara progressão e algumas tendências importantes.

A métrica *Equal Error Rate* (EER) é amplamente adotada nessas competições por refletir o ponto de equilíbrio entre os erros de falso positivo e falso negativo. Observa-se que abordagens clássicas, que utilizam características como CQCC com classificadores GMM, estabeleceram as primeiras linhas de base, mas foram consistentemente superadas por métodos de aprendizado profundo (KHANJANI *et al.*, 2023; DIXIT *et al.*, 2023).

Dentro do aprendizado profundo, uma tendência notável é a superioridade de arquiteturas *end-to-end*, que operam diretamente na forma de onda. Modelos como RawNet2 e, especialmente, o AASIST, que emprega Redes de Atenção de Grafo, demonstram um desempenho estado da arte, alcançando taxas de erro muito baixas em cenários competitivos (JUNG *et al.*, 2021; ALMUTAIRI; ELGIBREEN, 2022). Isso evidencia que o aprendizado direto de características a partir do sinal bruto é uma estratégia altamente promissora, embora outras abordagens, como a busca automática de arquiteturas (NAS), também apresentem resultados relevantes (SHAABAN *et al.*, 2023). Esse panorama reforça a importância de se investir em modelos mais sofisticados e adaptáveis para lidar com a crescente qualidade dos AD.

4 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica que serve de base para o desenvolvimento deste trabalho. A Seção 4.1 aborda os conceitos da Transformada Wavelet, detalhando a técnica de extração de características utilizada. Em seguida, a Seção 4.2 explora os fundamentos das Redes Neurais Convolucionais, a arquitetura empregada para a classificação dos dados. Por fim, a Seção 4.3 discute os trabalhos relacionados, contextualizando a presente pesquisa no estado da arte da detecção de *deepfakes* de áudio.

4.1 Transformada Wavelet

A Transformada Wavelet (do inglês, *Wavelet Transform* (WT)) é uma ferramenta matemática estabelecida para a análise de sinais em múltiplas resoluções, demonstrando particular eficácia no processamento de sinais não estacionários, categoria à qual pertencem os sinais de áudio (GRAPS, 1995; RIOUL; VETTERLI, 1991). Diferentemente da tradicional Transformada de Fourier (do inglês, *Fourier Transform* (FT)), que decompõe um sinal em uma soma de funções senoidais de duração infinita e, portanto, perde a informação temporal localizada, a WT utiliza funções base de curta duração, denominadas *wavelets* mães. Essas *wavelets* mães são versões escaladas (dilatadas ou comprimidas) e transladadas de uma função protótipo, permitindo uma análise que localiza simultaneamente a energia do sinal no domínio do tempo e da frequência (GRAPS, 1995). A escolha da *wavelet mãe* é um aspecto crucial, pois diferentes famílias de *wavelets* (como Haar, Daubechies, Morlet, Mexican Hat, entre outras) possuem distintas propriedades de suavidade, simetria e número de momentos nulos, o que influencia sua capacidade de representar diferentes tipos de transientes e componentes no sinal (ZHANG, 2019).

A ideia central da análise por *wavelets* consiste em examinar o sinal em diferentes escalas ou níveis de resolução. Ao dilatar a *wavelet mãe*, analisa-se componentes de baixa frequência (eventos lentos) do sinal com boa resolução em frequência, mas menor resolução temporal. Ao comprimi-la, analisam-se componentes de alta frequência (eventos rápidos ou transientes) com alta resolução temporal, mas menor resolução em frequência (GRAPS, 1995). Essa capacidade de *zoom* adaptável no plano tempo-frequência é o que torna a WT especialmente poderosa para capturar características locais e transitórias de sinais complexos como a fala, incluindo variações sutis de entonação, ritmo e a presença de artefatos que podem ser indicativos

de síntese artificial.

4.1.1 Análise Multirresolução e Decomposição

O conceito de *Análise Multirresolução* (AMR), formalizado por MALLAT (MALLAT, 1989), fornece o embasamento teórico para a WT. A AMR descreve um sinal como uma série de aproximações sucessivas em diferentes escalas. A cada nível de resolução, os detalhes que diferenciam uma aproximação da aproximação na escala imediatamente superior são capturados por coeficientes *wavelet*. Esse processo de decomposição pode ser visualizado como uma passagem do sinal através de um banco de filtros, separando-o em diferentes sub-bandas de frequência (RIOUL; VETTERLI, 1991).

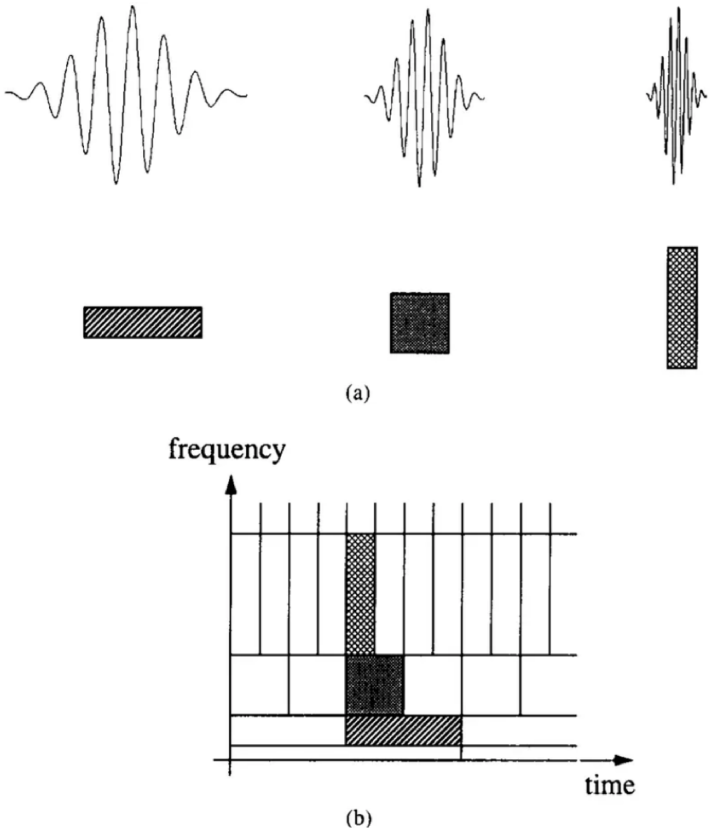
Para um sinal $f(t)$, sua representação por meio da WT pode ser expressa como uma combinação linear de funções *wavelet* $\psi_{j,k}(t)$, que são versões escaladas (pelo fator j) e transladadas (pelo fator k) da *wavelet mãe* $\psi(t)$:

$$f(t) \approx \sum_{j,k} c_{j,k}\, \psi_{j,k}(t) \tag{4.1}$$

onde $c_{j,k}$ são os coeficientes *wavelet*, representando a contribuição da componente $\psi_{j,k}(t)$ para o sinal original. Esses coeficientes quantificam a similaridade entre o sinal e a *wavelet* em uma determinada escala e posição temporal.

A aplicação da Transformada Wavelet Discreta (DWT) em um sinal bidimensional, como uma imagem, ilustra de forma clara o processo de Análise Multirresolução. A Figura 1 demonstra o resultado de uma decomposição de um nível, onde a imagem é separada em componentes de diferentes frequências.

Figura 1 – Decomposição de uma imagem em um nível utilizando a DWT.



Fonte: (RANGEL, 2020).

A figura ilustra o processo de decomposição em multirresoluções, onde cada região da imagem resultante é "composta"por wavelets filhas — versões escaladas e transladadas da wavelet mãe. A área de aproximação (canto superior esquerdo) contém as componentes de baixa frequência e pode ser vista como uma versão em menor resolução da imagem original. As demais áreas contêm os coeficientes de detalhe, que capturam as componentes de alta frequência, como as bordas e texturas em diferentes orientações (RANGEL, 2020). Este processo de filtragem pode ser aplicado de forma recursiva sobre a componente de aproximação para obter múltiplos níveis de decomposição.

4.1.2 Continuous Wavelet Transform (CWT), Discrete Wavelet Transform (DWT) e a Geração de Escalogramas

A WT pode ser implementada principalmente de duas formas:

- A Transformada Wavelet Contínua (CWT) calcula os coeficientes *wavelet* para um *continuum* de escalas e posições temporais. Embora resulte em uma representação altamente redundante, essa redundância é vantajosa para a análise detalhada e a visualização da

distribuição de energia do sinal no plano tempo-frequência. A representação gráfica dos módulos ao quadrado dos coeficientes da CWT é conhecida como *escalograma*. Os escalogramas são mapas tempo-frequência que podem ser tratados como imagens bidimensionais, tornando-os adequados como entrada para Redes Neurais Convolucionais (CNNs) em tarefas de classificação de sinais (KISKIN *et al.*, 2020; CANTüRK; GüNAY, 2024).

- A Transformada Wavelet Discreta (DWT) opera sobre um subconjunto discreto de escalas e posições, tipicamente em uma grade diádica. Isso resulta em uma representação não redundante (ou minimamente redundante) e computacionalmente mais eficiente, sendo amplamente utilizada em aplicações como compressão de dados, remoção de ruído e codificação de sinais (RIOUL; VETTERLI, 1991). Para imagens, a DWT bidimensional é frequentemente empregada para extrair características texturais em diferentes orientações e resoluções (ZHANG, 2019).

Para os propósitos deste trabalho, a CWT é a ferramenta de escolha devido à sua capacidade de gerar escalogramas ricos em informação visual, que podem revelar padrões discriminativos entre a fala autêntica e a sintética quando analisados por CNNs. A CWT inerentemente oferece uma resolução temporal fina para componentes de alta frequência e uma resolução em frequência fina para componentes de baixa frequência, adaptando-se bem à estrutura dos sinais de fala, onde eventos transientes (como consoantes oclusivas) coexistem com componentes mais estacionários (como vogais).

4.1.3 Wavelet Versus Abordagens Tradicionais

A principal limitação da FT é sua incapacidade de localizar temporalmente os componentes de frequência do sinal; ela informa quais frequências estão presentes, mas não quando elas ocorrem. A Transformada de Fourier de Curto Termo (do inglês, *Short-Time Fourier Transform* (STFT)), que aplica a FT a segmentos janelados do sinal, tenta contornar essa limitação. No entanto, a STFT utiliza uma janela de análise de tamanho fixo, resultando em uma resolução tempo-frequência uniforme em todo o espectro. Isso implica um compromisso: uma janela estreita melhora a resolução temporal, mas piora a resolução em frequência, e vice-versa, conforme o princípio da incerteza de Gabor (GRAPS, 1995).

A WT, com sua abordagem multirresolução, supera essa limitação ao adaptar dinamicamente a largura da janela de análise: janelas curtas para altas frequências (boa resolução temporal) e janelas longas para baixas frequências (boa resolução espectral). Essa característica é

particularmente vantajosa para sinais de áudio, que frequentemente contêm componentes de alta frequência de curta duração e componentes de baixa frequência de longa duração (RIOUL; VETTERLI, 1991). Consequentemente, a WT é mais adequada para detectar anomalias, transientes ou padrões sutis que podem ser indicativos de manipulação em áudios sintéticos.

