UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE SOBRAL

CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

VITOR HUGO MUNIZ DE SOUSA SANTOS

ESTRATÉGIAS DE AMPLIAÇÃO E BALANCEAMENTO DE DADOS PARA A DETECÇÃO DE ÁUDIOS VERDADEIROS E FALSOS

SOBRAL

2025

VITOR HUGO MUNIZ DE SOUSA SANTOS

ESTRATÉGIAS DE AMPLIAÇÃO E BALANCEAMENTO DE DADOS PARA A DETECÇÃO DE ÁUDIOS VERDADEIROS E FALSOS

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

S239e Santos, Vitor Hugo Muniz de Sousa.
ESTRATÉGIAS DE AMPLIAÇÃO E BALANCEAMENTO DE DADOS PARA A DETECÇÃO DE ÁUDIOS VERDADEIROS E FALSOS / Vitor Hugo Muniz de Sousa Santos. – 2025.
52 f. : il.

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

1. Detecção de manipulação de áudio. 2. Voz sintética. 3. Ampliação de dados. 4. Balanceamento de classes. 5. Classificação supervisionada. I. Título.

CDD 621.39

VITOR HUGO MUNIZ DE SOUSA SANTOS

ESTRATÉGIAS DE AMPLIAÇÃO E BALANCEAMENTO DE DADOS PARA A DETECÇÃO DE ÁUDIOS VERDADEIROS E FALSOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Computação do Centro de Sobral da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de 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. Ialis de Paula Cavalcanti Júnior
Universidade Federal do Ceará (UFC)

Eng. Me. Lucas Pedrosa Valente
Chipus Microeletrônica

À minha família, pelo amor, apoio e confiança em cada etapa desta jornada, e a todos que, de alguma forma, contribuíram para a realização deste sonho.

AGRADECIMENTOS

A conclusão deste Trabalho de Conclusão de Curso marca o fim de uma jornada repleta de aprendizados, desafios e crescimento pessoal. Para que esse momento fosse possível, contei com o apoio de pessoas fundamentais, às quais deixo aqui minha profunda gratidão.

À minha família, pelo amor, apoio constante e ensinamentos, especialmente aos meus pais. À minha namorada, por estar comigo nos momentos mais difíceis e felizes dessa caminhada.

Ao meu orientador, Prof. Dr. Marcelo Marques Simões de Souza, pelas contribuições técnicas e críticas que elevaram a qualidade deste trabalho.

Aos professores do curso de Engenharia de Computação, pelo conhecimento transmitido. Aos colegas e amigos da Engenharia de Computação, por compartilharem essa jornada comigo.

A todos que, direta ou indiretamente, fizeram parte dessa história, meu muito obrigado.

“A mente que se abre a uma nova ideia jamais voltará ao seu tamanho original.”

(Albert Einstein)

RESUMO

Este trabalho investiga o impacto de estratégias de ampliação e balanceamento de dados na melhoria do desempenho de modelos de classificação supervisionada na detecção de áudios de fala manipulados por sistemas de conversão de texto em fala (TTS). Utilizou-se a base de dados Fake or Real (FoR), desenvolvida pela Lassonde School of Engineering, composta por amostras de fala autênticas e sintéticas, geradas por diferentes arquiteturas de conversão de texto em fala. Foram aplicadas cinco técnicas de ampliação de dados: adição de ruído, mudança de velocidade, mudança de tom, mascaramento tempo-frequência e equalização espectral aleatória. Para mitigar o desbalanceamento entre classes, foram utilizadas as técnicas de sobreamostragem de minoria sintética (SMOTE) e subamostragem aleatória. Os sinais de áudio foram representados por coeficientes cepstrais na escala de frequência Mel (MFCCs), e esses últimos foram utilizados como entrada para os classificadores floresta aleatória (RF), máquina de aumento de gradiente leve (LGBM), Naive Bayes, rede neural de longo e curto prazo (LSTM) e algoritmo k-vizinhos próximos (KNN). A combinação da ampliação de dados com o balanceamento via SMOTE proporcionou o melhor desempenho geral. Nesse cenário, o modelo Floresta Aleatória (RF) destacou-se, alcançando acurácia e F1-score de 95 %.

Palavras-chave: Detecção de manipulação de áudio. Voz sintética. Ampliação de dados. Balanceamento de classes. Classificação supervisionada.

ABSTRACT

This work investigates the impact of data augmentation and class balancing strategies on improving the performance of supervised classification models in the detection of speech audio manipulated by text-to-speech (TTS) systems. The Fake or Real (FoR) dataset, developed by the Lassonde School of Engineering, was used. It consists of authentic and synthetic speech samples generated by different TTS architectures. Five data augmentation techniques were applied: noise addition, speed variation, pitch shifting, time-frequency masking, and random spectral equalization. To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) and random undersampling were used. The audio signals were represented by Mel-Frequency Cepstral Coefficients (MFCCs), which served as input to the following classifiers: Random Forest (RF), Light Gradient Boosting Machine (LGBM), Naive Bayes, Long Short-Term Memory (LSTM) neural network, and k-Nearest Neighbors (KNN). The combination of data augmentation with balancing via SMOTE provided the best overall performance. In this scenario, the Random Forest (RF) model stood out, achieving accuracy and F1-score of 95 %.

Keywords: Audio manipulation detection. Synthetic voice. Data augmentation. Class balancing. Supervised classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DCT	Transformada Discreta do Cosseno (do inglês, *Discrete Cosine Transform*)
ELA	Esclerose Lateral Amiotrófica
FoR	falsas ou verdadeiras (do inglês, *Fake or Real*)
KNN	K-Vizinhos Mais Próximos (do inglês, *K-Nearest Neighbors*)
LGBM	Máquina de Aumento de Gradiente Leve (do inglês, *Light Gradient Boosting Machine*)
LSTM	Redes Neurais de Memória de Longo Curto Prazo (do inglês, *Long Short-Term Memory*)
MFCCs	Coeficientes Cepstrais de Frequência Mel (do inglês, *Mel Frequency Cepstral Coefficients*)
NLP	Processamento de Linguagem Natural (do inglês, *Natural Language Processing*)
RF	Floresta Aleatória (do inglês, *Random Forest*)
SMOTE	Técnica de Sobreamostragem de Minoria Sintética (do inglês, *Synthetic Minority Over-sampling Technique*)
TTS	conversão de texto para fala (do inglês, *Text-to-Speech*)

LISTA DE SÍMBOLOS

α	Fator de intensidade do ruído adicionado
β	Fator de reamostragem para mudança de tom
Δf	Largura da faixa mascarada na frequência
Δt	Duração da região mascarada no tempo
$F\{s(t)\}$	Transformada de Fourier do sinal
F^{-1}	Transformada Inversa de Fourier
$G(t,f)$	Máscara de ganho aleatória no domínio tempo-frequência
$H(f)$	Função de equalização aleatória no domínio da frequência
$H_m(k)$	Filtro triangular na escala Mel
$ISTFT\{\cdot\}$	Transformada Inversa de Fourier de Curto Prazo
j	Unidade imaginária, tal que $j^2 = -1$
k	Índice da frequência discreta
M	Número total de filtros Mel
m	Índice do filtro Mel
n_p	Número de semitons de variação de tom
n	Índice do coeficiente cepstral desejado
N	Número de amostras no frame
r	Vetor de ruído
s	Sinal de áudio original no domínio do tempo
s'	Sinal de áudio modificado após alguma técnica de ampliação
$s(t)$	Representação do sinal de áudio em função do tempo
$S(f,t)$	Espectrograma do sinal original
$S'(f,t)$	Espectrograma após mascaramento no tempo
$S''(f,t)$	Espectrograma após mascaramento no tempo e na frequência
$STFT\{s(t)\}$	Transformada de Fourier de Curto Prazo do sinal
t	Tempo contínuo (variável temporal)

t_0	Ponto inicial do mascaramento no tempo
f	Frequência (Hz)
f_0	Ponto inicial do mascaramento na frequência
$x(n)$	Frame do sinal de áudio no tempo
$X(k)$	Espectro de frequência do frame
$\log S_m$	Saída logarítmica do filtro m
$MFCC_n$	Coeficiente cepstral de frequência Mel

SUMÁRIO

1 INTRODUÇÃO

A comunicação sempre foi um elemento fundamental na organização e desenvolvimento das sociedades humanas. Desde os primeiros registros históricos, tem sido uma ferramenta essencial para expressar ideias, relatar eventos e promover conexões entre as pessoas.

O advento da internet potencializou a produção e disseminação de informações em diversos formatos midiáticos, sendo a exposição à desinformação online uma das principais ameaças à integridade da sociedade atual. Isso exige uma constante revisão dos mecanismos de verificação e validação de informações (TANDOC *et al.*, 2018).

O avanço das tecnologias de manipulação de áudio, especialmente aquelas baseadas em redes neurais profundas, tornou possível a sintetização de vozes humanas com elevado grau de realismo. Esse fato levanta questionamentos importantes quanto à confiabilidade de conteúdos sonoros e ao seu uso malicioso, o que compromete a noção de verdade nos meios digitais (CHESNEY; CITRON, 2019).

Uma abordagem que contribui para mitigar o problema em questão é o desenvolvimento de modelos de aprendizado de máquina para identificação de áudios manipulados e autênticos. No entanto, desenvolver tais modelos requer bases de dados com diversidade de amostras de vozes, em termos de idiomas, paisagens sonoras, sotaques, técnicas de manipulação, ruído, etc. Bases sem essas características restringem a capacidade de generalização do modelo e sua eficiência em cenários reais (MÜLLER *et al.*, 2021).

Esse estudo concentra-se na avaliação de técnicas específicas de ampliação de dados (HE; GARCIA, 2008), aplicáveis ao problema em questão, como a adição de ruído (do inglês, *Noise Injection*), mudança de velocidade (do inglês, *Speed Perturbation*), mudança de tom (do inglês, *Pitch Shifting*), mascaramento tempo-frequência (do inglês, *Time-Frequency Masking*) e equalização aleatória (do inglês, *Random Equalization*).

Adicionalmente, são exploradas técnicas de balanceamento de dados, como abordagem para mitigação do desequilíbrio entre as classes, para que se possa treinar modelos de aprendizado de máquina sem viéses. Assim, foram exploradas as Técnica de Sobreamostragem de Minoria Sintética (do inglês, *Synthetic Minority Over-sampling Technique*) (SMOTE) (CHAWLA *et al.*, 2002) e subamostragem aleatória (do inglês, *Random Undersampling*) (RATNASARI, 2024).

Os métodos de aumento de dados e de balanceamento de classes supracitadas foram avaliados em experimentos de validação cruzada associados com os seguintes modelos de

classificação: Floresta Aleatória (do inglês, *Random Forest*) (RF), Máquina de Aumento de Gradiente Leve (do inglês, *Light Gradient Boosting Machine*) (LGBM), Bayes Ingênuo (do inglês, *Naïve Bayes*), Redes Neurais de Memória de Longo Curto Prazo (do inglês, *Long Short-Term Memory*) (LSTM) e K-Vizinhos Mais Próximos (do inglês, *K-Nearest Neighbors*) (KNN). Esses algoritmos clássicos são amplamente reconhecidos por sua eficiência em tarefas de classificação (WAINER, 2016).

Ressalta-se, no entanto, que alguns desses modelos não são especificamente otimizados para tarefas com dados temporais ou sequenciais, como é o caso do áudio, podendo, portanto, apresentar métricas de desempenho inferiores. Ainda assim, sua inclusão nesta pesquisa se justifica pela intenção de contemplar diferentes abordagens, como as baseadas em árvores, estatísticas, redes neurais e métodos de instância, de modo a obter uma análise comparativa mais abrangente e representar a diversidade de estratégias comumente utilizadas em problemas de classificação. As métricas adotadas para fins de comparação e avaliação de desempenho dos modelos são a acurácia, precisão, revocação e *F1-score*.

As amostras utilizadas foram extraídas da base de dados falsas ou verdadeiras (do inglês, *Fake or Real*) (FoR) da *Lassonde School of Engineering* (BIOLOGIACALLY INSPIRED LEARNING LAB, 2021), especificamente sua versão original. Essa base contém gravações de falas reais e sintetizadas, balanceadas em termos de gênero e classe, incluindo falas geradas por sistemas de conversão de texto para fala (do inglês, *Text-to-Speech*) (TTS), *Deep Voice 3* e *Google Wavenet*, além de falas reais oriundas de fontes como o *Arctic Dataset* (CARNEGIE MELLON UNIVERSITY, 2004), *LJSpeech Dataset* (ITO, 2017) e *VoxForge Dataset* (VOXFORGE, 2020). Esse conjunto diversificado de dados oferece uma base robusta para o treinamento e avaliação de modelos voltados à detecção de manipulação de áudio.

1.1 Objetivos

O objetivo geral é investigar como metodologias de ampliação de dados e balanceamento de dados influenciam o desempenho de modelos de classificação supervisionada, na tarefa de identificação de manipulação de áudio de vozes. Como objetivos específicos, destacamos:

- Investigar como diferentes técnicas de ampliação de dados: adição de ruído, mudança de velocidade, mudança de tom, mascaramento tempo-frequência e equalização aleatória afetam o desempenho de modelos de classificação supervisionada, utilizando validação cruzada com a base de dados *FoR*.

- Analisar o impacto das estratégias de balanceamento de classes, especificamente o uso de *SMOTE* e da subamostragem aleatória, sobre o desempenho dos modelos de aprendizado de máquina considerados.
- Comparar o desempenho de diferentes modelos de classificação supervisionada: *Random Forest (RF)*, *LightGBM (LGBM)*, *Naive Bayes*, *Long Short-Term Memory (LSTM)* e *K-Nearest Neighbors (KNN)* diante de diferentes cenários de dados (original, ampliado e balanceado).

1.2 Organização do Trabalho

Este trabalho está estruturado em sete capítulos:

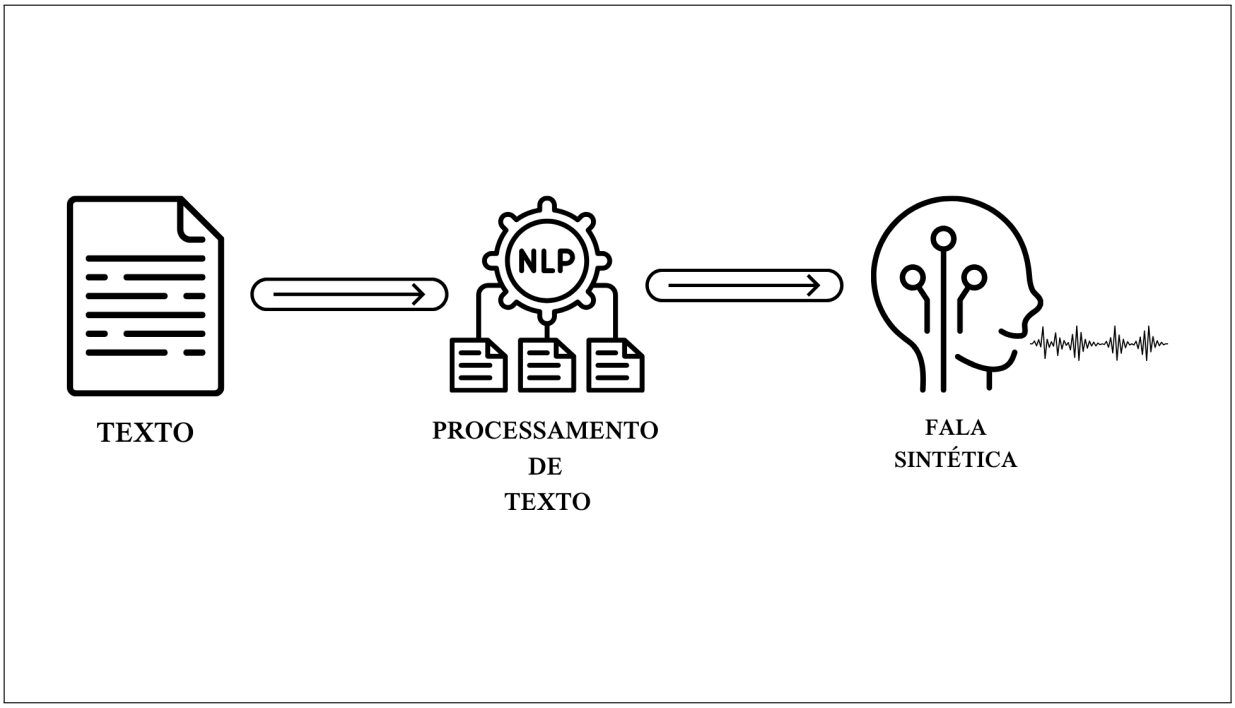
- Capítulo 1 – Contextualiza o problema, a motivação para o estudo, o objetivo geral da pesquisa e a organização do trabalho.
- Capítulo 2 – Discorre sobre o impacto da Manipulação de Áudio na sociedade trazendo o panorama atual da manipulação de áudios digitais, suas implicações e os principais desafios associados à detecção de áudios falsificados.
- Capítulo 3 – Detalha as técnicas de ampliação de dados aplicadas para aumentar a diversidade do conjunto de treinamento.
- Capítulo 4 – Apresenta as estratégias de balanceamento de classes utilizadas para corrigir o desequilíbrio entre as amostras de áudios verdadeiros e falsos.
- Capítulo 5 – Descreve as etapas metodológicas adotadas, incluindo a coleta e pré-processamento dos modelos de aprendizado de máquina utilizados.
- Capítulo 6 – Apresenta os resultados obtidos com as técnicas de aumento de dados e balanceamento aplicadas, detalhando o desempenho dos modelos em diferentes cenários.
- Capítulo 7 – Resume e conclui os principais achados do estudo, discute suas limitações e propõe possíveis direções para trabalhos futuros.

2 MANIPULAÇÃO DE ÁUDIO

O contexto atual da inteligência artificial democratizou a manipulação de voz, anteriormente restrita a especialistas, tornando-a acessível aos usuários leigos. Porém, tais facilidades têm promovido a proliferação crescente e preocupante de conteúdos manipulados por voz (ZHANG *et al.*, 2021). Ao contrário das edições em imagens ou vídeos, que frequentemente deixam vestígios, as alterações em arquivos de áudio passam despercebidas pelo ouvido humano, tornando o desafio de verificar sua autenticidade ainda mais complexo (DAS *et al.*, 2020).

Ferramentas baseadas em tecnologias de TTS, como ilustrado na Figura 1, permitem a conversão de texto em fala com naturalidade, o que pode ser explorado de maneira maliciosa. O processo inicia-se com a entrada textual, seguida por uma etapa de processamento linguístico para ajustar entonação, pausas e ritmo, além de extrair características relevantes que alimentam modelos responsáveis pela geração de voz realista, utilizando para isso técnicas de Processamento de Linguagem Natural (do inglês, *Natural Language Processing*) (NLP). Por fim, o sistema gera a fala sintética, que pode ser disponibilizada em diferentes mídias de saída, como arquivos de áudio, espectrogramas ou vídeos.

Figura 1 – Processo de TTS



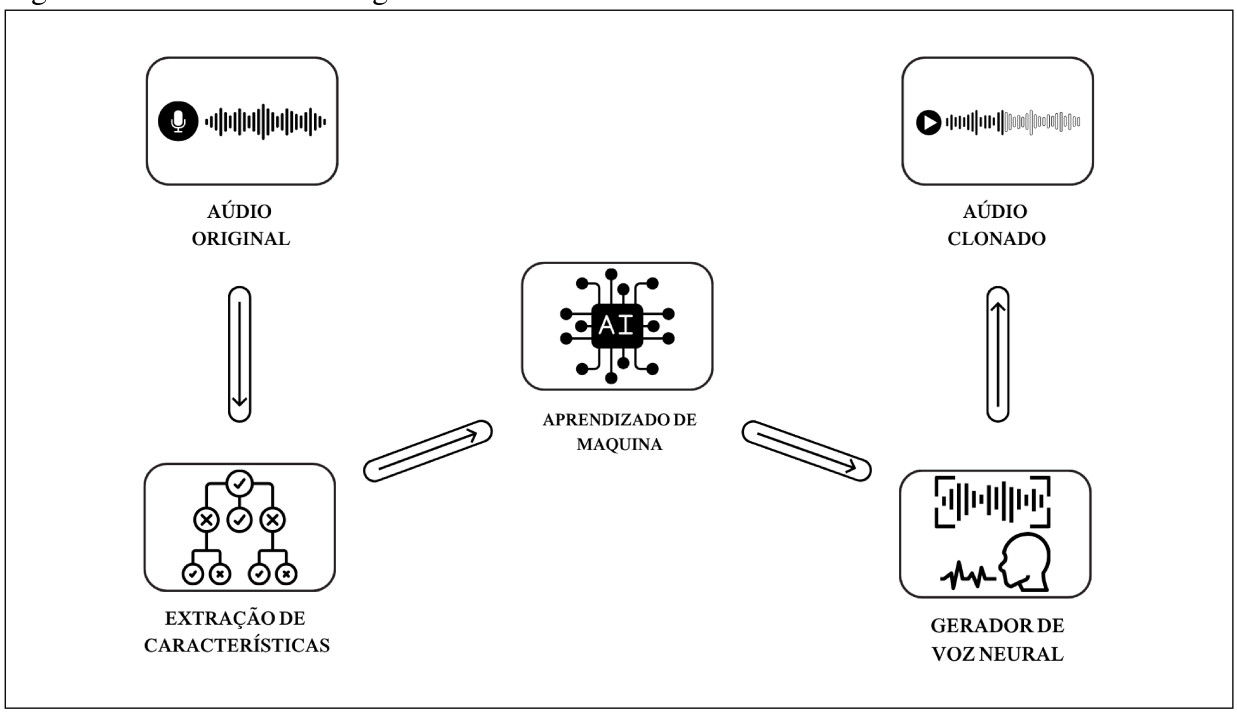
Fonte: Autor

Um caso emblemático ocorreu em 2020, quando uma empresa sediada nos Emirados Árabes Unidos foi alvo de um golpe sofisticado. Criminosos utilizaram clonagem de voz

para imitar o diretor da organização. Enganado pela verossimilhança da fala, um gerente bancário autorizou a transferência de US$35 milhões para contas controladas pelos fraudadores. Esse episódio ilustra de maneira contundente o impacto real e potencialmente devastador da manipulação de voz mediada por inteligência artificial (KUNDU, 2021).

A clonagem de voz é a reprodução artificial da voz de uma pessoa utilizando técnicas de inteligência artificial. O processo envolve três etapas principais: extração, modelagem e síntese. Inicialmente, características linguísticas e acústicas são extraídas da fala original, representando os traços únicos da voz. Essas informações alimentam um modelo de aprendizado de máquina, que aprende os padrões específicos da fala da fonte original. Por fim, um gerador de voz neural sintetiza uma nova fala com propriedades vocais semelhantes, porém com conteúdo distinto. A clonagem pode ser executada por meio de abordagens baseadas em *TTS*, em que o conteúdo da fala é gerado a partir de texto, ou por técnicas de *Voice Conversion*, que transformam diretamente a voz de um locutor na de outro, mantendo o conteúdo original. A ilustração desse processo está apresentada na Figura 2.

Figura 2 – Processo de clonagem de voz



Fonte: Autor

Diante desse cenário, a detecção eficaz de áudios manipulados torna-se imperativo para prevenir fraudes financeiras, golpes políticos, entre outros. Vale destacar que esses são apenas exemplos das diversas ameaças existentes. Sem métodos robustos de verificação, transações,

operações bancárias e até mesmo interações institucionais tornam-se altamente vulneráveis a ataques cibernéticos.

Nesse contexto, estudos que explorem estratégias baseadas em pré-processamento de sinais, extração de características e técnicas de ampliação de dados para o treinamento de modelos de inteligência artificial podem aprimorar significativamente o desempenho de sistemas automáticos de detecção, aumentando a segurança e garantindo a integridade das transações (MICHELSANTI *et al.*, 2021).

Apesar dos riscos inerentes e da urgência por mecanismos de defesa, é fundamental reconhecer que a tecnologia de manipulação de áudio, quando utilizada eticamente, representa uma força transformadora com uma vasta gama de aplicações legítimas e benéficas. Longe de se restringir a usos controversos, inovações como a clonagem vocal estão redefinindo paradigmas em áreas essenciais como saúde, educação, entretenimento e segurança, promovendo melhorias substanciais na comunicação, otimizando o aprendizado e fortalecendo análises investigativas.