4.1.4 Aplicações e Relevância para Detecção de Audio Deepfakes

A teoria das *wavelets* unificou e estendeu diversas técnicas clássicas de processamento de sinais, como codificação em sub-bandas e análise de transientes (RIOUL; VETTERLI, 1991). Suas aplicações são vastas, incluindo compressão de sinais e imagens, denoização, análise de texturas, detecção de bordas e descontinuidades, e extração de características para reconhecimento de padrões em diversas áreas, desde geofísica até diagnósticos médicos (GRAPS, 1995).

No domínio da detecção de AD, o emprego da WT, especificamente a CWT, para gerar escalogramas tem se mostrado uma estratégia promissora. Os escalogramas transformam o sinal de áudio unidimensional em uma representação bidimensional rica em informação, que pode capturar artefatos sutis ou padrões não lineares introduzidos pelos algoritmos de síntese de voz. Esses padrões visuais nos escalogramas podem ser mais facilmente aprendidos por arquiteturas de CNN, que são proficientes na identificação de características espaciais e texturais em imagens (KISKIN *et al.*, 2020). Pesquisas recentes têm explorado essa combinação para diversas tarefas de classificação de sinais acústicos, incluindo a detecção de doenças a partir da voz (CANTüRK; GüNAY, 2024) e a detecção de sons em bioacústica (KISKIN *et al.*, 2020), demonstrando a versatilidade e eficácia da abordagem.

4.1.5 Considerações Finais

Em suma, a WT oferece um arcabouço matemático robusto e flexível para a análise de sinais em múltiplas escalas de tempo e frequência. Sua capacidade intrínseca de se adaptar às características locais de um sinal, aliada à possibilidade de gerar representações visuais informativas como os escalogramas, torna essa técnica particularmente valiosa para a tarefa de detecção de áudios falsos. Espera-se que os padrões finos e, por vezes, não lineares, que diferenciam a fala humana autêntica da fala sinteticamente gerada, sejam efetivamente destacados nos escalogramas e subsequentemente identificados pelas CNNs.

4.2 Redes Neurais Convolucionais

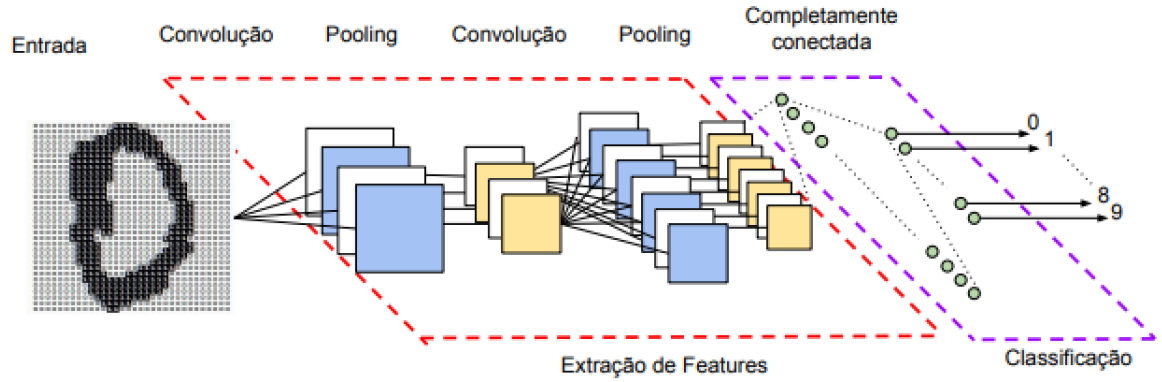
As Redes Neurais Convolucionais (*Convolutional Neural Network* (CNN)) constituem uma classe especializada de redes neurais profundas que se tornaram o padrão de excelência em diversas tarefas de visão computacional e processamento de sinais que podem ser representados visualmente (BHATT *et al.*, 2021; LI *et al.*, 2022). Sua arquitetura é bioinspirada, emulando aspectos do córtex visual humano, e é projetada para aprender automaticamente representações hierárquicas de características a partir de dados brutos, como *pixels* de uma imagem (VARGAS *et al.*, 2016). Essa capacidade de aprendizado de características dispensa, em grande medida, a necessidade de extração manual de atributos, permitindo que a rede identifique padrões relevantes — desde bordas e texturas simples nas camadas iniciais até formas e objetos complexos nas camadas mais profundas — diretamente dos dados de entrada (DESAI; SHAH, 2021). Tal propriedade as torna particularmente adequadas para tarefas como classificação, detecção de objetos e segmentação de imagens.

A eficácia das CNNs reside em três ideias principais: campos receptivos locais, compartilhamento de pesos (ou parâmetros) e subamostragem espacial (*pooling*) (LI *et al.*, 2022). Essas características permitem que as CNNs explorem a forte correlação espacial presente em dados como imagens e, por extensão, em representações tempo-frequência como os escalogramas. A evolução contínua das CNNs, com o desenvolvimento de novas camadas, funções de ativação e arquiteturas inovadoras, tem expandido seu domínio de aplicação para áreas como diagnósticos médicos, reconhecimento facial e veículos autônomos, demonstrando sua robustez e adaptabilidade a cenários complexos do mundo real (BHATT *et al.*, 2021; LI *et al.*, 2022).

4.2.1 Camadas da Arquitetura CNN

Uma CNN típica é construída pela composição sequencial de diferentes tipos de camadas, cada uma desempenhando um papel específico na extração e transformação de características. As camadas mais proeminentes são as camadas de convolução, de *pooling* (subamostragem) e totalmente conectadas, conforme ilustrado na Figura 2.

Figura 2 – Topologia básica de uma CNN



Fonte: (VARGAS *et al.*, 2016).

4.2.1.1 Convolução

A camada de convolução é o bloco construtivo central de uma CNN. Nesta camada, um conjunto de filtros (ou *kernels*), que são pequenas matrizes de pesos, desliza sobre a imagem de entrada, realizando uma operação de convolução em cada posição. Essa operação calcula o produto escalar entre os pesos do filtro e a região local da entrada (campo receptivo), produzindo um mapa de ativação ou mapa de características (*feature map*) (BHATT *et al.*, 2021). Cada filtro aprende a detectar um padrão específico, como uma borda, uma cor ou uma textura. Uma característica fundamental é o compartilhamento de pesos: o mesmo filtro é aplicado em toda a extensão da imagem, o que reduz drasticamente o número de parâmetros a serem aprendidos (comparado a uma rede totalmente conectada) e confere à rede uma propriedade de equivariância a translações — um padrão aprendido em uma parte da imagem pode ser detectado em qualquer outra parte (LI *et al.*, 2022). A extração hierárquica de características ocorre à medida que múltiplas camadas convolucionais são empilhadas: as primeiras camadas aprendem características de baixo nível, enquanto camadas subsequentes combinam essas características para detectar padrões mais complexos e abstratos (DESAI; SHAH, 2021).

A operação de convolução bidimensional discreta é definida como:

$$S(i,j) = (X * K)(i,j) = \sum_m \sum_n X(i+m, j+n) \cdot K(m,n) \tag{4.2}$$

onde X é a matriz de entrada (por exemplo, um canal de um escalograma), K é o filtro convolucional, e $S(i,j)$ é o elemento na posição (i,j) do mapa de características resultante.

Variantes como convoluções deformáveis, que adaptam dinamicamente o formato do campo receptivo, e convoluções grupais ou separáveis em profundidade (*depthwise separable*

convolutions), que otimizam o processo convolucional, são exemplos de inovações que buscam aumentar a eficiência e a capacidade de representação dessas camadas (LI *et al.*, 2022; BHATT *et al.*, 2021).

4.2.1.2 Ativação

Imediatamente após cada operação de convolução (e também nas camadas totalmente conectadas), aplica-se uma função de ativação não linear aos mapas de características. Essa não-linearidade é crucial, pois permite que a rede aprenda mapeamentos mais complexos do que seria possível com transformações puramente lineares. A Unidade Linear Retificada (*Rectified Linear Unit* (ReLU)) e suas variantes (como Leaky ReLU, Parametric ReLU) são amplamente utilizadas devido à sua simplicidade computacional e por mitigarem o problema do desaparecimento do gradiente (*vanishing gradient problem*) durante o treinamento de redes profundas (BHATT *et al.*, 2021; LI *et al.*, 2022). A ReLU, por exemplo, é definida como $f(x) = \max(0, x)$.

4.2.1.3 Pooling (Subamostragem)

As camadas de *pooling* têm como objetivo principal reduzir a dimensionalidade espacial (largura e altura) dos mapas de características, diminuindo assim o número de parâmetros, o custo computacional e controlando o sobreajuste (*overfitting*) (VARGAS *et al.*, 2016). Ao reduzir a resolução espacial, a camada de *pooling* também contribui para que as representações de características se tornem mais robustas a pequenas variações (translações ou distorções) na posição dos padrões na entrada (LI *et al.*, 2022). As operações de *pooling* mais comuns são o *max pooling*, que seleciona o valor máximo de uma vizinhança no mapa de características, e o *average pooling*, que calcula a média dos valores na vizinhança.

4.2.1.4 Totalmente Conectada (Fully Connected Layer)

Após uma sequência de camadas convolucionais, de ativação e de *pooling*, os mapas de características resultantes, que já codificam informações de alto nível sobre a entrada, são geralmente achatados (*flattened*) em um vetor unidimensional. Esse vetor é então alimentado a uma ou mais camadas totalmente conectadas. Nessas camadas, cada neurônio está conectado a todas as ativações da camada anterior, de forma similar às redes neurais tradicionais (DESAI; SHAH, 2021). A camada totalmente conectada final usa uma função de ativação apropriada

para a tarefa em questão (por exemplo, *softmax* para classificação multiclasse ou sigmoide para classificação binária) para produzir a saída da rede.

4.2.2 Evoluções Arquiteturais Notáveis em CNNs

O campo das CNNs tem testemunhado uma rápida evolução arquitetural. Marcos importantes incluem o desenvolvimento de redes como AlexNet, VGG, e, mais notavelmente, *GoogLeNet* (com seus módulos *Inception*) e *ResNet* (Redes Residuais). Os módulos *Inception* introduziram a ideia de blocos com múltiplos caminhos (*multi-path*) contendo filtros de diferentes tamanhos em paralelo, permitindo a captura de características em diversas escalas simultaneamente (BHATT *et al.*, 2021). As *ResNets* abordaram o problema da degradação do desempenho no treinamento de redes muito profundas por meio da introdução de conexões de atalho ou residuais (*skip connections*), que permitem que o gradiente flua mais facilmente através da rede, facilitando o treinamento de arquiteturas com centenas ou até milhares de camadas (LI *et al.*, 2022; DESAI; SHAH, 2021). Essas inovações foram cruciais para alcançar avanços significativos em tarefas de visão computacional.

4.3 Trabalhos Relacionados

Nesta seção, são apresentados trabalhos da literatura que contextualizam e fundamentam a presente pesquisa. Tais trabalhos analisam o cenário da detecção de AD e seus desafios, o desenvolvimento de bases de dados para a área, e estudos que aplicam modelos de aprendizado profundo, como CNNs, sobre representações tempo-frequência. Tais representações, como espectrogramas e escalogramas, convertem o sinal de áudio para um formato de imagem, permitindo a identificação de padrões de manipulação. O conjunto desses trabalhos estabelece o contexto do problema, a importância dos recursos empregados e a validade da estratégia metodológica aqui proposta.