No setor de segurança pública, investigação criminal e perícia forense, a manipulação de áudio tem sua aplicação na detecção de áudios gerados artificialmente. Ferramentas baseadas em inteligência artificial têm-se mostrado essenciais para a análise forense de gravações, permitindo verificar a autenticidade do material sonoro e identificar se um conteúdo foi manipulado ou produzido por sistemas de clonagem de voz (CASEY, 2009).

Órgãos como o FBI têm investido em soluções de análise forense de áudio para combater crimes digitais, extorsões, fraudes e campanhas de desinformação, reconhecendo o crescente potencial nocivo das tecnologias de manipulação de voz e *deepfakes* (FBI, 2023). Além disso, sistemas de monitoramento modernos vêm incorporando detectores de voz falsa com o objetivo de evitar o uso de áudios manipulados em operações ilegais ou em processos judiciais. Esses avanços demonstram que, quando aplicada de forma ética, a tecnologia de clonagem e detecção de voz pode fortalecer o trabalho investigativo, proteger vítimas e garantir maior confiança nas comunicações e evidências oficiais (KHANJANI *et al.*, 2021).

Na área da saúde e da comunicação assistiva, por exemplo, a clonagem personalizada de voz representa um avanço significativo, especialmente para pessoas com deficiências na fala causadas por doenças degenerativas. Tecnologias como o *Project Euphonia*, do *Google*, e o *Voice Banking*, da *Acapela Group*, possibilitam que os usuários gravem e preservem suas próprias vozes antes da perda da fala.

O *Project Euphonia*, por exemplo, auxiliou Steve Saling, diagnosticado com Escle-

rose Lateral Amiotrófica (ELA), a interagir com dispositivos domésticos inteligentes e participar de atividades cotidianas, como torcer em jogos esportivos, por meio de sons não verbais e gestos faciais. Esses sistemas armazenam os padrões vocais do indivíduo, permitindo que sua identidade vocal seja mantida em dispositivos de comunicação assistiva. Assim, mesmo após a perda da capacidade de falar, é possível manter uma comunicação autêntica, promovendo dignidade, autonomia e continuidade na interação social (CATTIAU, 2021).

Em suma, os avanços na manipulação de áudio por inteligência artificial têm impulsionado aplicações em múltiplos setores. Técnicas como clonagem vocal, modulação de entonação e filtragem de ruídos já demonstram impactos positivos em diferentes contextos. Por outro lado, o uso malicioso dessas tecnologias representa um risco crescente, o que exige medidas preventivas eficazes. O desafio está em impulsionar o progresso tecnológico sem comprometer a segurança e a ética. Para isso, é essencial investir em métodos de detecção mais sofisticados e estabelecer diretrizes que assegurem um uso responsável.

3 AMPLIAÇÃO DOS DADOS

A ampliação de dados, ou data augmentation, consiste em manipular as amostras de dados disponíveis para geração de novas amostras, incrementando a quantidade e a variedade dos dados para o treinamento de modelos de classificação de áudio de falas. Essa abordagem é utilizada em cenários onde os conjuntos de dados originais são limitados em tamanho ou diversidade, o que pode comprometer a capacidade de generalização dos modelos. Ao empregar tais técnicas, busca-se mitigar essas limitações, fomentando o desenvolvimento de modelos capazes de generalizar adequadamente os dados, mostrando capacidade preditiva em cenários reais (BIANCO *et al.*, 2019).

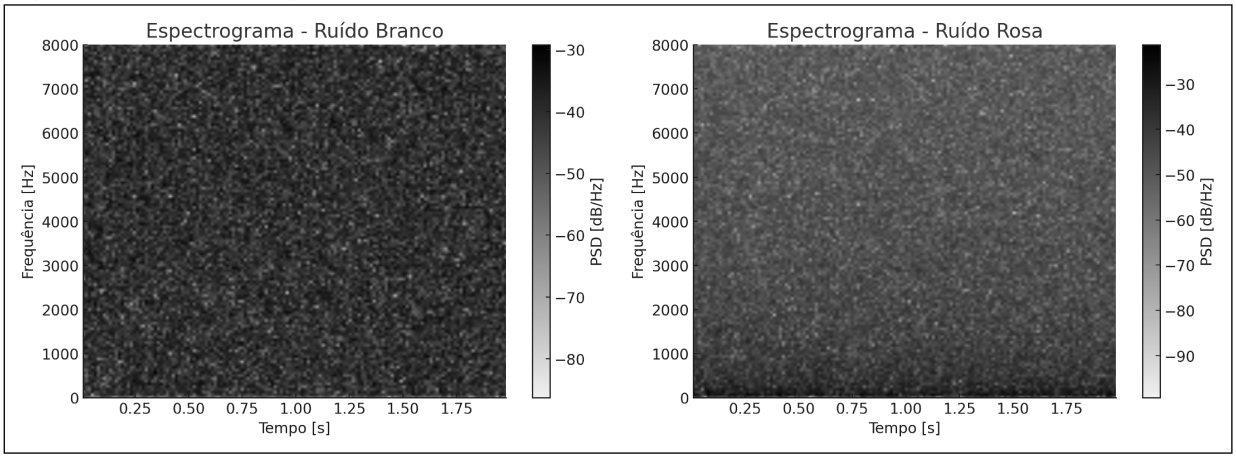
Além de enriquecer os conjuntos de dados, a ampliação de dados melhora o desempenho de modelos em cenários complexos e variados (PARK *et al.*, 2019). No contexto de áudios de vozes, foram selecionadas cinco técnicas distintas de ampliação de dados: adição de ruído, mudança de velocidade, mudança de tom, mascaramento tempo-frequência e equalização aleatória. Cada técnica simula diferentes tipos de variações presentes em gravações de áudio de falas, a partir das quais buscou-se avaliar como cada uma delas impacta na capacidade do modelo de generalizar. As subseções seguintes detalham cada uma dessas técnicas.

3.1 Adição de Ruído

A adição de ruído aos sinais de áudio foi implementada buscando simular a presença de ruído ambiente ou interferências eletrônicas em gravações de áudios de falas, refletindo condições não ideais de captura e variações naturais dos canais de áudio (SALAMON; BELLO, 2017).

Para simular diferentes cenários de gravação, foram utilizados dois tipos de ruído: ruído rosa nos áudios reais e ruído branco nos áudios falsos. O ruído branco é caracterizado por uma distribuição espectral uniforme em todas as frequências, enquanto o ruído rosa é caracterizado por apresentar maior energia nas frequências baixas e decresce conforme aumenta a frequência. A Figura 3 ilustra essa diferença por meio dos espectrogramas Mel dos ruídos branco e rosa.

Figura 3 – Análise Espectral de Ruído Branco e Ruído Rosa



Fonte: Autor

A escolha por aplicar ruído branco aos áudios falsos e ruído rosa aos reais baseou-se nas características específicas desses ruídos e na intenção de criar cenários mais realistas. A adição de ruído branco impede que o modelo aprenda a associar a ausência de ruído a uma determinada classe, enquanto o ruído rosa desafia o modelo a reconhecer os padrões de voz autênticos mesmo sob condições acústicas adversas. Essa estratégia contribui para tornar o modelo menos sensível a variações aleatórias no sinal de entrada (KO *et al.*, 2015). A implementação da adição de ruído é descrita na Equação 3.1:

$$s' = s + \alpha \cdot r \tag{3.1}$$

em que:

- s é o sinal de áudio original;
- r é o vetor de ruído;
- α define a intensidade do ruído;
- s' é o sinal de áudio com ruído.

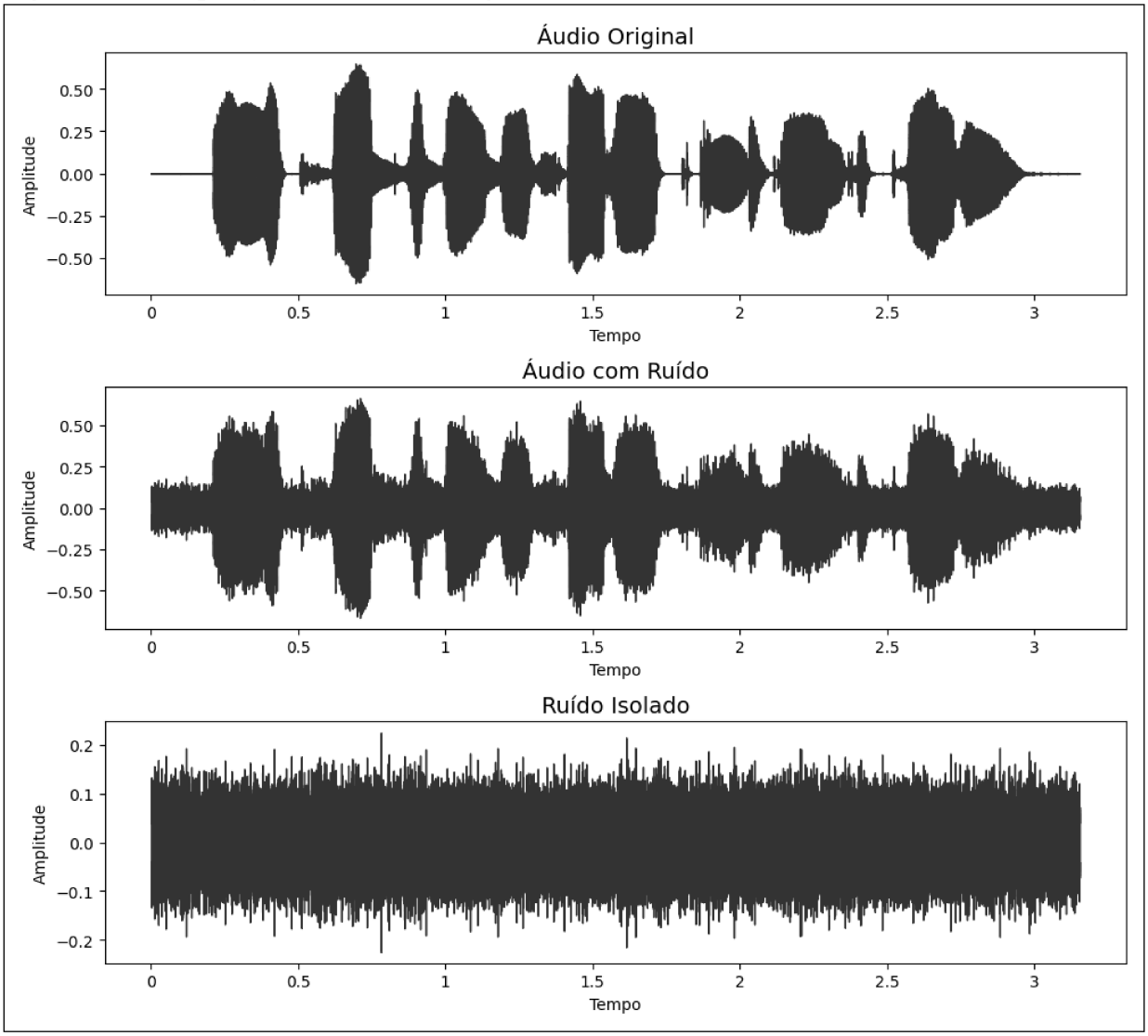
Essa equação descreve uma combinação linear entre o sinal original s e o vetor de ruído r, resultando no novo sinal com ruído adicionado s'. O fator α atua como um controlador da intensidade do ruído: se α for muito grande, o ruído pode dominar o sinal original, tornando o áudio uninteligível; por outro lado, se α for muito pequeno, o ruído pode se tornar imperceptível, reduzindo a eficácia da técnica.

Após a adição do ruído, o sinal resultante s' pode ultrapassar os limites dinâmicos aceitáveis para um sinal de áudio, causando distorções. Logo, o sinal é normalizado para garantir que seus valores permaneçam dentro de um intervalo seguro, tipicamente entre -1,0 e 1,0.

A implementação envolveu a utilização de uma função dedicada para sobrepor ruído ao sinal de áudio original, com um fator de intensidade de ruído (α) padrão estabelecido em 0,05. A escolha deste valor foi baseada em uma série de testes preliminares que consideraram tanto a percepção auditiva quanto o impacto no desempenho do modelo. O objetivo foi encontrar um equilíbrio onde o ruído introduzisse uma variação sonora perceptível sem comprometer a inteligibilidade do conteúdo da fala e mascarar as características fonéticas essenciais para a classificação.

A Figura 4 apresenta a comparação entre o sinal original e o sinal com ruído adicionado, permitindo visualizar o impacto da perturbação no domínio do tempo. Na representação de amplitude e tempo, o sinal original mostra suas variações de forma limpa. Já o sinal com ruído apresenta flutuações adicionais, mais irregulares e de menor amplitude, sobrepostas à forma de onda original. Essa comparação mostra como o ruído altera a forma do sinal, exigindo que o modelo classificador seja capaz de identificar a informação relevante mesmo em meio à interferência.

Figura 4 – Comparação entre o sinal original e o sinal com ruído



Fonte: Autor

3.2 Mudança de Velocidade

Uma estratégia complementar de aumento de dados é alterar a velocidade dos áudios, com o intuito de simular variações naturais no ritmo e na cadência da fala humana. Essa abordagem expõe o modelo de classificação a um espectro mais amplo de velocidades de fala, tornando-o menos dependente de um padrão específico. Com isso, o modelo aprende a focar em características mais intrínsecas da voz e do conteúdo linguístico, independentemente do ritmo da fala (ALEX *et al.*, 2023).

A dilatação temporal, na detecção de áudios manipulados, parte do pressuposto de que métodos de síntese vocal introduzem padrões rítmicos artificiais, como cadências excessivamente regulares ou pausas inconsistentes. Incorporar amostras desaceleradas durante

o treinamento permite que o modelo aprenda a identificar essas discrepâncias como possíveis indícios de falsificação (KORSHUNOV; MARCEL, 2018). A implementação da mudança de velocidade pode ser descrita matematicamente pela seguinte equação 3.2:

$$s'(t) = s(\gamma \cdot t) \tag{3.2}$$

em que:

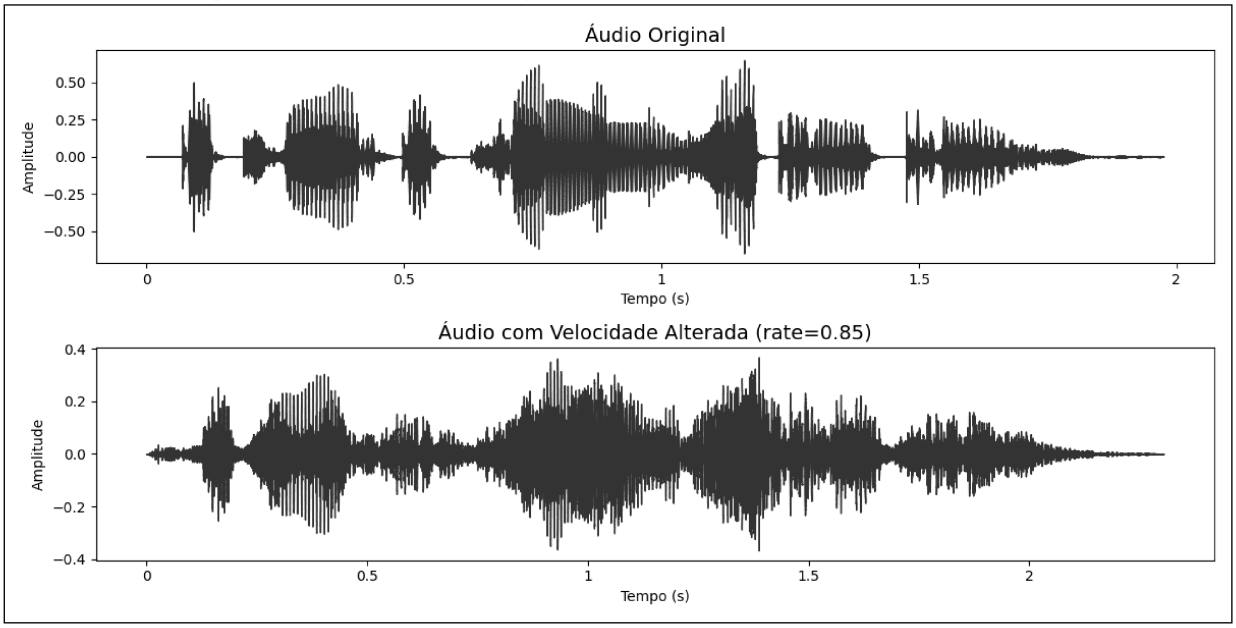
- $s(t)$ é o sinal original no tempo;
- $s'(t)$ é o sinal com velocidade modificada;
- γ define o quanto a velocidade do áudio será alterada (neste caso, $\gamma = 0{,}85$).

Essa equação representa uma transformação temporal do sinal de áudio. O sinal $s'(t)$ é obtido ao modificar os valores do sinal original $s(t)$ em uma escala de tempo ajustada pelo fator γ. O fator γ controla a velocidade de reprodução: quando $\gamma < 1$, o áudio é reproduzido mais lentamente; já se $\gamma > 1$, o áudio é acelerado.

Para nossa pesquisa, a implementação da mudança de velocidade foi realizada produzindo uma ligeira desaceleração do áudio. Com base em meus testes preliminares que avaliaram a naturalidade e o impacto na inteligibilidade, adotou-se um fator de variação de velocidade de $v = 0{,}85 \cdot v_0$, representando uma redução de 15% em relação à velocidade original v_0. Isso efetivamente simula uma fala mais pausada. Importante destacar que essa modificação foi implementada através de algoritmos de *time-stretching*, que alteram a duração do áudio sem modificar significativamente seu tom, portanto, preservando as características espectrais da voz.

A Figura 5 apresenta a comparação entre um sinal de áudio original e sua versão com velocidade reduzida no domínio do tempo. Na representação de amplitude e tempo, o sinal original mostra suas variações com espaçamento regular, enquanto o sinal com velocidade reduzida apresenta um alongamento no eixo temporal, fazendo com que a forma de onda se estenda por uma duração maior, exigindo que o modelo classificador reconheça as mesmas informações mesmo para uma fala mais lenta.

Figura 5 – Comparação entre o sinal original e o sinal com velocidade reduzida em 15%



Fonte: Autor

3.3 Mudança de Tom

A mudança de tom, conhecida como *pitch shifting* é uma técnica de processamento digital de sinais de áudio que permite simular variações da altura de um som, sem alterar a velocidade ou a duração total. O tom (do inglês, *pitch*) está intrinsecamente ligado à frequência das ondas sonoras, e consiste no atributo psicoacústico que define se um som é percebido como mais agudo ou mais grave (OLIVEIRA, 2020).

Entre suas aplicações mais comuns está a correção de afinação vocais, como ocorre com o uso de auto-tune, a criação de efeitos sonoros em músicas e filmes, e a geração de vozes sintéticas com diferentes características. Além dessas finalidades, a técnica também vem sendo adotada como estratégia de aumento de dados em sistemas de aprendizado de máquina, especialmente em tarefas que envolvem sinais de áudio (CANTU, 2023).

A implementação desta técnica foi realizada por meio de uma função que aplica a variação de tom medida em semitons. Foi adotado um valor de ±2 semitons, intervalo capaz de introduzir variações audíveis no tom da voz, simulando com naturalidade a diversidade de características vocais humanas. A escolha de ±2 semitons busca simular a variabilidade vocal natural sem comprometer a qualidade perceptiva do áudio. Alterações maiores poderiam inserir artefatos que atrapalham o aprendizado do modelo ou levam à geração de dados artificialmente distorcidos. A fórmula matemática que representa essa transformação é apresentada na equação 3.3.

$$s'(t) = s(t/\beta) \quad \text{com} \quad \beta = 2^{n_p/12} \tag{3.3}$$

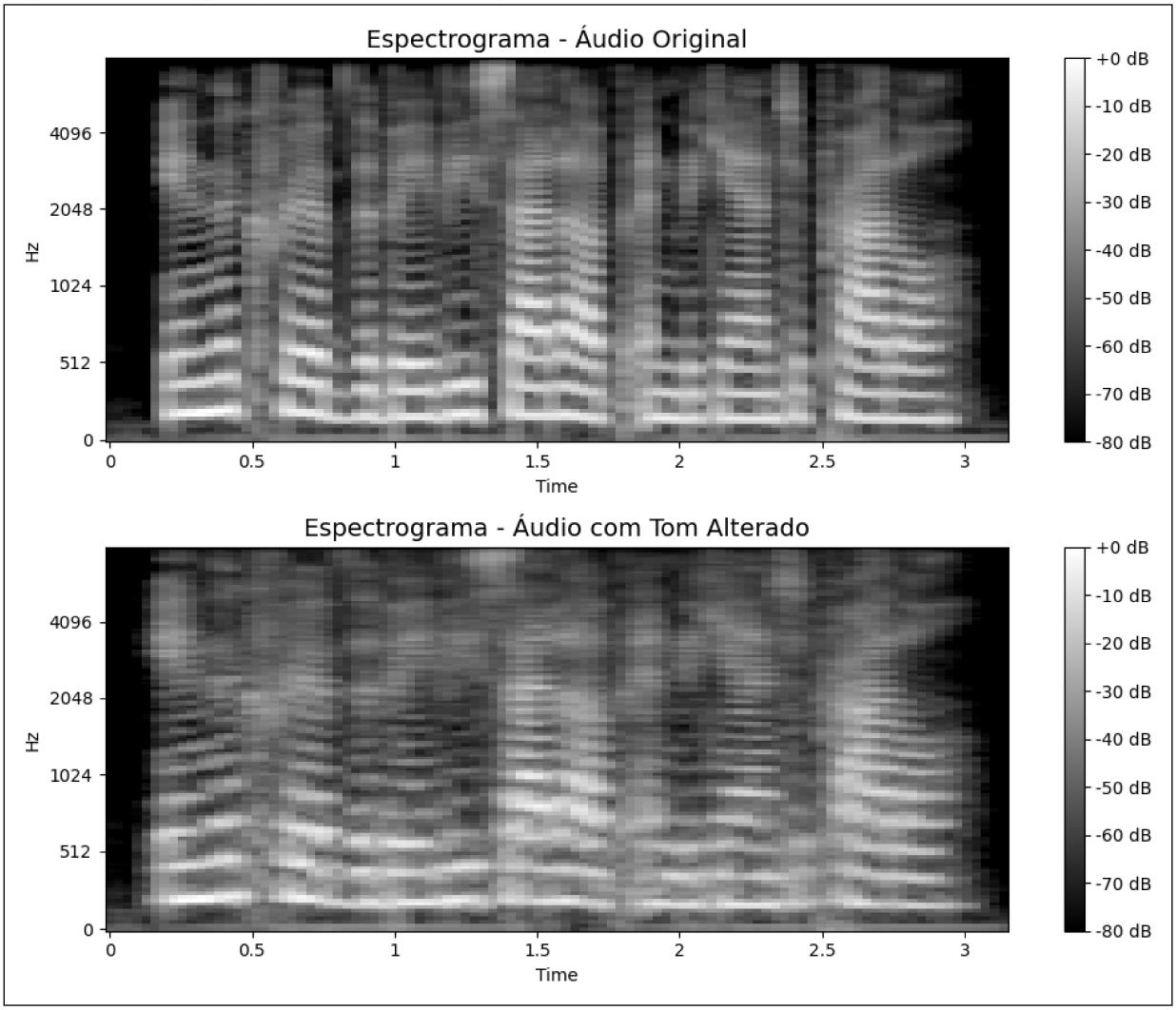
em que:

- $s(t)$ é o sinal original no tempo;
- $s'(t)$ é o sinal com tom alterado;
- n_p é o número de semitons de deslocamento;
- β é o fator de modificação correspondente.

Essa equação representa uma transformação na escala de tempo com o objetivo de alterar a frequência do sinal. É importante destacar que essa operação é realizada sem alterar significativamente a duração perceptível do áudio. O sinal $s'(t)$ é obtido ao modificar os valores do sinal original $s(t)$ em uma escala de tempo comprimida ou expandida, dependendo do fator β. Quando $\beta > 1$, o sinal é comprimido no tempo, resultando em um aumento de tom; já se $\beta < 1$, o sinal é expandido no tempo, resultando em uma diminuição de tom.

A Figura 5 apresenta a comparação entre o sinal original e o sinal com tom alterado, permitindo visualizar o impacto dessa transformação tanto no domínio do tempo quanto no domínio da frequência, por meio dos espectrogramas Mel. No espectrograma do sinal original, as componentes harmônicas e a frequência fundamental aparecem distribuídas de forma estável ao longo do tempo. Já no sinal com tom alterado, essas componentes são deslocadas para faixas de frequência mais altas, no caso de aumento do tom, ou mais baixas, no caso de redução. Essa comparação mostra como a técnica modifica o conteúdo espectral do sinal sem alterar sua duração, exigindo que o modelo classificador reconheça padrões mesmo com mudanças na altura tonal.