Um dos levantamentos mais relevantes é o de KHANJANI *et al.* (KHANJANI *et al.*, 2023), que apresenta um panorama sobre *deepfakes*, com foco especial nos de áudio — uma área, segundo os autores, negligenciada em comparação com as manipulações de vídeo. O estudo cataloga as principais técnicas de geração e detecção, destacando que abordagens baseadas em GANs e CNNs são comuns em ambos os processos. A principal contribuição desse trabalho é evidenciar a necessidade de mais pesquisas dedicadas à detecção de AD, justificando

a importância de estudos como o que é proposto nesta dissertação.

Na mesma linha, ALMUTAIRI; ELGIBREEN (ALMUTAIRI; ELGIBREEN, 2022) apresentam uma revisão dos métodos modernos de detecção de AD, analisando os desafios e as direções futuras da área. Os autores concluem que, embora os métodos baseados em DL sejam promissores, a capacidade de generalização para novos tipos de ataques e a robustez em ambientes ruidosos ainda são desafios significativos. O trabalho reforça a necessidade de desenvolver modelos de detecção mais adaptáveis, o que motiva a abordagem comparativa adotada neste trabalho para encontrar combinações de características e classificadores mais eficazes.

Para que o desenvolvimento e a avaliação de novas técnicas sejam possíveis, é necessária a existência de conjuntos de dados adequados. O trabalho de REIMAO; TZERPOS (REIMAO; TZERPOS, 2019) aborda essa questão ao apresentar a base de dados FoR. Os autores criaram este *dataset* para suprir a carência de bases públicas que contivessem áudios sintéticos gerados por algoritmos de DL de última geração. A robustez do conjunto foi corroborada pelos próprios autores, que em seus experimentos de validação alcançaram acurácias de até 99,96% na distinção entre fala real e sintética utilizando modelos de DL. Esses resultados atestam a qualidade do *dataset*, consolidando-o como um *benchmark* relevante para a área.

PHAM *et al.* (PHAM *et al.*, 2024) realizaram um estudo comparativo abrangente de diferentes representações espectrais para a detecção de AD. Os autores transformaram os sinais de áudio em espectrogramas utilizando métodos como a Transformada de Fourier de Curto Termo (STFT), a Transformada Q-Constante (do inglês, Constant-Q Transform, (CQT)) e a própria Transformada Wavelet (WT). Em seguida, avaliaram uma vasta gama de modelos de aprendizado profundo, incluindo CNNs, RNNs e arquiteturas pré-treinadas. O melhor sistema, um *ensemble* de diferentes modelos e características, alcançou um resultado de ponta no *dataset* ASVspoof 2019, com um EER de 0,03. Este trabalho reforça a validade de se explorar representações tempo-frequência e, especificamente, o uso da WT, como uma estratégia eficaz para a extração de características para a detecção de áudios falsos.

Em trabalho similar, VALENTE *et al.* (VALENTE *et al.*, 2024) propuseram a aplicação de Redes Neurais Convolucionais e espectrogramas Mel para a detecção de vozes geradas artificialmente. Foram conduzidos experimentos com diversas bases de dados, incluindo a FoR, ASVspoof e WaveFake, com o objetivo de encontrar uma topologia de CNN eficaz. O estudo demonstrou a viabilidade da abordagem, alcançando uma acurácia de 99% para a base de

dados FoR. Este resultado é particularmente relevante, pois corrobora a eficácia da estratégia de utilizar CNNs para a classificação de representações tempo-frequência no mesmo *dataset* empregado neste TCC.

Em (SCARPINITI *et al.*, 2024), foi proposta uma metodologia para a classificação automática de sons em canteiros de obras. O estudo abordou a limitação da representação por espectrogramas, que, por serem gerados via STFT, possuem uma resolução fixa que impõe um compromisso entre a localização temporal e a espectral dos eventos acústicos. Para superar essa limitação, os autores propuseram o uso de escalogramas, gerados a partir da CWT, como entrada para uma CNN. A justificativa apresentada é que a CWT, por sua natureza de análise multirresolução, gera uma representação mais detalhada e robusta, capaz de capturar com mais fidelidade as características de sinais do mundo real. Os resultados experimentais demonstraram a eficácia da abordagem, que superou o desempenho de outras soluções de ponta baseadas em espectrogramas. Este trabalho é relevante, pois valida a escolha de utilizar escalogramas como uma representação de características potencialmente superior para a análise de sinais de áudio complexos como a fala.

KISKIN *et al.* (KISKIN *et al.*, 2020) propôs uma abordagem baseada em CNN condicionada por transformadas wavelet para a detecção de eventos acústicos em contextos de baixa relação sinal-ruído. Os autores demonstraram que representações geradas via CWT superam aquelas obtidas por meio da STFT, tanto em termos de desempenho quanto de capacidade de generalização entre diferentes conjuntos de dados. Mesmo em cenários com poucos dados rotulados, o modelo em questão foi capaz de superar classificadores tradicionais e até mesmo anotadores humanos, com precisão superior a 90% na detecção de mosquitos e espécies de pássaros.

Em outra outro trabalho, CANTüRK; GüNAY (CANTüRK; GüNAY, 2024) utilizaram escalogramas gerados via CWT de sinais de voz para diagnosticar a Doença de Parkinson com auxílio de técnicas de aprendizado profundo. O estudo avaliou arquiteturas de CNNs, incluindo AlexNet, GoogleNet e ResNet50, além de propor um sistema híbrido baseado em votação majoritária. Os melhores obtidos foram pela abordagem de *deep feature fusion*, que emprega as CNNs DenseNet e NasNet como extratores de características, e KNN como classificador final. Essa configuração de modelo alcançou acurácia e *F1-score* de 0,95, evidenciando o potencial dos escalogramas como representações discriminativas eficazes.

Em suma, a literatura existente estabelece tanto a urgência do problema de detecção

de AD quanto a viabilidade da abordagem metodológica aqui proposta. Os levantamentos indicam uma área de pesquisa ativa e com desafios abertos, enquanto estudos de caso, tanto no domínio específico do problema quanto em áreas correlatas, demonstram o sucesso da combinação de CWT e CNNs. Este trabalho se insere nesse contexto, aplicando e avaliando sistematicamente essa metodologia promissora no domínio da detecção de fala sintética.

5 METODOLOGIA

A metodologia proposta foi dividida nas etapas de preparação dos dados e desenvolvimento do modelo. As seções seguintes detalham essas etapas. A Seção 5.1 oferece uma visão geral do fluxo de trabalho. A Seção 5.2 descreve a base de dados utilizada, o pré-processamento dos áudios, a geração dos escalogramas com diferentes *wavelets* e a organização dos conjuntos de dados. A Seção 5.3 detalha o modelo de detecção, apresentando a arquitetura de Rede Neural Convolucional adotada e as configurações de seu treinamento. As métricas utilizadas para a avaliação de desempenho são explicadas na Seção 5.4. Por fim, a Seção 5.5 lista as ferramentas e o ambiente computacional empregados nos experimentos.

5.1 Visão Geral

A metodologia combina a Transformada *Wavelet* (WT) para extração de características dos sinais de áudio e uma Rede Neural Convolucional (CNN) para classificá-los como reais ou falsos. O objetivo é analisar o desempenho do modelo ao ser treinado e validado com áudios representados por diferentes famílias de *wavelets*, a fim de identificar a abordagem mais eficiente. A Figura 3 apresenta o fluxo geral da metodologia.

Figura 3 – Fluxo geral da metodologia proposta

Fonte: próprio autor.

5.2 Preparação dos Dados

5.2.1 Base de Dados FoR

A base de dados *Fake or Real* (FoR) foi desenvolvida para suprir lacunas encontradas em outras bases quanto à disponibilidade de dados mais abrangentes e representativos. Na época de sua criação, muitas dessas bases não contemplavam os algoritmos de síntese considerados mais modernos e apresentavam um volume limitado de amostras — insuficiente para treinar modelos complexos de DL —, ou eram primariamente focadas em ataques de reprodução (*replay*) em vez de fala puramente sintética (REIMAO; TZERPOS, 2019).

Essa base contêm cerca de 198.000 instâncias, dos quais cerca de 87.000 são sintéticos e cerca de 111.000 são reais. Essa vasta coleção de dados é crucial, pois algoritmos de DL, especialmente redes neurais profundas, requerem grandes volumes de dados para alcançar

generalização e desempenho ótimos (REIMAO; TZERPOS, 2019). As amostras sintéticas foram geradas a partir de 33 vozes provenientes de diferentes sistemas de conversão de texto em fala (TTS), incluindo tecnologias de código aberto e comerciais como DeepVoice3, Google Wavenet, Amazon Polly e Microsoft Azure. Essa diversidade de fontes de síntese expõe os modelos de detecção a uma variedade de artefatos e características típicas das falas geradas artificialmente. Em contrapartida, os enunciados reais foram coletados de fontes públicas reconhecidas, como LJSpeech, Arctic e VoxForge, além de gravações de vídeos educacionais (provenientes de plataformas como TED Talks e YouTube (REIMAO; TZERPOS, 2019)), proporcionando vozes de diversos gêneros, idades, sotaques e qualidade de gravação. Essa heterogeneidade nos dados reais ajuda a construir detectores mais robustos e menos suscetíveis a falsos positivos causados por variações naturais da fala.

A base FoR foi submetida a um cuidadoso processo de balanceamento e normalização para minimizar vieses e garantir a comparabilidade dos experimentos. Todos os áudios foram convertidos para o formato WAV, com normalização de volume, padronização da taxa de amostragem para 16kHz e conversão para canal monofônico. A base foi disponibilizada em quatro versões distintas (sendo utilizada neste trabalho a for-2seconds) atendendo a diferentes necessidades e cenários de ataque:

- *for-original*: contém os enunciados em seu formato original de coleta, sem modificações;
- *for-norm*: apresenta os áudios normalizados, convertidos para o formato padrão e balanceados em termos de classe (real/falso) e gênero;
- *for-2seconds*: uma versão onde os áudios da *for-norm* são truncados ou preenchidos para terem exatamente dois segundos de duração, facilitando o processamento em *batches* por modelos de DL e incluindo balanceamento adicional;
- *for-rerecorded*: composta por áudios da versão *for-2seconds* que foram regravados utilizando um sistema de alto-falante e microfone, simulando um cenário de ataque onde o áudio sintético é reproduzido e recapturado em um ambiente físico.

A Tabela 2 resume as principais características, a composição e os parâmetros de padronização da base de dados FoR.

Tabela 2 – Base de dados FoR.

Característica	**Descrição / Quantidade**
Total de Enunciados	$\approx$ 198.000
Enunciados Sintéticos	$\approx$ 87.000
Enunciados Reais	$\approx$ 111.000
Parâmetros de Normalização	
Formato	WAV
Canais	Monofônico
Taxa de Amostragem	16 kHz
Amplitude	Normalização de volume

Fonte: (REIMAO; TZERPOS, 2019).