Figura 6 – Comparação entre o sinal original e o sinal com tom alterado de ±2 semitons



Fonte: Autor

3.4 Mascaramento Tempo-Frequência

O mascaramento tempo-frequência é uma técnica aplicada sobre o espectrograma de sinais de áudio, na qual regiões específicas ao longo dos eixos temporal e frequencial são ocultadas por meio da substituição de seus valores por zero ou por um valor neutro (YU; LI, 2021).

Ao ser exposto, durante o treinamento, a entradas com informação acústica parcialmente degradadas, o modelo de aprendizado de máquina é forçado a desenvolver mecanismos de inferência que permitam a extração de padrões mesmo na ausência de partes do sinal, evitando assim a superdependência de características locais muito específicas (LI *et al.*, 2020). A implementação pode ser descrita pelas seguintes equações:

$$S'(f,t) = \begin{cases} 0, & \text{se } t \in [t_0, t_0 + \Delta t] \\ S(f,t), & \text{caso contrário} \end{cases} \tag{3.4}$$

em que:

- $S(f,t)$ é o espectrograma original,
- $S'(f,t)$ é o espectrograma modificado com mascaramento no tempo,
- t_0 representa o ponto inicial do mascaramento no eixo do tempo,
- Δt define a duração da região mascarada no tempo.

$$S''(f,t) = \begin{cases} 0, & \text{se } f \in [f_0, f_0 + \Delta f] \\ S'(f,t), & \text{caso contrário} \end{cases} \tag{3.5}$$

em que:

- $S'(f,t)$ é o espectrograma com mascaramento no tempo,
- $S''(f,t)$ é o espectrograma final com mascaramento no tempo e na frequência,
- f_0 representa o ponto inicial do mascaramento no eixo de frequência,
- Δf define a largura da região mascarada no eixo de frequência.

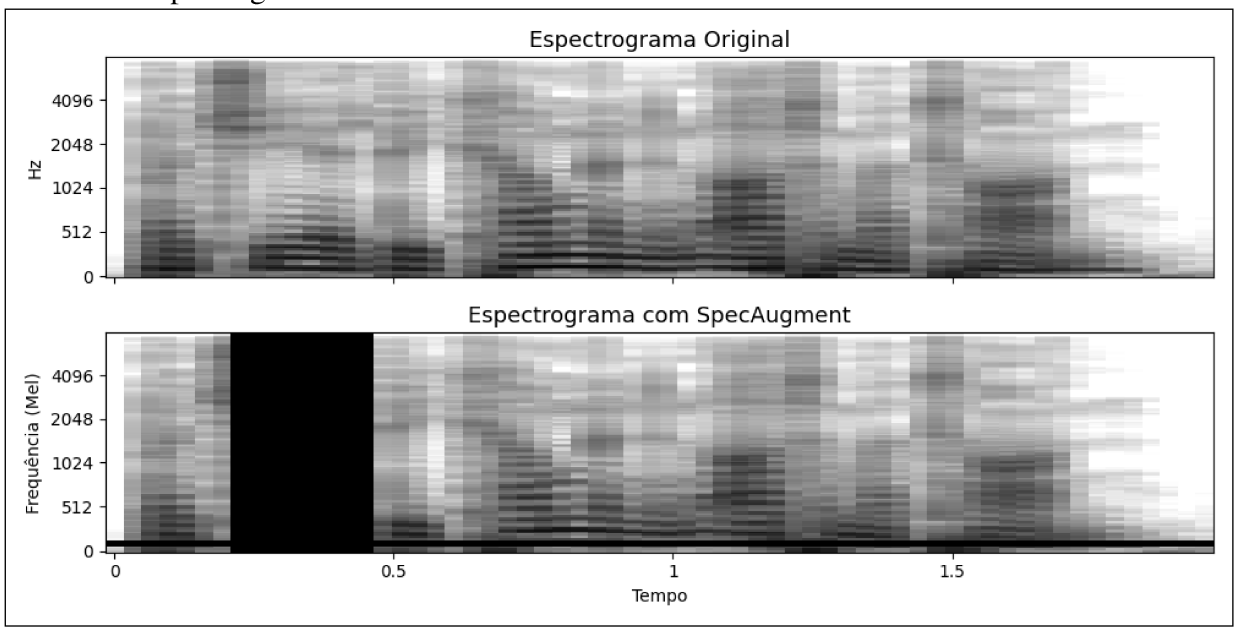
No mascaramento no tempo, o espectrograma modificado $S'(f,t)$ é obtido ao substituir os valores do espectrograma original $S(f,t)$ por zero dentro de um intervalo temporal específico, mantendo as frequências inalteradas. Os valores de $S(f,t)$ são anulados para todos os instantes t dentro do intervalo $[t_0, t_0 + \Delta t]$, onde t_0 representa o instante de início da máscara e Δt a sua duração.

Analogicamente, no mascaramento em frequência, $S''(f,t)$ é obtido ao substituir os valores em $S(f,t)$ por zero dentro de um intervalo de frequência, mantendo o eixo do tempo inalterado. Nesse caso, os valores de $S(f,t)$ são zerados para todas as frequências f no intervalo $[f_0, f_0 + \Delta f]$, em que f_0 representa o instante de início da máscara e Δf a sua largura.

A Figura 7 apresenta a comparação entre os espectrogramas mascarados pela técnica de mascaramento tempo-frequência, permitindo visualizar os efeitos de distorções simuladas no domínio tempo-frequência, por meio dos espectrogramas Mel. No espectrograma original, o conteúdo espectral é exibido de forma contínua ao longo do tempo e da frequência, enquanto que no espectrograma mascarado, observam-se regiões retangulares ocultadas, representadas por faixas escuras que indicam a anulação parcial do sinal. A faixa vertical corresponde ao

mascaramento no tempo, afetando todas as frequências em um determinado intervalo temporal, enquanto a faixa horizontal representa o mascaramento na frequência, anulando uma banda espectral ao longo de toda a duração. Essa comparação mostra como a técnica simula perdas no sinal, exigindo que o modelo classificador mantenha a capacidade de identificar padrões mesmo com informações ausentes.

Figura 7 – Comparação entre o sinal original e o sinal com o efeito do *SpecAugment* sobre o espectrograma



Fonte: Autor

3.5 Equalização Aleatória

A equalização aleatória simula a variabilidade espectral natural que ocorre nos sinais sonoros devido à diversidade de dispositivos de gravação, ambientes acústicos e canais de transmissão. Modificando aleatoriamente a intensidade de determinadas faixas de frequência do áudio, essa técnica cria versões alternativas do mesmo sinal, preservando suas características semânticas, mas alterando sua representação espectral (PEREZ-LOPEZ; SERRA, 2019).

A equalização aleatória produz tal variação espectral, aplicando filtros digitais aleatórios ao espectro de frequência do sinal. Isso gera uma função de equalização a partir de pontos de controle distribuídos ao longo do espectro audível. O resultado é um filtro com resposta espectral plausível, que, ao ser aplicado ao sinal de entrada, produz uma versão "colorida"do áudio original. A implementação pode ser representada pela seguinte equação:

$$s'(t) = ISTFT\{G(t,f) \cdot STFT\{s(t)\}\} \tag{3.6}$$

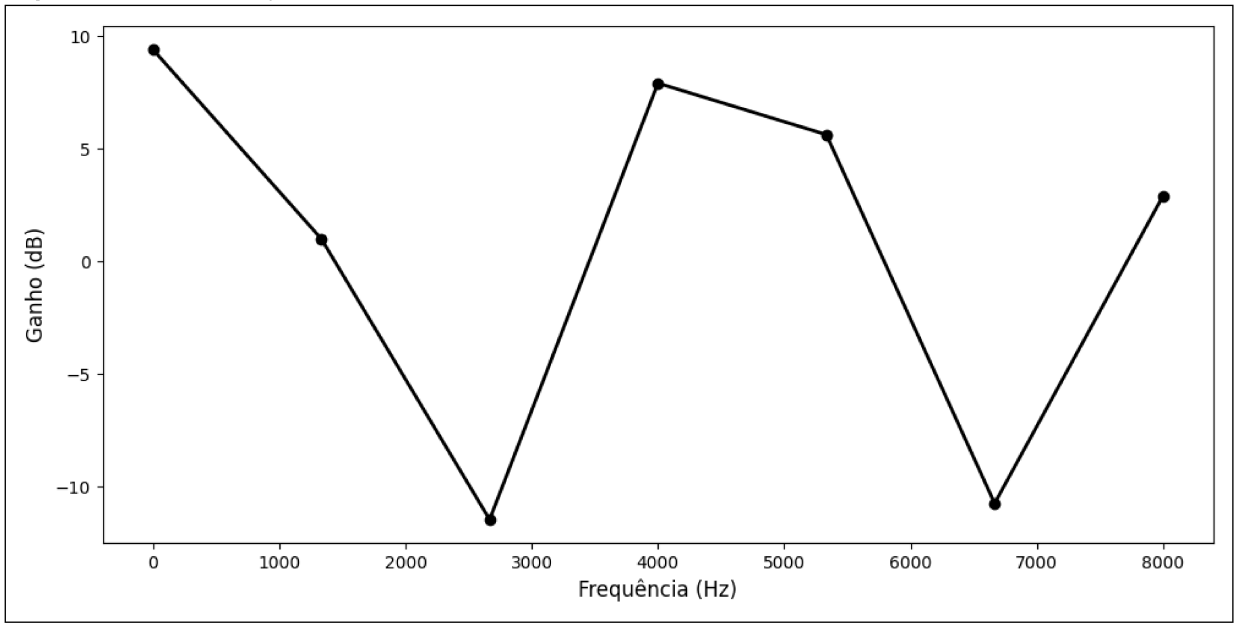
em que:

- $s(t)$ é o sinal de áudio original no domínio do tempo,
- $STFT\{s(t)\}$ representa a Transformada de Fourier de Curto Prazo do sinal,
- $G(t,f)$ é uma máscara de ganho aleatória aplicada no domínio tempo-frequência,
- $s'(t)$ é o sinal reconstruído após aplicar a Transformada Inversa $ISTFT$.

O sinal original, representado no domínio do tempo, é transformado por meio da Transformada de Fourier de Curto Prazo (STFT), que gera uma matriz complexa descrevendo a evolução das frequências ao longo do tempo. A aplicação da STFT e de sua inversa (ISTFT) envolve um equilíbrio entre resolução temporal e espectral: ao se obter maior precisão em frequência, perde-se em resolução no tempo, e vice-versa. No entanto, no contexto deste trabalho, essa perda é muito pequena e pode ser considerada desprezível.

Em seguida, aplica-se uma máscara de ganho aleatória, chamada $G(t,f)$, que contém fatores multiplicativos distribuídos aleatoriamente. Essa máscara atua sobre diferentes regiões do espectrograma, amplificando ou atenuando seletivamente componentes específicas do sinal em determinadas janelas de tempo e faixas de frequência. A distribuição dos ganhos é ilustrada na Figura 8.

Figura 8 – Distribuição da Máscara de Ganhos Aleatórios

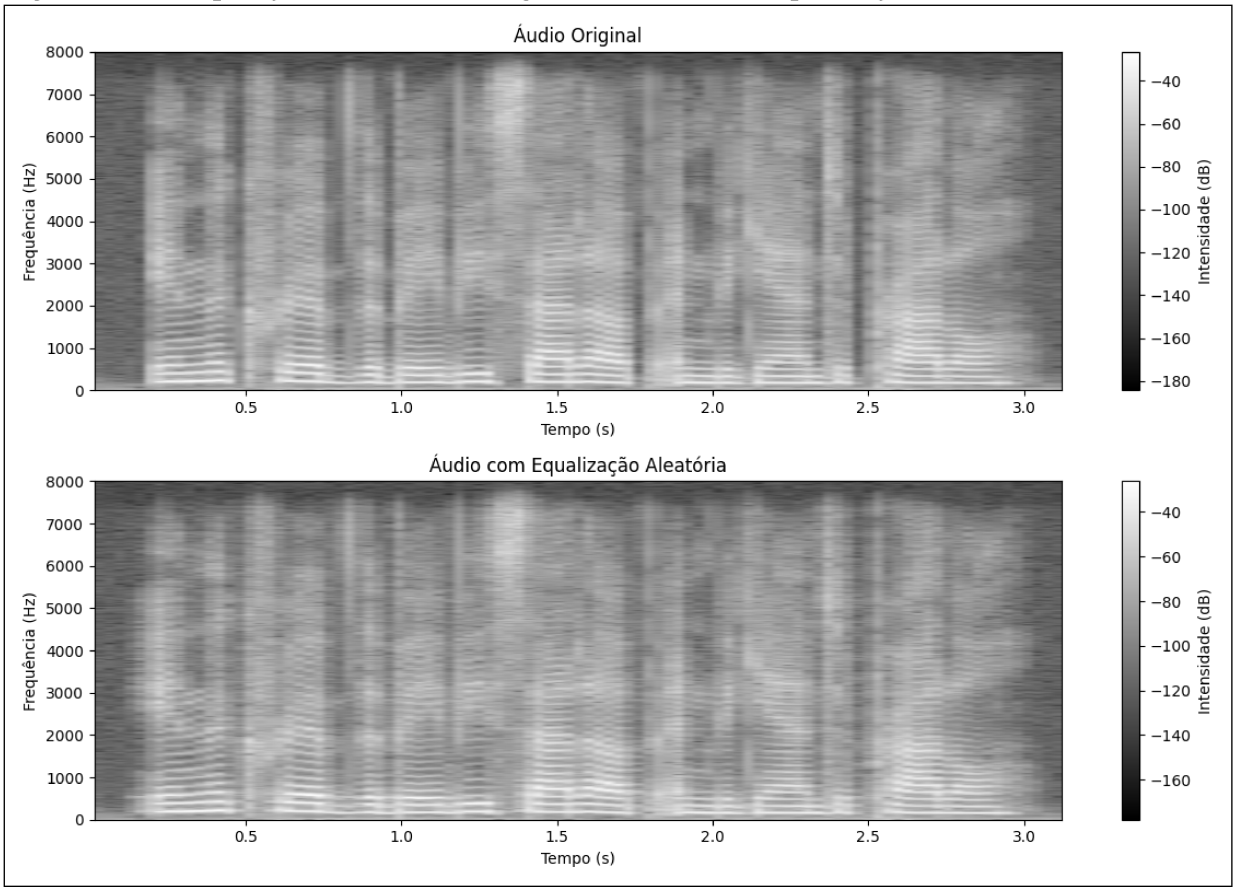


Fonte: Autor

Por fim, o sinal resultante é reconstruído no domínio do tempo por meio da Transformada Inversa (ISTFT). Essa forma de equalização aleatória simula variações reais introduzidas por diferentes condições de gravação, dispositivos ou ambientes, sendo uma estratégia eficaz de aumento de dados para melhorar a robustez de modelos de aprendizado em tarefas como detecção de manipulações ou classificação de áudios.

A Figura 9 apresenta o espectrograma de um sinal de áudio e sua versão após a aplicação de equalização aleatória, por meio dos espectrogramas Mel. Essa modificação permite visualizar o impacto dessa mudança na distribuição espectral ao longo do tempo. No espectrograma original, as faixas de frequência mantêm uma distribuição de energia característica e relativamente estável. Já no espectrograma equalizado, observa-se uma alteração nessa distribuição, com algumas regiões mais intensas e outras menos pronunciadas. Assim, a equalização aleatória modifica o balanço espectral do sinal, exigindo que o modelo classificador seja robusto a variações na ênfase de diferentes bandas de frequência.

Figura 9 – Comparação entre o sinal original e o sinal com equalização aleatória



Fonte: Autor

4 BALANCEAMENTO DOS DADOS

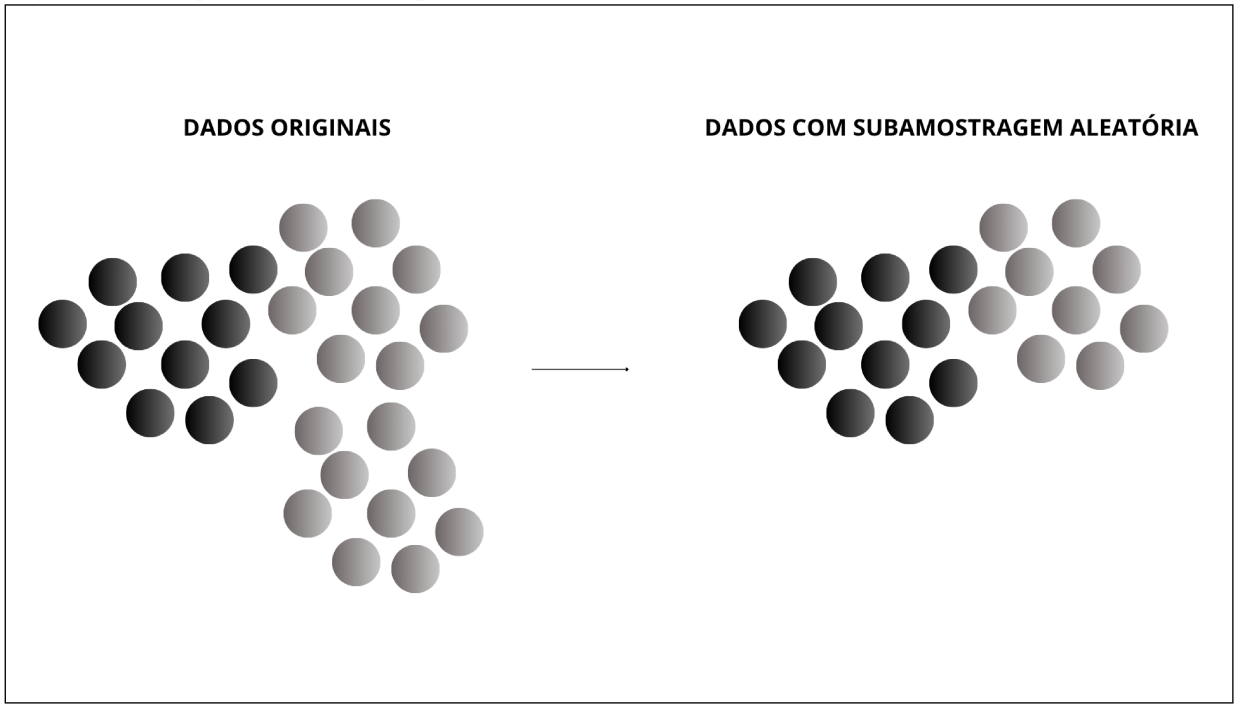
Modelos treinados com conjuntos de dados desbalanceados tendem a aprender padrões de forma enviesada, favorecendo predominantemente a classe majoritária durante o processo de classificação (BRAGA *et al.*, 2007). No âmbito desta pesquisa, há um desequilíbrio na base de dados utilizada, com uma quantidade significativamente maior de amostras rotuladas como reais, em comparação àquelas classificadas como falsas. Tal assimetria pode comprometer o desempenho do modelo, especialmente no que se refere à capacidade de identificar corretamente os áudios manipulados, resultando em menor sensibilidade à classe minoritária. Para mitigar os efeitos desse viés e promover uma distribuição mais equitativa entre as classes, empregam-se estratégias de balanceamento de dados.

Avaliamos nesse trabalho as estratégias de balanceamento por subamostragem aleatória da classe majoritária e de superamostragem da minoria com *SMOTE*. Isso permite comparar o impacto de cada técnica no desempenho do modelo e identificar qual delas proporciona os melhores resultados na tarefa de classificação. (BATISTA GUSTAVO E. A. P. A., 2004).

4.1 Subamostragem Aleatória

A subamostragem aleatória trata o problema do desbalanceamento entre classes, reduzindo a cardinalidade da classe majoritária por meio da remoção aleatória de amostras. A mecânica desse processo de subamostragem aleatória está representada na Figura 10. Ao diminuir a predominância de uma classe sobre a outra, busca-se mitigar o viés do modelo, favorecendo uma aprendizagem mais equilibrada entre as categorias.

Figura 10 – Representação da Aplicação da Técnica de Subamostragem Aleatória



Fonte: Autor

A classe majoritária apresenta uma quantidade excessiva de amostras em comparação à classe minoritária. Com a aplicação da subamostragem, parte dessas amostras é removida, reduzindo sua densidade no espaço de atributos. Esse processo mantém a representatividade dos dados originais da classe majoritária, mas diminui seu domínio, promovendo um equilíbrio mais adequado entre as classes.

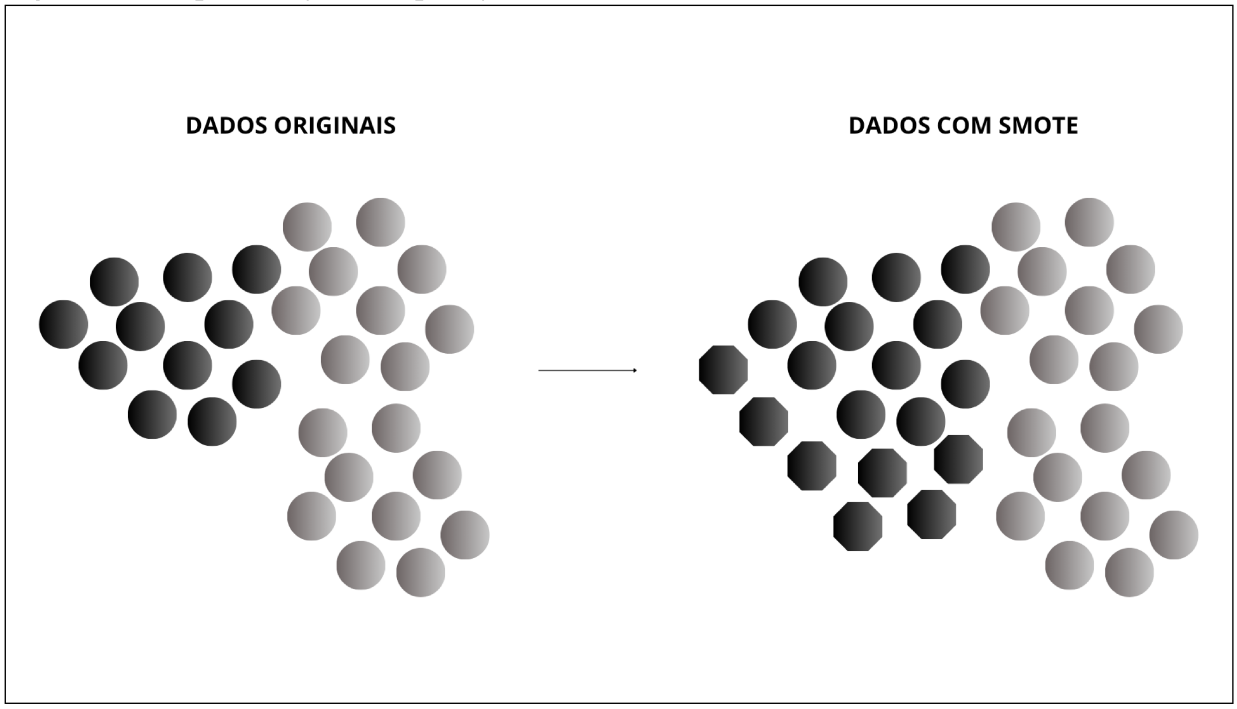
4.2 *SMOTE (Synthetic Minority Over-sampling Technique)*

O SMOTE é uma técnica que realiza superamostragem para corrigir o desbalanceamento entre classes. Ela atua no espaço vetorial das características, criando novas amostras da classe minoritária por meio de interpolação linear entre uma amostra e uma de suas vizinhas mais próximas, escolhida aleatoriamente (CHAWLA et al., 2002).

A interpolação é feita nas características extraídas do áudio, que são números que representam o som. Por isso, o SMOTE gera novos exemplos numéricos que são uma mistura suave das amostras originais. Mesmo sem criar áudios reais, o SMOTE funciona porque essas características são o que os modelos usam para aprender. Ao gerar variações plausíveis da classe minoritária, o SMOTE ajuda o modelo a entender melhor as diferenças e semelhanças, melhorando seu desempenho. O processo de geração dessas amostras sintéticas está ilustrado na

Figura 11.

Figura 11 – Representação da Aplicação da Técnica SMOTE



Fonte: Autor

A classe minoritária possui uma quantidade reduzida de amostras em relação à majoritária. Com a aplicação do SMOTE, novas amostras sintéticas são geradas, expandindo sua distribuição no espaço de atributos. Esse processo insere essas amostras ao redor dos dados originais, tornando a classe minoritária mais densa e promovendo um equilíbrio mais adequado entre as classes.