5.2.2 Pré-processamento e Geração dos Escalogramas

Os arquivos de áudio da versão *for-2seconds* foram processados individualmente. Conforme a implementação, cada sinal foi inicialmente normalizado para que sua amplitude máxima estivesse no intervalo [-1, 1]. Então, a Transformada Wavelet Contínua (CWT) foi aplicada para gerar as representações tempo-frequência dos sinais, na forma de escalogramas. A escolha por esta representação, em detrimento dos espectrogramas gerados via Transformada de Fourier de Curto Termo (STFT), se justifica pela capacidade da CWT de superar a limitação de resolução fixa da STFT. Ao oferecer uma análise multirresolução, o escalograma pode fornecer uma representação mais robusta para a classificação de sinais de áudio complexos com CNNs (SCARPINITI *et al.*, 2024). Para esta transformação, utilizou-se um conjunto de 127 escalas, permitindo a análise do sinal em múltiplos níveis de detalhe.

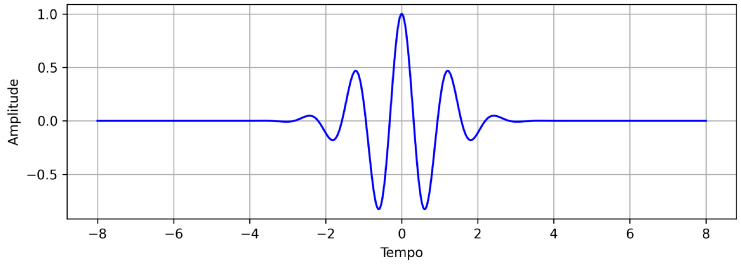
Para realizar uma análise comparativa, foram escolhidas três famílias distintas de *wavelets*, cada uma com propriedades matemáticas específicas para extrair diferentes tipos de características do sinal:

- Morlet (morl): Uma das *wavelets* mais utilizadas, é conhecida por oferecer um excelente equilíbrio entre a resolução no tempo e na frequência. Sua forma se assemelha a uma onda senoidal modulada por um envelope Gaussiano.
- Mexican Hat (mexh): Matematicamente, corresponde à segunda derivada de uma função Gaussiana. Seu formato de "chapéu"é particularmente eficaz na detecção de transientes, picos e descontinuidades no sinal, podendo destacar eventos de curta duração.
- Complex Morlet (cmor1.5-1.0): Uma variante da Morlet que é explicitamente complexa. Foi escolhida para investigar se suas propriedades de representação tempo-frequência

mais ricas poderiam revelar artefatos de síntese distintos daqueles capturados pelas outras *wavelets*.

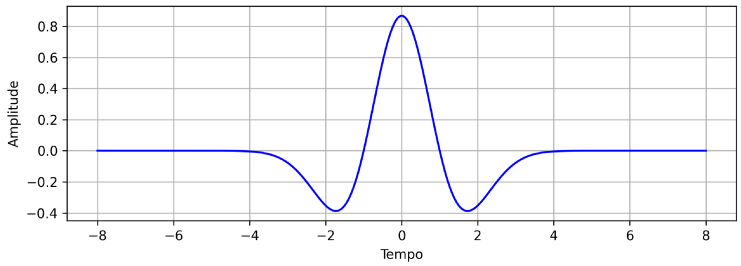
As formas de onda típicas das wavelets Morlet e Mexican Hat, que definem como elas analisam o sinal, são ilustradas nas Figuras 4 e 5, respectivamente. A wavelet Complex Morlet não é exibida, pois sua parte real é análoga à da wavelet Morlet.

Figura 4 – Forma de onda da wavelet Morlet.



Fonte: próprio autor.

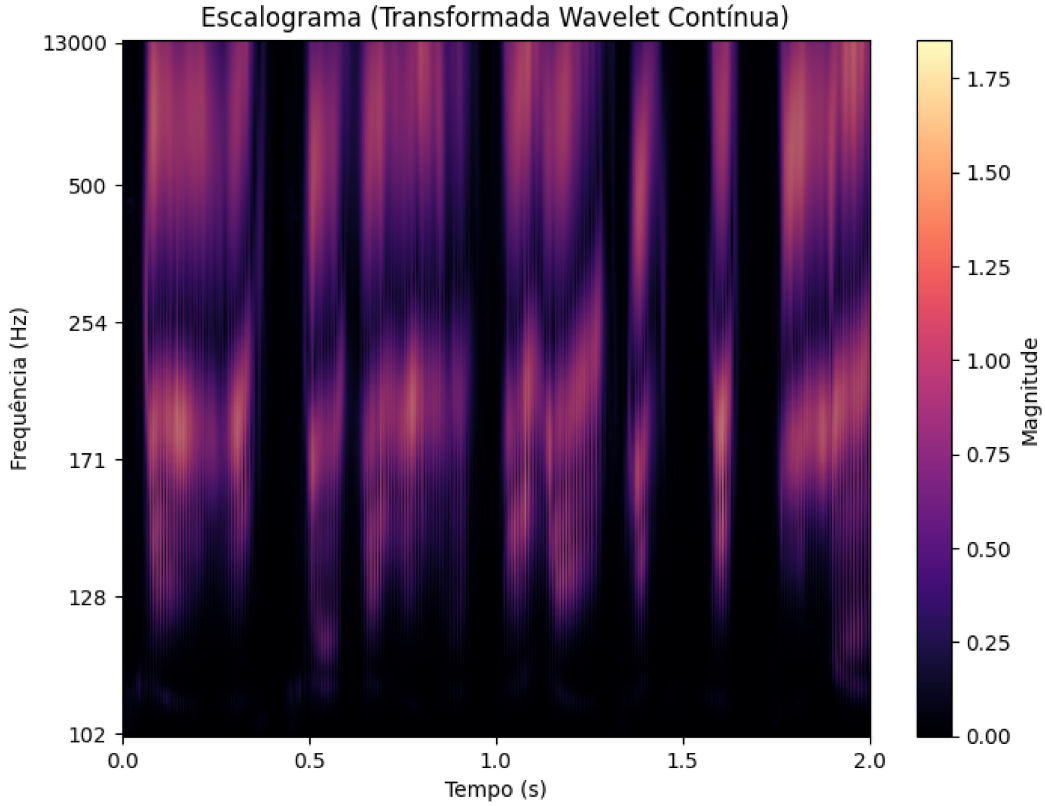
Figura 5 – Forma de onda da wavelet Mexican Hat.



Fonte: próprio autor.

O resultado da CWT é uma matriz de coeficientes. A magnitude desses coeficientes, obtida pelo cálculo de seu valor absoluto, representa a energia do sinal em cada ponto do plano tempo-frequência e é utilizada para gerar a imagem do escalograma. É importante notar que, como o sinal de áudio de entrada foi normalizado, os valores de magnitude resultantes são relativos e padronizados. Para a tarefa de classificação com CNNs, o mais importante não são os valores numéricos absolutos da magnitude, mas sim os padrões visuais e as texturas formadas pelas áreas de alta e baixa energia, que podem conter as "assinaturas"da manipulação sintética. A Figura 6 ilustra um exemplo de escalograma gerado, onde os eixos foram mantidos para fins de visualização e análise. No entanto, para o treinamento do modelo, os escalogramas foram gerados sem eixos ou margens, contendo apenas a informação visual da matriz de coeficientes.

Figura 6 – Escalograma gerado a partir de um sinal de áudio



Fonte: próprio autor.

Dessa forma, cada arquivo de áudio da base de dados originou três imagens de escalogramas, uma para cada família de *wavelet*, formando três conjuntos de dados distintos para a etapa de treinamento.

5.2.3 Organização dos Conjuntos de Dados

A organização dos dados seguiu a estrutura de divisão pré-definida pela própria base de dados FoR, que já separa os arquivos de áudio em conjuntos de treinamento, validação e teste. Os escalogramas foram gerados a partir desses subconjuntos originais, respeitando a divisão para garantir uma avaliação consistente com a proposta da base. A distribuição de imagens resultante foi a seguinte:

- Treinamento (70%): constituído por 13.956 imagens (6.978 de cada classe), utilizado para o ajuste dos pesos da rede neural.
- Validação (15%): composto por 2.826 imagens (1.413 de cada classe), utilizado para monitorar o desempenho e ajustar hiperparâmetros durante o treinamento.
- Teste (15%): totalizando 1.088 imagens (544 de cada classe), utilizado para a avaliação

final do modelo com dados não vistos anteriormente.

5.3 Modelo de Detecção Proposto

Para a tarefa de classificação dos escalogramas, foi utilizada a arquitetura de CNN MobileNet por meio da técnica de aprendizagem por transferência (*transfer learning*).

5.3.1 Arquitetura e Aprendizagem por Transferência

A aprendizagem por transferência é uma técnica que reutiliza modelos de redes neurais pré-treinados em uma tarefa para que possam operar em um novo conjunto de dados. O principal benefício dessa abordagem é a economia de recursos computacionais, uma vez que o conhecimento do modelo é, em grande parte, reaproveitado. As estratégias podem variar desde o retreinamento completo da rede até o ajuste de apenas partes específicas, como as camadas de classificação (TSALERA *et al.*, 2021).

O benefício de utilizar a aprendizagem por transferência, mesmo entre domínios aparentemente distintos como imagens naturais e escalogramas de áudio, reside no fato de que as camadas iniciais de uma CNN aprendem a detectar características de baixo nível, como bordas, texturas e formas. Essas características fundamentais são genéricas o suficiente para serem úteis na identificação de padrões visuais em representações tempo-frequência, o que acelera o treinamento e pode melhorar a capacidade de generalização do modelo (TSALERA *et al.*, 2021; KISKIN *et al.*, 2020). A arquitetura escolhida foi a MobileNet, reconhecida por sua eficiência computacional.

O modelo foi construído da seguinte forma:

1. A base convolucional da MobileNet, pré-treinada no conjunto de dados ImageNet para a tarefa de classificação de imagens naturais, foi utilizada como extrator de características. Suas camadas superiores, responsáveis pela classificação original, foram removidas.
2. Um novo classificador (ou "cabeça") foi adicionado no topo da base. Esse classificador é composto por uma camada de *GlobalAveragePooling2D*, seguida por uma camada densa com 128 neurônios e ativação ReLU, uma camada de *BatchNormalization*, uma camada de *Dropout* com taxa de 0.5 para regularização, e, por fim, uma camada de saída com um único neurônio e ativação sigmoide, adequada para a classificação binária.

5.3.2 Estratégia de Treinamento

O treinamento do modelo foi realizado em duas etapas distintas para garantir uma adaptação estável e eficaz à nova tarefa:

- Etapa 1: Treinamento do Classificador. Inicialmente, os pesos da base convolucional da MobileNet foram "congelados"(definidos como não treináveis). Apenas os pesos do novo classificador adicionado foram treinados com os dados de escalogramas.
- Etapa 2: Ajuste Fino (*Fine-Tuning*). Após o treinamento inicial do classificador, as 30 camadas superiores da base convolucional foram "descongeladas"(tornadas treináveis). O modelo foi então re-compilado com uma taxa de aprendizado muito baixa e o treinamento foi retomado para permitir que o modelo ajuste sutilmente suas características de mais alto nível.

5.3.3 Hiperparâmetros

O processo de treinamento e validação foi configurado com os hiperparâmetros detalhados na Tabela 3. Para evitar o sobreajuste (*overfitting*), foi utilizada a técnica de parada antecipada (*early stopping*), monitorando a acurácia no conjunto de validação (*val_accuracy*) para salvar e reter o modelo com o melhor desempenho.

Tabela 3 – Hiperparâmetros utilizados no treinamento do modelo MobileNet

Hiperparâmetro	**Valor / Configuração**
Arquitetura Base	MobileNet (pré-treinada com ImageNet)
Otimizador	Adam
Função de Perda	Entropia Cruzada Binária (*Binary Crossentropy*)
Batch Size	32
— Etapa 1: Treinamento do Classificador —	
Taxa de Aprendizado	Padrão do Adam (0.001)
Épocas Máximas	15
Patience (Parada Antecipada)	5 épocas
Métrica Monitorada	Acurácia de Validação (*val_accuracy*)
— Etapa 2: Fine-Tuning —	
Taxa de Aprendizado	1e-5 (0.00001)
Épocas Máximas	50 (total)
Patience (Parada Antecipada)	8 épocas
Patience (Redução de LR)	3 épocas
Métrica Monitorada	Acurácia de Validação (*val_accuracy*)
— Configuração de Entrada —	
Tamanho de Entrada (Escalograma)	224 x 224 pixels
Canais de Entrada (Escalograma)	3 (RGB)

Fonte: Extraído do código-fonte do projeto.

5.4 Métricas de Avaliação de Desempenho

O desempenho do modelo foi avaliado através de métricas padrão para problemas de classificação. Para garantir que a avaliação final seja realizada com o melhor resultado de treinamento, os pesos salvos no ponto de maior acurácia de validação são recarregados antes da predição no conjunto de teste. A análise dos resultados é, então, baseada na matriz de confusão e em métricas derivadas.

A matriz de confusão permite visualizar o desempenho de um algoritmo de classificação, organizando as predições em quatro categorias: Verdadeiro Positivo (VP), Falso Positivo (FP), Verdadeiro Negativo (VN) e Falso Negativo (FN). A Tabela 4 ilustra a estrutura de uma matriz de confusão.

Tabela 4 – Estrutura da matriz de confusão

		Predito pelo Modelo	
		Positivo	Negativo
Real	Positivo	VP (Verdadeiro Positivo)	FN (Falso Negativo)
	Negativo	FP (Falso Positivo)	VN (Verdadeiro Negativo)

Fonte: próprio autor.

A partir dos quatro parâmetros da matriz de confusão, derivam as métricas presentes na Tabela 5. A comparação final entre os modelos treinados com as diferentes famílias de *wavelets* será baseada nessas métricas, nas predições feitas com o conjunto de teste.

Tabela 5 – Métricas de avaliação derivadas da matriz de confusão

Métrica	**Fórmula**	**Descrição**
Acurácia	$\frac{VP+VN}{VP+FP+VN+FN}$	Mede a proporção geral de previsões corretas.
Precisão	$\frac{VP}{VP+FP}$	Das amostras classificadas como positivas, indica a proporção das que eram de fato positivas.
Revocação (Recall)	$\frac{VP}{VP+FN}$	Das amostras que eram de fato positivas, indica a proporção que o modelo conseguiu identificar.
F1-Score	$2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}}$	Média harmônica entre Precisão e Revocação.
AUC-ROC	–	Mede a área sob a curva ROC (Receiver Operating Characteristic), que avalia a capacidade do modelo de distinguir entre as classes.

Fonte: Próprio autor.

5.5 Ferramentas e Ambiente de Execução

Todos os modelos e experimentos foram implementados e executados no ambiente Google Colab Pro, implementados em linguagem Python (versão 3.11). Para a manipulação dos sinais de áudio e da Transformada Wavelet Contínua (CWT), foram utilizadas as bibliotecas SciPy (VIRTANEN *et al.*, 2020) para algoritmos científicos e PyWavelets (LEE *et al.*, 2019) para a implementação das transformadas. Para o carregamento dos arquivos de áudio, empregou-se a biblioteca Librosa (MCFEE *et al.*, 2015).

A geração de todas as visualizações gráficas, como os escalogramas e as matrizes de confusão, foi realizada com o Matplotlib (HUNTER, 2007), com o suporte da biblioteca Seaborn (WASKOM, 2021) para a criação de gráficos estatísticos.

Para a construção e o treinamento dos modelos de Rede Neural Convolucional, foi utilizada a plataforma TensorFlow (ABADI *et al.*, 2016) e sua API de alto nível Keras (CHOLLET *et al.*, 2015). O cálculo das métricas finais de avaliação de desempenho, como acurácia, precisão, revocação e F1-Score, foi realizado com as funções da biblioteca Scikit-learn (PEDREGOSA *et al.*, 2011).

6 RESULTADOS E DISCUSSÃO

Os resultados foram organizados em seções, cada uma dedicada a uma das três famílias de *wavelets* utilizadas para gerar os escalogramas. Ao final do capítulo, uma análise comparativa é realizada para determinar qual família de *wavelet* proporcionou a representação de características mais discriminativa para a tarefa.

6.1 Wavelet Morlet

O primeiro experimento utilizou os escalogramas gerados com a família de *wavelets* Morlet. A Tabela 6 resume as métricas de desempenho do modelo no conjunto de teste.

Tabela 6 – Resultados do modelo MobileNet com escalogramas Morlet

Métrica	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**	**AUC-ROC**
Valor	0.8401	0.8304	0.8548	0.8424	0.9149

Fonte: próprio autor.

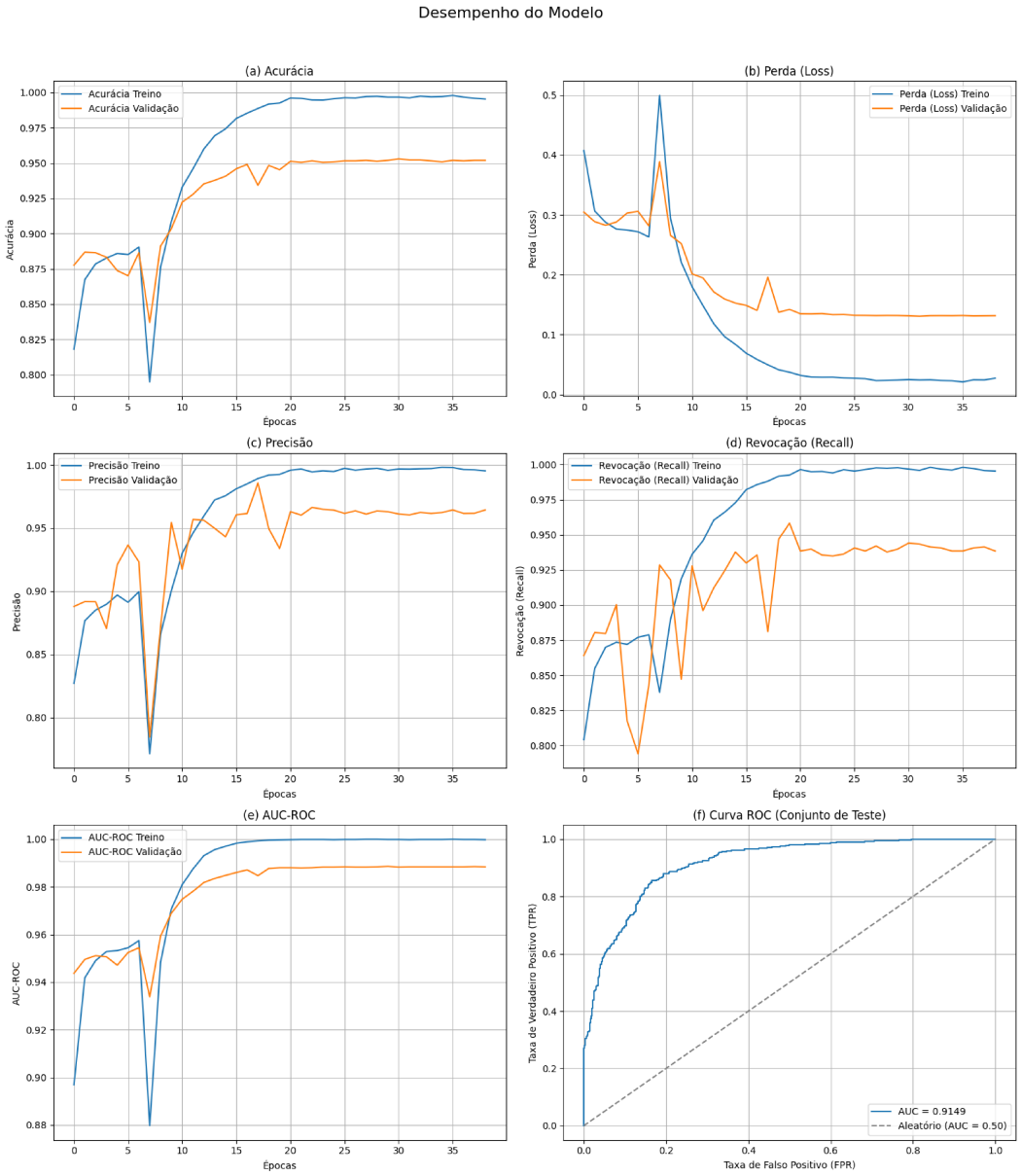
A *wavelet* Morlet demonstrou um desempenho geral consistente e equilibrado, alcançando uma acurácia de 84,01% e um F1-Score de 84,24%. A análise detalhada das curvas de treinamento, exibidas na Figura 7, fornece mais detalhes sobre o comportamento do modelo. As curvas de acurácia (a) e perda (b) mostram que, após uma instabilidade inicial que coincide com o início da fase de *fine-tuning*, por volta da época 7, o modelo converge de forma estável, com a performance de validação acompanhando a de treino, embora com um pequeno *gap* que indica um leve e controlado sobreajuste. As curvas de precisão (c) e revocação (d) seguem um padrão similar de estabilização pós-instabilidade.

A matriz de confusão (Figura 8), corrobora o bom equilíbrio do modelo, com um número similar de falsos positivos (95) e falsos negativos (79). Isso indica que o modelo não possui um viés forte para um tipo específico de erro, o que é refletido nos valores balanceados de Precisão (83,04%) e Revocação (85,48%). Finalmente, o poder discriminativo do classificador é atestado pela curva ROC do conjunto de teste (Figura 7(f)), que apresenta uma área sob a curva (AUC) de 0,9149, confirmando a alta capacidade do modelo em separar as classes real e falsa com esta representação.