5 METODOLOGIA

Neste capítulo, são apresentados os materiais, procedimentos adotados e os experimentos realizados para alcançar os objetivos estabelecidos nesta pesquisa. Na Seção 5.1, é descrito o processo de coleta e seleção dos dados a partir da base FoR. A Seção 5.2 aborda as etapas necessárias de pré-processamento dos sinais de áudio, incluindo padronização, normalização e remoção de arquivos inválidos. Na Seção 5.3, são detalhadas as técnicas de extração de características dos sinais de áudio. A Seção 5.4 trata da divisão dos dados, apresentação dos modelos de aprendizado de máquina e a descrição das métricas para avaliação do desempenho dos modelos.

5.1 Coleta dos Dados

A base de dados que utilizamos neste estudo é a versão original da *FoR Dataset*, desenvolvida pela *Lassonde School of Engineering* (BIOLOGIACALLY INSPIRED LEARNING LAB, 2021). Ela contém aproximadamente 195 mil áudios monofônicos, com duração média de 2 segundos, taxa de amostragem de 16 kHz e formato WAV. Voltada para pesquisas relacionadas à detecção de manipulação de áudio, essa base contém gravações de fala autênticas e sintetizadas. Seu conjunto de dados inclui amostras de falas reais extraídas de outras bases como *Arctic Dataset*(CARNEGIE MELLON UNIVERSITY, 2004), *LJSpeech Dataset* (ITO, 2017) e *VoxForge Dataset* (VOXFORGE, 2020), além de falas geradas por sistemas de síntese de voz baseados em *TTS*.

Para compor o conjunto de dados utilizado no estudo, foram selecionadas 500 amostras reais e 200 amostras falsas a partir da base da *Lassonde School of Engineering*, configurando uma distribuição desbalanceada que simula cenários reais com disponibilidade limitada de dados manipulados. Essa configuração permite avaliar a robustez dos modelos diante da menor proporção de exemplos falsificados, possibilitando a análise do desempenho em condições que refletem os desafios típicos da detecção em ambientes práticos (NAKASHIMA *et al.*, 2024).

5.2 Pré-processamento

O pré-processamento é uma etapa que garante que os sinais de áudio estejam em um formato adequado para a extração de características. Áudios, coletados e rotulados como

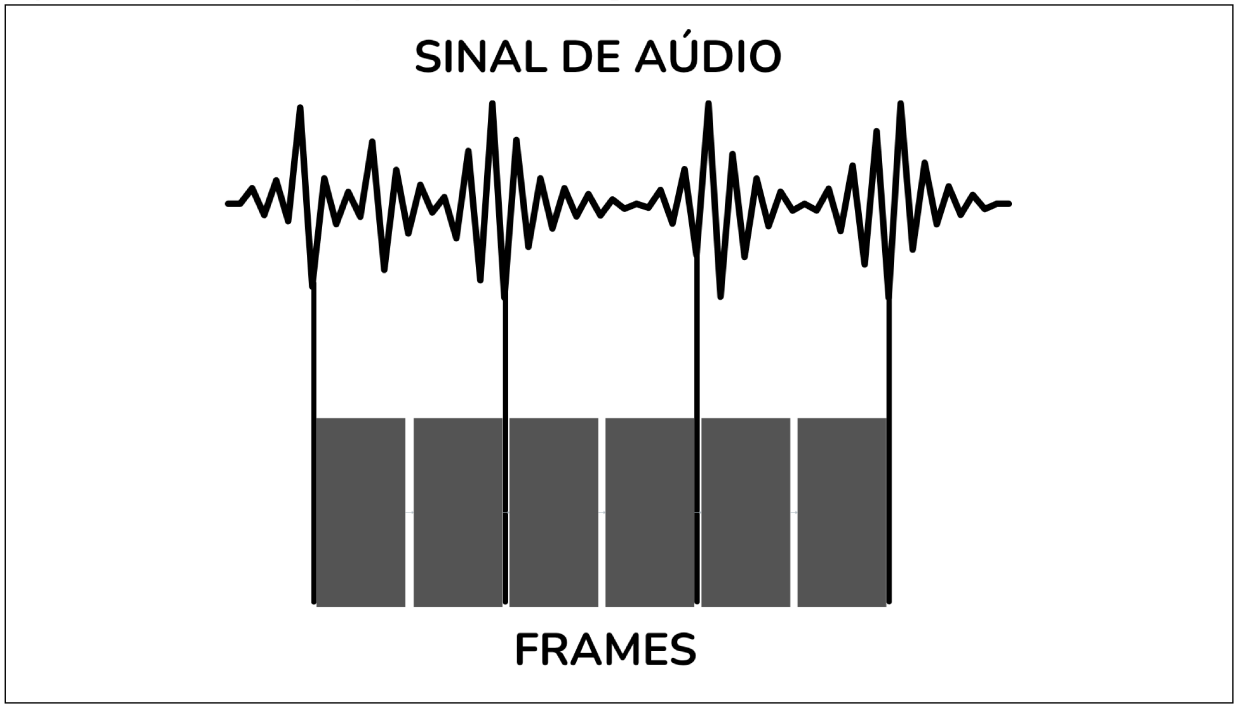
reais ou falsos, foram inicialmente padronizados quanto ao formato e à taxa de amostragem, empregando-se a biblioteca `librosa` (LIBROSA, 2022). Dessa forma, a taxa de amostragem dos arquivos foi redefinida para 16 kHz, assegurando uniformidade nos dados analisados.

Ademais, foi realizado um tratamento automatizado de exceções para identificar e descartar arquivos corrompidos, com extensões incompatíveis ou que não contivessem dados válidos. Por fim, normalizaram-se os sinais para eliminar discrepâncias de volume entre diferentes gravações, permitindo que os modelos de aprendizado se concentrem em padrões espectrais relevantes à tarefa de classificação e minimizando a influência de variações de intensidade sonora que não carregam informações úteis para a distinção entre classes.

5.3 Extração de Características

A extração de características do sinal de áudio pré-processado foi realizada a partir do Coeficientes Cepstrais de Frequência Mel (do inglês, *Mel Frequency Cepstral Coefficients*) (MFCCs), com o auxílio da biblioteca (LIBROSA, 2022). Esses coeficientes são empregados nas áreas de reconhecimento de fala e processamento de sinais acústicos, pois capturam informações perceptualmente relevantes do espectro de frequência (PEREIRA *et al.*, 2017). Sua extração envolve a segmentação do sinal em pequenas janelas temporais ou *frames*, geralmente parcialmente sobrepostas entre si. Esse último aspecto contribui para a preservação da continuidade espectral e uma representação mais precisa das variações dinâmicas do áudio. Esse processo está ilustrado na figura 12.

Figura 12 – Processo de Segmentação do Sinal para obtenção dos *frames*



Fonte: Autor

Inicialmente, é aplicada a transformada de Fourier a cada *frames* do sinal de áudio, convertendo o *frames* do domínio do tempo para o domínio da frequência:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j\frac{2\pi}{N}kn} \tag{5.1}$$

em que:

- $X(k)$ representa o espectro de frequência do sinal;
- $x(n)$ é o sinal no domínio do tempo;
- N é o número de amostras no *frame*;
- k é o índice da frequência discreta;
- j é a unidade imaginária, tal que $j^2 = -1$.

Em seguida, aplica-se um banco de filtros Mel, que simula a resposta do sistema auditivo humano. A saída de cada filtro é submetida a uma operação logarítmica, aproximando o sinal da percepção humana de intensidade sonora:

$$\log S_m = \log \left(\sum_{k=k_{m-1}}^{k_{m+1}} |X(k)|^2 \cdot H_m(k) \right) \tag{5.2}$$

em que:

- S_m representa o valor agregado pelo filtro m;
- $H_m(k)$ é o filtro na escala de Mel;
- $|X(k)|^2$ é a densidade espectral de potência.

Por fim, os valores $\log S_m$ são processados pela Transformada Discreta do Cosseno (do inglês, *Discrete Cosine Transform*) (DCT) para reduzir a correlação entre os coeficientes e gerar os *MFCCs*:

$$\text{MFCC}_n = \sum_{m=1}^{M} \log S_m \cdot \cos\left(\frac{\pi n}{M}(m-0.5)\right) \tag{5.3}$$

em que:

- MFCC_n é o coeficiente cepstral de frequência Mel;
- $\log S_m$ é o valor logarítmico da saída do filtro m;
- M é o número total de filtros Mel utilizados;
- m é o índice do filtro Mel, com $1 \leq m \leq M$;
- n é o índice do coeficiente cepstral desejado, com $1 \leq n \leq N_c$.

Ficou estabelecido o uso de 12 coeficientes *MFCCs* por *frame*, número considerado suficiente para capturar as informações espectrais essenciais sem introduzir ruído ou redundância (HEGDE *et al.*, 2015). Sendo essa configuração também adequada para os objetivos e características do presente trabalho. Ademais, calculou-se a média de cada coeficiente para compor uma única representação vetorial por áudio, reduzindo a variabilidade entre *frames* e resultando em uma entrada mais estável para os modelos de aprendizado.

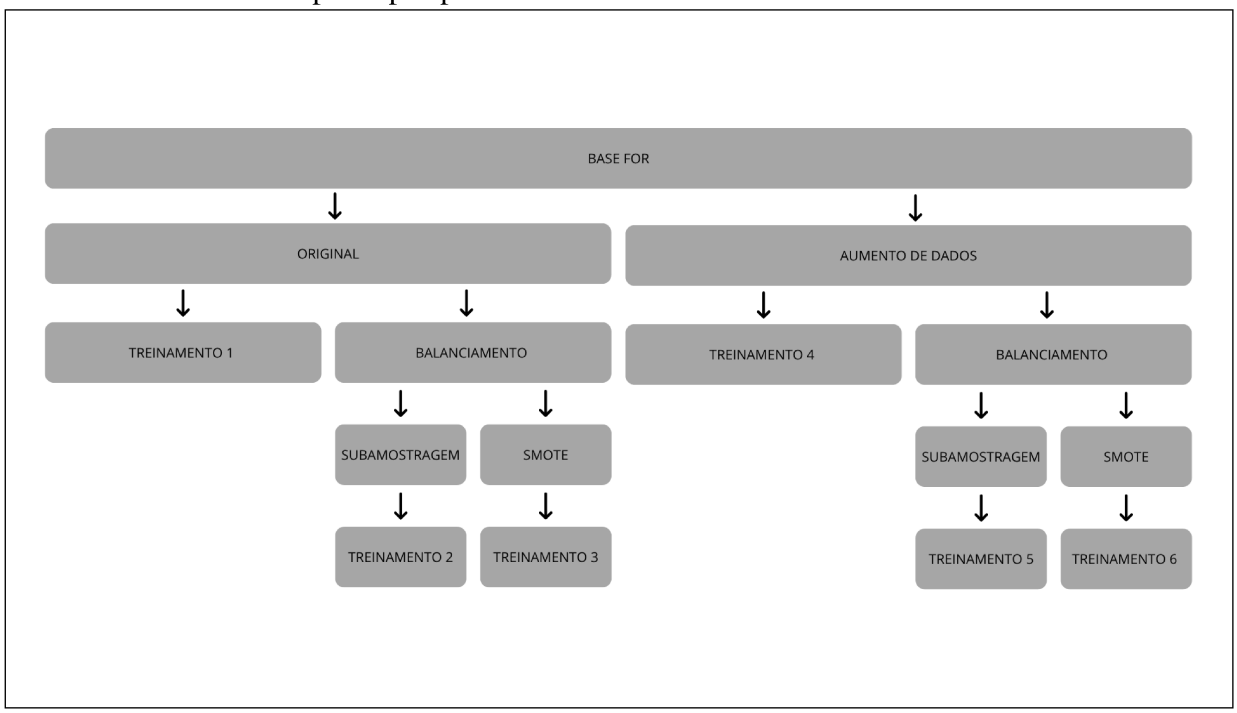
5.4 Topologia e Treinamento dos Modelos de Classificação

Os modelos avaliados neste estudo incluem algoritmos clássicos de aprendizado supervisionado: *Random Forest* (RF), *Light Gradient Boosting Machine* (LGBM), *Naive Bayes*, *Long Short-Term Memory* (LSTM) e *K-Nearest Neighbors* (KNN). Essa seleção contempla diferentes paradigmas de classificação, como métodos baseados em árvores, abordagens probabilísticas e redes neurais, permitindo avaliar a detecção de áudios falsos em cenários variados.