Esses resultados correspondem às expectativas teóricas para a *wavelet* Morlet. Conforme discutido anteriormente, sua principal característica é o bom equilíbrio entre a resolução

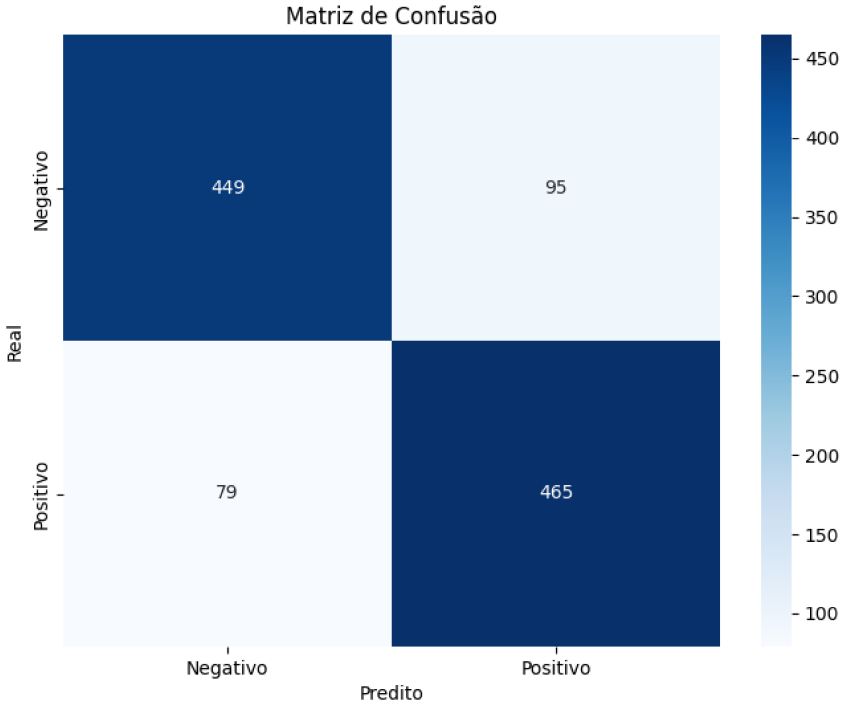
no tempo e na frequência, o que a torna ideal para analisar sinais com características oscilatórias e distribuídas, como a fala. Esperava-se que essa propriedade fosse mais adequada para capturar as anomalias sutis da fala sintética do que *wavelets* especializadas em transientes, como a Mexican Hat. O desempenho superior e o perfil de erro balanceado obtidos experimentalmente validam essa hipótese, indicando que a representação gerada pela Morlet foi a que melhor extraiu as características discriminativas para a CNN.

Figura 7 – Curvas de desempenho do modelo durante o treinamento com escalogramas Morlet



Fonte: próprio autor.

Figura 8 – Matriz de Confusão para o modelo com escalogramas Morlet



Fonte: próprio autor.

6.2 Wavelet Complex Morlet

O segundo experimento utilizou os escalogramas gerados pela *wavelet* Complex Morlet. Os resultados no conjunto de teste são apresentados na Tabela 7.

Tabela 7 – Resultados do modelo MobileNet com escalogramas Complex Morlet

Métrica	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**	**AUC-ROC**
Valor	0.8208	0.7555	0.9485	0.8411	0.9233

Fonte: próprio autor.

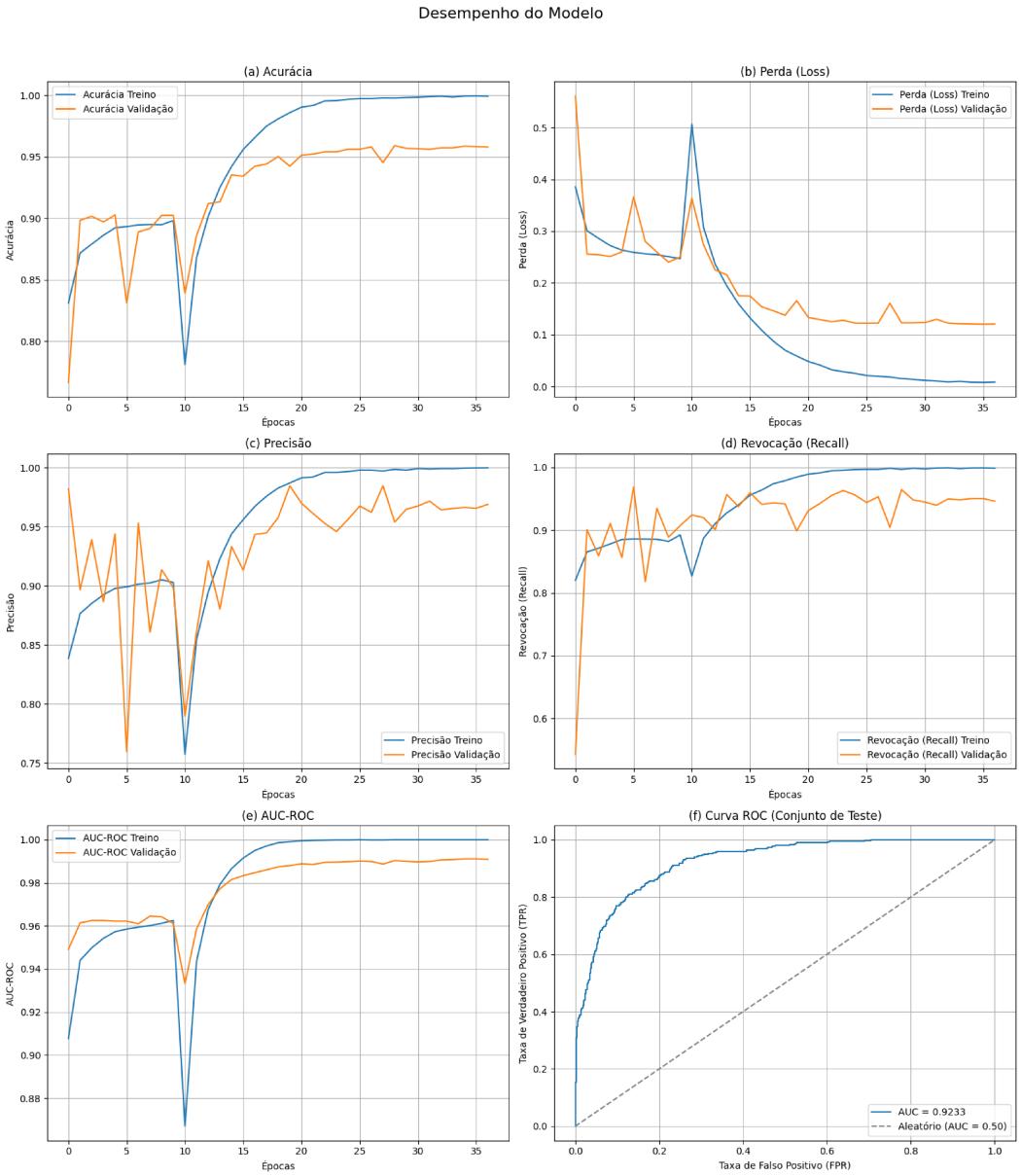
A *wavelet* Complex Morlet revelou um modelo com um perfil de desempenho bastante particular. As curvas de treinamento (Figura 9) foram as mais voláteis entre os três experimentos, especialmente as de validação, como pode ser visto nas curvas de precisão (c) e revocação (d). Essa instabilidade sugere que o modelo teve mais dificuldade em convergir para uma solução estável utilizando as características extraídas por esta *wavelet*.

A matriz de confusão (Figura 10) mostra que o modelo final se especializou na identificação de áudios falsos. Ele alcançou a maior taxa de Revocação (94,85%) entre todos os experimentos, cometendo apenas 28 Falsos Negativos. Contudo, essa alta sensibilidade ocorreu

ao custo de uma Precisão mais baixa (75,55%), com um número elevado de Falsos Positivos (167). Este perfil de erro desbalanceado caracteriza um modelo "vigilante", que prioriza a detecção de ameaças sobre a correta classificação de amostras reais. O resultado mais notável é visto na curva ROC (Figura 9(f)): apesar do desequilíbrio nos erros, o modelo atingiu a maior AUC (0,9233), indicando que esta representação possui o maior potencial discriminativo.

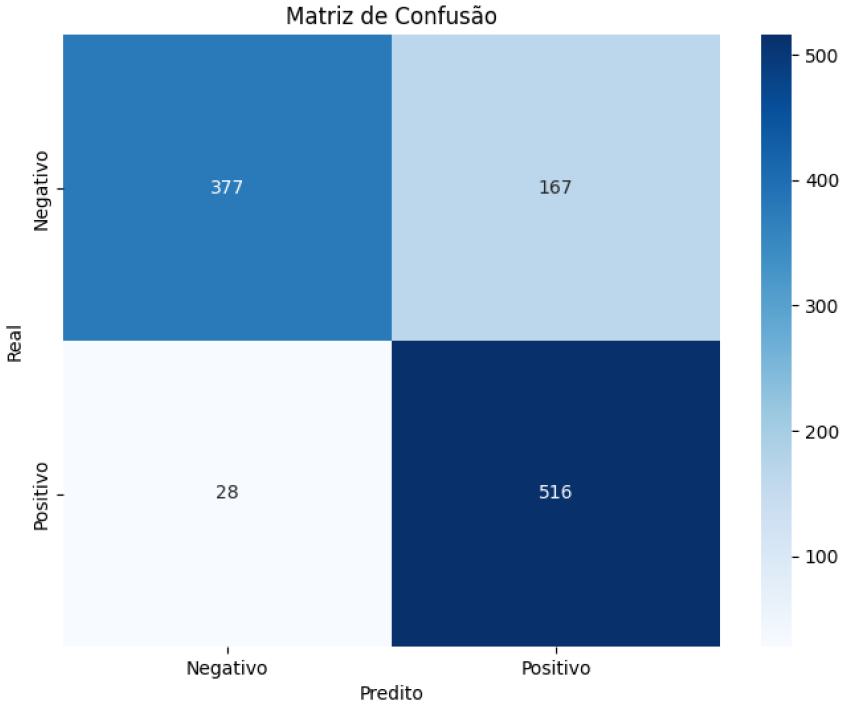
O perfil de erro de um classificador define sua adequação para diferentes cenários práticos. Em contextos de alta segurança, como análise forense ou detecção de fraudes, o erro mais crítico é o Falso Negativo — falhar em detectar um áudio falso, ou "inocentar um culpado". Com sua baixíssima taxa de Falsos Negativos, o modelo treinado com a *wavelet* Complex Morlet se mostra ideal para essas aplicações, pois prioriza a detecção de ameaças. Contudo, em outras situações, como a moderação de conteúdo em plataformas de mídia, o erro mais grave poderia ser o Falso Positivo — "acusar um inocente"—, que causaria transtornos desnecessários. Para esses cenários, o comportamento "vigilante"deste modelo, com seu alto número de Falsos Positivos, não seria a escolha mais adequada.

Figura 9 – Curvas de desempenho do modelo durante o treinamento com escalogramas Complex Morlet



Fonte: próprio autor.

Figura 10 – Matriz de Confusão para o modelo com escalogramas Complex Morlet



Fonte: próprio autor.

6.3 Wavelet Mexican Hat

O terceiro experimento utilizou os escalogramas gerados pela *wavelet* Mexican Hat. A Tabela 8 sintetiza os resultados.

Tabela 8 – Resultados do modelo MobileNet com escalogramas Mexican Hat

Métrica	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**	**AUC-ROC**
Valor	0.6765	0.6200	0.9118	0.7381	0.8436

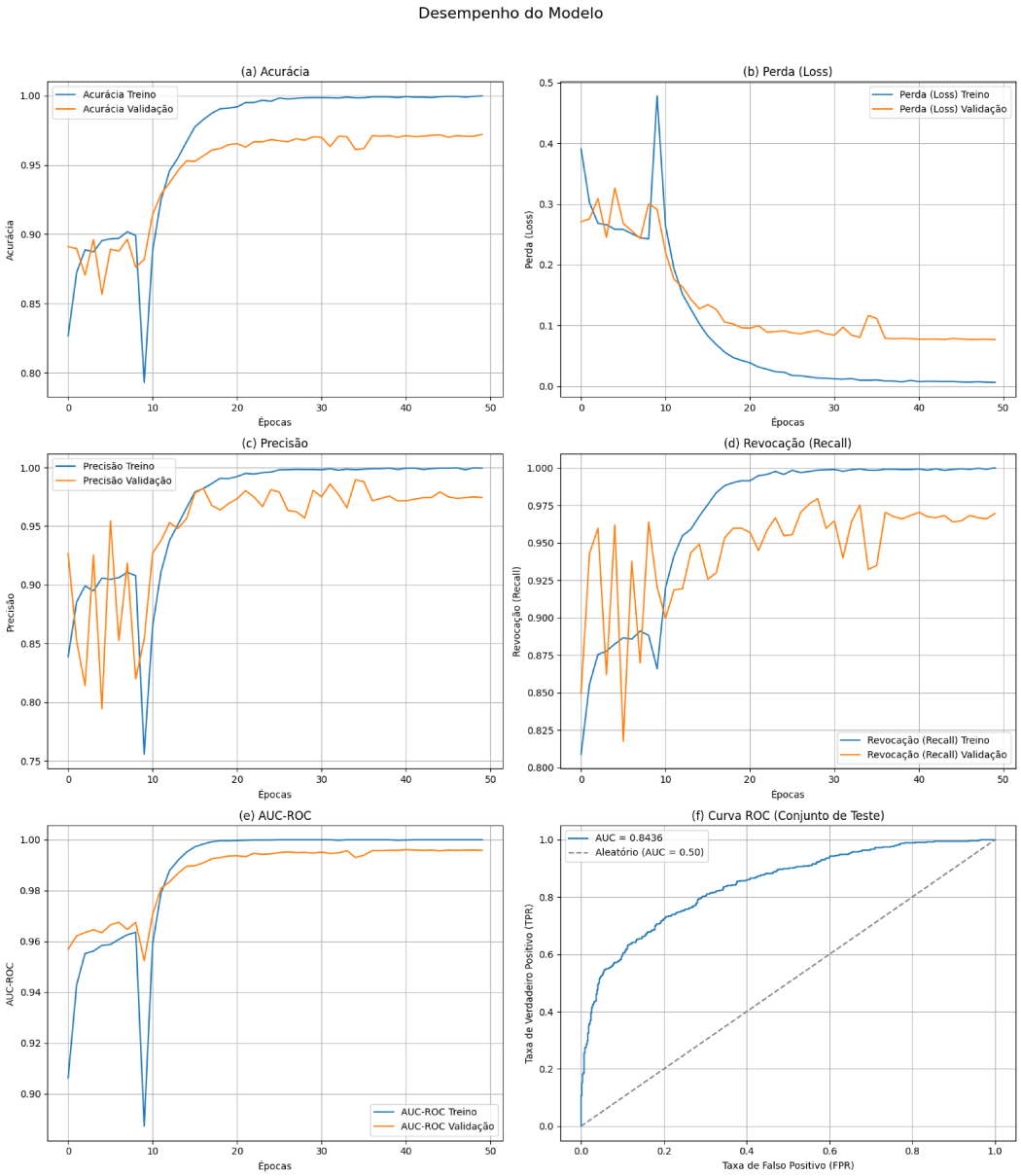
Fonte: próprio autor.

Os resultados alcançados demonstram que este foi o modelo de menor desempenho, alcançando uma acurácia de 67,65% e um F1-Score de 73,81%, um desfecho que pode ser considerado esperado dadas as características da *wavelet* Mexican Hat. Conforme discutido na metodologia, esta *wavelet* é particularmente eficaz na detecção de transientes e descontinuidades abruptas. No entanto, os artefatos da fala sintética são frequentemente mais sutis e distribuídos por todo o espectro. A especialização da Mexican Hat em picos e eventos de curta duração parece não ter sido adequada para capturar esses padrões complexos, o que se refletiu no desempenho inferior do classificador. As curvas de treinamento (Figura 11) ajudam a diagnosticar o problema,

onde se observa um grande e persistente *gap* entre as métricas de treino e validação, um sinal de sobreajuste acentuado.

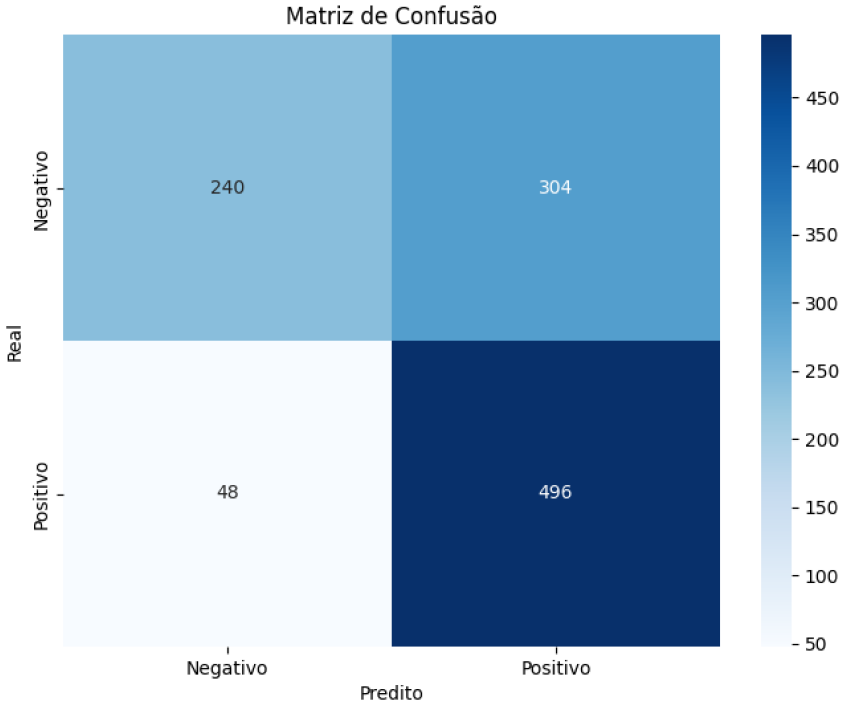
A matriz de confusão (Figura 12) esclarece a consequência desse aprendizado deficiente. O modelo desenvolveu um forte viés para classificar amostras como falsas, resultando em um número extremo de Falsos Positivos (304) e uma Especificidade de apenas 44%. Em uma situação real de perícia forense, as consequências de utilizar este modelo seriam graves: ele acusaria incorretamente mais da metade dos áudios autênticos como sendo falsificados. Um sistema com uma taxa tão alta de alarmes falsos seria impraticável e perderia rapidamente a credibilidade, pois exigiria verificação manual excessiva e poderia levar a conclusões equivocadas. A alta taxa de Revocação (91,18%) torna-se, nesse contexto, uma métrica enganosa, pois é alcançada não por uma real capacidade de distinção, mas por uma tendência do modelo a prever majoritariamente uma única classe. A curva ROC (Figura 11(f)) e o valor de AUC de 0,8436 confirmam que as características extraídas pela Mexican Hat foram as menos eficazes para a tarefa.

Figura 11 – Curvas de desempenho do modelo durante o treinamento com escalogramas Mexican Hat



Fonte: próprio autor.

Figura 12 – Matriz de Confusão para o modelo com escalogramas Mexican Hat



Fonte: próprio autor.

6.4 Análise Comparativa

A Tabela 9 resume os valores de acurácia e F1-Score obtidos com cada família de *wavelet*.

Tabela 9 – Síntese das métricas do desempenho por família de wavelet

Família de Wavelet	**Acurácia Final**	**F1-Score Final**
Morlet	0.8401	0.8424
Complex Morlet	0.8208	0.8411
Mexican Hat	0.6765	0.7381

Fonte: próprio autor.

Esses resultados demonstram que a escolha da família de *wavelet* é um fator determinante no desempenho do modelo de detecção. A *wavelet* Morlet se destacou como a abordagem mais equilibrada e de melhor performance geral, alcançando a maior acurácia e o maior F1-Score. A estabilidade do seu treinamento e o perfil de erro balanceado sugerem que sua característica de boa resolução conjunta em tempo e frequência é a mais adequada para capturar as anomalias sutis e distribuídas da fala sintética.

A abordagem com a Complex Morlet é um caso interessante de compromisso (*trade-off*). Embora com uma acurácia geral menor, obteve a maior capacidade de identificar áudios falsos (maior revocação e maior AUC-ROC), mas ao custo de classificar erroneamente um número considerável de áudios reais. Tal modelo seria a escolha preferencial em aplicações de segurança crítica, onde a prioridade máxima é não permitir que nenhum falso negativo passe pelo sistema. Em uma situação real de perícia forense, por exemplo, a consequência de um falso negativo seria aceitar um áudio forjado como evidência genuína, o que poderia levar a uma falha judicial grave. Em contraste, um falso positivo significaria que um áudio autêntico seria sinalizado para uma verificação humana adicional. Embora isso gere mais trabalho para o perito, é um resultado muito menos danoso do que deixar uma evidência falsa passar despercebida.

Já a Mexican Hat mostrou-se inadequada para esta tarefa. Sua especialidade em detectar transientes e descontinuidades abruptas não performou bem na identificação dos padrões mais complexos e distribuídos da fala sintética, levando o modelo a aprender um viés simplista e pouco eficiente. Este resultado, ainda que negativo, é valioso por demonstrar que nem todas as representações tempo-frequência são igualmente úteis para o problema.

7 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho desenvolveu uma metodologia para a detecção de *deepfakes* de áudio, baseada em escalogramas *wavelet* por meio de Redes Neurais Convolucionais (CNNs). Três famílias de *wavelets* (Morlet, Complex Morlet e Mexican Hat) foram empregadas na geração das representações tempo-frequência, utilizadas no treinamento de um modelo MobileNet com aprendizagem por transferência.

A análise e comparação dos resultados indicam a viabilidade da abordagem proposta, demonstrando que a escolha da função *wavelet* impacta o desempenho e o perfil de erros do classificador. A representação com a *wavelet* Morlet resultou em um modelo com desempenho equilibrado, alcançando um F1-Score de 84,24% e uma acurácia de 84,01%. Já a *wavelet* Complex Morlet produziu o modelo com a maior taxa de revocação (94,85%) e AUC-ROC (0,9233), apesar da acurácia menor, indicando seu potencial para aplicações que priorizam a detecção de todas as amostras falsas, mesmo que isso aumente o número de alarmes falsos. A *wavelet* Mexican Hat mostrou-se a menos adequada, para a aplicação em questão, com um desempenho muito inferior às demais: acurácia de 67,65% e F1-Score de 73,81%

Os resultados deste estudo sugerem que a metodologia de conversão de sinais de áudio em escalogramas, seguida de classificação com CNNs pré-treinadas, é tecnicamente viável e promissora. Além disso, os resultados evidenciam que diferentes famílias de *wavelets* podem gerar modelos com características operacionais distintas, reforçando a importância de tratar a seleção da *wavelet* mãe como um hiperparâmetro relevante no projeto de sistemas de detecção de AD. Essa escolha influencia não apenas a precisão final, mas também o tipo de erro predominante do sistema, permitindo sua adaptação a diferentes cenários de aplicação, como segurança digital, validação de identidade ou análise forense.

7.1 Trabalhos Futuros

Dentre as diversas linhas de pesquisa futuras possíveis de serem exploradas para expandir e aprimorar a metodologia proposta destacamos para trabalhos futuros:.

- validar o melhor modelo encontrado em outras bases de dados de referência, como a ASVspoof para verificar a capacidade de generalização do sistema para diferentes tipos de ataques de síntese e condições acústicas, que é um dos principais desafios da área.
- investigar outras arquiteturas de CNN, como os modelos EfficientNet, ResNet ou até

mesmo arquiteturas mais recentes, poderia revelar se redes com maior capacidade de representação são capazes de alcançar um desempenho ainda superior na classificação dos escalogramas.

- apronfudar o processo de otimização de hiperparâmetros da rede neural, como taxa de aprendizado e otimizador, outras famílias de *wavelets*, como Daubechies ou Symlet, ou diferentes configurações de escalas na transformada.
- aumentar a robustez estatística da avaliação, através de validação cruzada, em substituição à atual divisão única em treino, validação e teste, para obter uma estimativa mais confiável da performance do modelo.

REFERÊNCIAS

ABADI, M.; BARHAM, P.; CHEN, J.; CHEN, Z.; DAVIS, A.; DEAN, J.; DEVIN, M.; GHEMAWAT, S.; IRVING, G.; ISARD, M. *et al.* **TensorFlow: Large-scale machine learning on heterogeneous distributed systems**. 2016. Versão do software utilizada: 2.19.0. Disponível em: <https://arxiv.org/abs/1603.04467>.

ALMUTAIRI, Z.; ELGIBREEN, H. A review of modern audio deepfake detection methods: Challenges and future directions. **Algorithms**, v. 15, n. 5, 2022. ISSN 1999-4893. Disponível em: <https://www.mdpi.com/1999-4893/15/5/155>.

BHATT, D.; PATEL, C.; TALSANIA, H.; PATEL, J.; VAGHELA, R.; PANDYA, S.; MODI, K.; GHAYVAT, H. Cnn variants for computer vision: History, architecture, application, challenges and future scope. **Electronics**, v. 10, n. 20, 2021. ISSN 2079-9292. Disponível em: <https://www.mdpi.com/2079-9292/10/20/2470>.

CANTüRK, ; GüNAY, O. Investigation of scalograms with a deep feature fusion approach for detection of parkinson's disease. **Cognitive Computation**, v. 16, n. 3, p. 1198–1209, 2024. ISSN 1866-9964. Disponível em: <https://doi.org/10.1007/s12559-024-10254-8>.

CHOLLET, F. *et al.* **Keras**. [S.l.]: GitHub, 2015. <https://github.com/keras-team/keras>. Versão do software utilizada: 3.8.0.

DAUER, L. **Inteligência artificial: deepfake já foi usada em eleições pelo mundo**. 2024. Acessado em 15 de maio de 2025. Disponível em: <https://noticias.uol.com.br/confere/ultimas-noticias/2024/03/03/deepfake-uso-inteligencia-artificial-eleicoes-argentina-estados-unidos.htm>.

DESAI, M.; SHAH, M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (mlp) and convolutional neural network (cnn). **Clinical eHealth**, v. 4, p. 1–11, 2021. ISSN 2588-9141. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2588914120300125>.

DIXIT, A.; KAUR, N.; KINGRA, S. Review of audio deepfake detection techniques: Issues and prospects. **Expert Systems**, v. 40, n. 8, p. e13322, 2023. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/exsy.13322>.

GRAPS, A. An introduction to wavelets. **IEEE Computational Science and Engineering**, v. 2, n. 2, p. 50–61, 1995.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science & Engineering**, v. 9, n. 3, p. 90–95, 2007. Versão do software utilizada: 3.10.0. Disponível em: <https://doi.org/10.1109/MCSE.2007.55>.

JUNG, J. weon; HEO, H.-S.; TAK, H.; SHIM, H. jin; CHUNG, J. S.; LEE, B.-J.; YU, H.-J.; EVANS, N. **AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks**. 2021. Disponível em: <https://arxiv.org/abs/2110.01200>.

KHANJANI, Z.; WATSON, G.; JANEJA, V. P. Audio deepfakes: A survey. **Frontiers in Big Data**, v. 5, 2023. ISSN 2624-909X. Disponível em: <https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2022.1001063>.

KISKIN, I.; ZILLI, D.; LI, Y.; SINKA, M.; WILLIS, K.; ROBERTS, S. Bioacoustic detection with wavelet-conditioned convolutional neural networks. **Neural Computing and Applications**, v. 32, n. 4, p. 915–927, 2020. ISSN 1433-3058. Disponível em: <https://doi.org/10.1007/s00521-018-3626-7>.

LEE, G. R.; GOMMERS, R.; WASELEWSKI, F.; WOHLFAHRT, K.; O'LEARY, A. PyWavelets: A Python package for wavelet analysis. **Journal of Open Source Software**, v. 4, n. 36, p. 1237, 2019. Versão do software utilizada: 1.8.0. Disponível em: <https://doi.org/10.21105/joss.01237>.

LI, Z.; LIU, F.; YANG, W.; PENG, S.; ZHOU, J. A survey of convolutional neural networks: Analysis, applications, and prospects. **IEEE Transactions on Neural Networks and Learning Systems**, v. 33, n. 12, p. 6999–7019, 2022.

MALLAT, S. A theory for multiresolution signal decomposition: the wavelet representation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 11, n. 7, p. 674–693, 1989.

MAWALIM, C. O.; WANG, Y.; ADILA, A.; OKADA, S.; UNOKI, M. Robust multilingual audio deepfake detection through hybrid modeling. In: **Proceedings of the 2025 ACM Workshop on Information Hiding and Multimedia Security**. New York, NY, USA: Association for Computing Machinery, 2025. (IHMMSEC '25), p. 181–192. ISBN 9798400718878. Disponível em: <https://doi.org/10.1145/3733102.3736706>.

MCFEE, B.; RAFFEL, C.; LIANG, D.; ELLIS, D. P. W.; MCVICAR, M.; BATTENBERG, E.; NIETO, O. librosa: Audio and music signal analysis in Python. In: **Proceedings of the 14th Python in Science Conference**. [s.n.], 2015. p. 18–25. Versão do software utilizada: 0.11.0. Disponível em: <https://doi.org/10.25080/Majora-7b98e3ed-003>.

O GLOBO. **Neymar, Paolla Oliveira e outros famosos viram vítimas de deepfake: Saiba os riscos e como se proteger**. 2025. Acessado em 15 de maio de 2025. Disponível em: <https://oglobo.globo.com/ela/noticia/2025/04/05/neymar-paolla-oliveira-e-outros-famosos-viram-vitimas-de-deepfake-saiba-os-riscos-e-como-se-proteger.ghtml>.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Versão do software utilizada: 1.6.1. Disponível em: <http://jmlr.org/papers/v12/pedregosa11a.html>.

PHAM, L.; LAM, P.; NGUYEN, T.; NGUYEN, H.; SCHINDLER, A. **Deepfake Audio Detection Using Spectrogram-based Feature and Ensemble of Deep Learning Models**. 2024. Disponível em: <https://arxiv.org/abs/2407.01777>.

RANGEL, R. F. **Transformadas Wavelet em Visão Computacional**. 2020. Turing Talks - Medium. Acessado em: 7 de julho de 2025. Disponível em: <https://medium.com/turing-talks/transformadas-wavelet-em-visÃčo-computacional-94c72d57e049>.

REIMAO, R.; TZERPOS, V. For: A dataset for synthetic speech detection. In: **2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)**. [S.l.: s.n.], 2019. p. 1–10.

RIOUL, O.; VETTERLI, M. Wavelets and signal processing. **Signal Processing Magazine, IEEE**, v. 8, p. 14 – 38, 10 1991.

SCARPINITI, M.; PARISI, R.; LEE, Y.-C. A scalogram-based cnn approach for audio classification in construction sites. **Applied Sciences**, v. 14, n. 1, 2024. ISSN 2076-3417. Disponível em: <https://www.mdpi.com/2076-3417/14/1/90>.

SHAABAN, O. A.; YILDIRIM, R.; ALGUTTAR, A. A. Audio deepfake approaches. **IEEE Access**, v. 11, p. 132652–132682, 2023.

TSALERA, E.; PAPADAKIS, A.; SAMARAKOU, M. Comparison of pre-trained cnns for audio classification using transfer learning. **Journal of Sensor and Actuator Networks**, v. 10, n. 4, 2021. ISSN 2224-2708. Disponível em: <https://www.mdpi.com/2224-2708/10/4/72>.

VALENTE, L. P.; SOUZA, M. M. S. de; ROCHA, A. M. D. Speech audio deepfake detection via convolutional neural networks. In: **2024 IEEE International Conference on Evolving and Adaptive Intelligent Systems (EAIS)**. [S.l.: s.n.], 2024. p. 1–6.

VARGAS, A. C. G.; PAES, A.; VASCONCELOS, C. N. Um estudo sobre redes neurais convolucionais e sua aplicação em detecção de pedestres. In: SN. **Proceedings of the xxix conference on graphics, patterns and images**. [S.l.], 2016. v. 1, n. 4.

VIRTANEN, P.; GOMMERS, R.; OLIPHANT, T. E.; HABERLAND, M.; REDDY, T.; COURNAPEAU, D.; BUROVSKI, E.; PETERSON, P.; WECKESSER, W.; BRIGHT, J.; van der Walt, S. J.; BRETT, M.; WILSON, J.; MILLMAN, K. J.; MAYOROV, N.; NELSON, A. R. J.; JONES, E.; KERN, R.; LARSON, E.; CAREY, C.; POLAT, I.; FENG, Y.; MOORE, E. W.; VanderPlas, J.; LAXALDE, D.; PERKTOLD, J.; CIMRMAN, R.; HENRIKSEN, I.; QUINTERO, E. A.; HARRIS, C. R.; ARCHIBALD, A. M.; RIBEIRO, A. H.; PEDREGOSA, F.; van Mulbregt, P.; SciPy 1.0 Contributors. SciPy 1.0: Fundamental algorithms for scientific computing in python. **Nature Methods**, v. 17, n. 3, p. 261–272, 2020. Versão do software utilizada: 1.15.3. Disponível em: <https://doi.org/10.1038/s41592-019-0686-2>.

WASKOM, M. L. seaborn: statistical data visualization. **Journal of Open Source Software**, v. 6, n. 60, p. 3021, 2021. Versão do software utilizada: 0.13.2. Disponível em: <https://doi.org/10.21105/joss.03021>.

YI, J.; WANG, C.; TAO, J.; ZHANG, X.; ZHANG, C. Y.; ZHAO, Y. **Audio Deepfake Detection: A Survey**. 2023. Disponível em: <https://arxiv.org/abs/2308.14970>.

ZHANG, D. Wavelet transform. In: ____. **Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval**. Cham: Springer International Publishing, 2019. p. 35–44. ISBN 978-3-030-17989-2. Disponível em: <https://doi.org/10.1007/978-3-030-17989-2_3>.

APÊNDICE A – REPOSITÓRIO DO CÓDIGO-FONTE

Todo o código-fonte desenvolvido para este trabalho, incluindo os notebooks para a geração dos escalogramas e para o treinamento do modelo de detecção, está disponível publicamente no seguinte repositório do GitHub:

<https://github.com/EmanuelDevid/TCC-Deteccao-Deepfake-Audio>