As implementações foram realizadas utilizando os hiperparâmetros padrão das bibliotecas *Scikit-learn* e *Keras*. Os únicos ajustes manuais feitos foram a definição do parâmetro *random state* para garantir a reprodutibilidade e a divisão dos dados em dois subconjuntos, sendo

80% para treinamento e 20% para teste. A estrutura dos conjuntos de dados para treinamento está ilustrada na Figura 13.

Figura 13 – Conjuntos de dados para treinamento dos modelos construídos a partir da base de áudio *FoR* após o pré-processamento



Fonte: Autor

Seis conjuntos de treinamento foram produzidos a partir das amostras pré-processadas da base *FoR*. Inicialmente, a base foi dividida em dois ramos: um com as amostras originais e outro com amostras geradas por técnicas de aumento de dados.

A partir das amostras originais, foi criado o conjunto Treinamento 1, utilizando os dados sem alterações. Em seguida, aplicaram-se duas estratégias distintas de balanceamento de classes: a subamostragem aleatória da classe majoritária, que originou o Treinamento 2, e a técnica de sobreamostragem da classe minoritária com *SMOTE*, que resultou no Treinamento 3.

No ramo de aumento de dados, foram aplicadas cinco técnicas de forma isolada, com apenas uma transformação por amostra: adição de ruído, variação de velocidade, alteração de tom, mascaramento tempo-frequência e equalização espectral aleatória. Esse processo gerou uma nova versão da base, utilizada na construção do conjunto Treinamento 4. A partir dele, também foram aplicadas as estratégias de balanceamento de classes: subamostragem, gerando o Treinamento 5, e *SMOTE*, gerando o Treinamento 6.

A avaliação dos modelos foi realizada com base em quatro métricas amplamente utilizadas em tarefas de classificação binária: acurácia, precisão, revocação e *F1-score* (FERRI

et al., 2009). Essas métricas permitem analisar o desempenho dos modelos sob diferentes perspectivas, considerando desde o número total de acertos até a capacidade de identificar corretamente as classes positivas e negativas, bem como o equilíbrio entre essas medidas.

Para garantir uma avaliação robusta, foi utilizada a técnica de validação cruzada do tipo *Stratified K-Fold*, com 10 subdivisões. Assim, cada modelo é treinado e testado 10 vezes, utilizando diferentes partições dos dados, e as métricas finais correspondem à média dos resultados obtidos em cada rodada (HASTIE *et al.*, 2009).

6 RESULTADOS E DISCUSSÕES

Nesse capítulo, apresentamos os resultados obtidos após a aplicação da metodologia descrita no capítulo anterior. Os dados estão organizados em tabelas, com os desempenhos obtidos a partir dos dados originais e após a aplicação das técnicas de ampliação: adição de ruído, mudança de velocidade, mudança de tom, mascaramento tempo-frequência e equalização aleatória. Também são apresentados os resultados após a aplicação do balanceamento de dados usando estratégias de balanceamento por *SMOTE* e subamostragem, permitindo analisar o impacto na performance dos modelos.

6.1 Dados Originais

Nesta seção, apresentamos o desempenho e o desvio padrão dos modelos de classificação de áudio com os dados originais, sem a aplicação de qualquer técnica de manipulação ou modificação. Esse cenário é utilizado como referência para comparar os efeitos das transformações aplicadas posteriormente. As métricas correspondentes estão apresentadas na Tabela 1.

Tabela 1 – Desempenho e Desvios Padrão dos modelos com os dados originais

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
RF	$0{,}82 \pm 0{,}07$	$0{,}87 \pm 0{,}05$	$0{,}89 \pm 0{,}06$	$0{,}87 \pm 0{,}05$
LGBM	$0{,}72 \pm 0{,}01$	$0{,}72 \pm 0{,}01$	$1{,}00 \pm 0{,}00$	$0{,}83 \pm 0{,}00$
Naive Bayes	$0{,}82 \pm 0{,}06$	$0{,}88 \pm 0{,}04$	$0{,}87 \pm 0{,}08$	$0{,}87 \pm 0{,}05$
LSTM	$0{,}75 \pm 0{,}05$	$0{,}77 \pm 0{,}05$	$0{,}94 \pm 0{,}06$	$0{,}84 \pm 0{,}03$
KNN	$0{,}78 \pm 0{,}06$	$0{,}81 \pm 0{,}03$	$0{,}90 \pm 0{,}06$	$0{,}85 \pm 0{,}04$

O Random Forest apresentou bom desempenho geral, com variação moderada entre as dobras, indicando certa sensibilidade à distribuição dos dados. Já o LGBM teve comportamento assimétrico, priorizando a identificação da classe minoritária, o que resultou em muitos falsos positivos. Apesar da estabilidade entre as dobras, sua tendência à superclassificação comprometeu a precisão.

O Naive Bayes manteve resultados estáveis, possivelmente devido à sua abordagem probabilística, que o torna menos dependente de grandes volumes de dados. Por outro lado, o LSTM teve desempenho inferior aos demais. Como depende de sequências temporais e maior volume de dados, encontrou dificuldades para generalizar adequadamente, além de apresentar

instabilidade entre as dobras.

Por fim, o KNN mostrou desempenho intermediário, mas foi afetado pela distribuição desbalanceada: em cenários assim, a classe majoritária tende a dominar as vizinhanças, dificultando a correta classificação da classe minoritária.

6.2 Dados Originais com subamostragem

Nesta subseção, são apresentados os resultados e o desvio padrão dos dados originais com a técnica de balanceamento de dados por subamostragem. As métricas estão organizadas na Tabela 2.

Tabela 2 – Desempenho e desvio padrão dos modelos com os dados originais e com a técnica de balanceamento de dados por subamostragem

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
RF	$0{,}74 \pm 0{,}06$	$0{,}75 \pm 0{,}06$	$0{,}72 \pm 0{,}11$	$0{,}73 \pm 0{,}07$
LGBM	$0{,}74 \pm 0{,}07$	$0{,}75 \pm 0{,}07$	$0{,}73 \pm 0{,}10$	$0{,}73 \pm 0{,}07$
Naive Bayes	$0{,}83 \pm 0{,}06$	$0{,}85 \pm 0{,}07$	$0{,}80 \pm 0{,}07$	$0{,}82 \pm 0{,}05$
LSTM	$0{,}64 \pm 0{,}08$	$0{,}65 \pm 0{,}08$	$0{,}61 \pm 0{,}12$	$0{,}63 \pm 0{,}09$
KNN	$0{,}76 \pm 0{,}05$	$0{,}84 \pm 0{,}06$	$0{,}65 \pm 0{,}10$	$0{,}73 \pm 0{,}07$

O Random Forest apresentou redução no desempenho geral e um leve aumento no desvio padrão, indicando maior instabilidade nos resultados após a subamostragem da classe majoritária devido ao número menor de amostras. Por outro lado, O LGBM mostrou melhora relativa na precisão, pois a subamostragem ajudou a equilibrar sua tendência à superclassificação, ainda assim o desempenho global foi inferior ao observado no cenário sem subamostragem, com desvio padrão indicando variações moderadas.

O Naive Bayes, mesmo adotando uma abordagem probabilística que tende a ser menos afetada pela quantidade de dados, apresentou comportamento misto, ou seja, houve leve melhora na acurácia, favorecida pelo balanceamento das classes, mas sofreu queda nas demais métricas, além de um leve aumento no desvio padrão, indicando menor estabilidade. Já o LSTM apresentou desempenho inferior e maior variação entre execuções, refletida em desvio padrão elevado, resultado da diminuição dos dados de treinamento, o que dificultou sua capacidade de generalização.

Por fim, o KNN teve seu desempenho afetado pela redução do volume de dados, apresentando queda geral; contudo, a subamostragem diminuiu a dominância da classe majo-

ritária nas vizinhanças, o que contribuiu para um desempenho mais preciso, com estabilidade semelhante à observada no cenário original.

6.3 Dados Originais com *SMOTE*

Nesta subseção, são apresentados os resultados e o desvio padrão dos dados originais com a técnica de balanceamento de dados por *SMOTE*. As métricas estão organizadas na Tabela 3.

Tabela 3 – Desempenho e o desvio padrão dos modelos com os dados originais e com a técnica de balanceamento de dados por *SMOTE*

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
RF	0,83 ± 0,03	0,84 ± 0,03	0,81 ± 0,07	0,83 ± 0,04
LGBM	0,81 ± 0,05	0,83 ± 0,07	0,81 ± 0,08	0,81 ± 0,06
Naive Bayes	0,83 ± 0,03	0,85 ± 0,04	0,81 ± 0,03	0,83 ± 0,04
LSTM	0,69 ± 0,07	0,69 ± 0,07	0,71 ± 0,12	0,70 ± 0,08
KNN	0,80 ± 0,04	0,89 ± 0,05	0,69 ± 0,06	0,78 ± 0,05

O Random Forest apresentou melhora no desempenho geral e redução no desvio padrão, indicando maior estabilidade nos resultados após a aplicação do SMOTE, que aumentou a representatividade da classe minoritária. Além disso, o LGBM também mostrou melhora relativa em todas as métricas, especialmente na precisão, pois o balanceamento das classes pelo SMOTE ajudou a mitigar sua tendência à superclassificação, sem perda de informações, resultando em desempenho superior ao observado nos outros cenários, com desvio padrão menor e mais consistente.

O Naive Bayes apresentou comportamento positivo, com melhora na acurácia e estabilidade nas demais métricas, refletida em menor desvio padrão. O LSTM, por sua vez, apresentou desempenho superior e menor variação entre execuções, resultado do aumento da base de dados proporcionado pelo SMOTE, o que facilitou sua capacidade de generalização.

Por fim, o KNN também apresentou melhora geral, pois o balanceamento das classes diminuiu a dominância da classe majoritária nas vizinhanças, contribuindo para um desempenho mais equilibrado e maior estabilidade em relação ao cenário original.

6.4 Dados Aumentados

Nesta seção, apresentamos o desempenho e o desvio padrão dos modelos de classificação de áudio após a aplicação das técnicas de ampliação de dados. Esse cenário é utilizado como referência para comparar os efeitos das transformações aplicadas posteriormente. As métricas correspondentes estão apresentadas na Tabela 4.

Tabela 4 – Desempenho e o desvio padrão dos modelos com aumento dos dados

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
RF	0,92 ± 0,02	0,95 ± 0,01	0,95 ± 0,02	0,95 ± 0,01
LGBM	0,72 ± 0,00	0,72 ± 0,00	1,00 ± 0,00	0,83 ± 0,00
Naive Bayes	0,85 ± 0,03	0,89 ± 0,03	0,90 ± 0,03	0,89 ± 0,02
LSTM	0,83 ± 0,04	0,85 ± 0,05	0,94 ± 0,05	0,89 ± 0,02
KNN	0,88 ± 0,03	0,88 ± 0,02	0,96 ± 0,02	0,92 ± 0,02

O Random Forest foi o modelo mais beneficiado pelo aumento de dados, apresentando desempenho superior em todas as métricas e maior estabilidade entre execuções. A diversidade introduzida consolidou sua robustez e o colocou acima dos demais cenários. O LGBM, embora tenha mantido alta sensibilidade, não apresentou avanços significativos em acurácia ou precisão, o que evidencia limitações internas mesmo com a base ampliada.

O Naive Bayes teve ganhos mais equilibrados, com melhorias moderadas nas métricas e leve redução na variabilidade, mostrando-se mais estável do que no cenário original, mas ainda inferior ao desempenho obtido com SMOTE. O LSTM também evoluiu, beneficiado pela maior quantidade de exemplos, mas continuou apresentando certa instabilidade entre execuções, indicando sensibilidade à forma de ampliação.

Por fim, o KNN demonstrou melhora tanto em desempenho quanto em estabilidade. O aumento dos dados reduziu a influência da classe majoritária nas vizinhanças, favorecendo decisões mais equilibradas e resultados mais consistentes em comparação aos demais cenários.

6.5 Dados Aumentados com subamostragem

Nesta subseção, são apresentados os resultados e o desvio padrão com as técnicas de aumento dos dados e com o balanceamento por subamostragem. As métricas estão organizadas na Tabela 5.

Tabela 5 – Desempenho e o desvio padrão dos modelos com aumento dos dados e com a técnica de balanceamento de dados por subamostragem

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
RF	$0{,}89 \pm 0{,}03$	$0{,}91 \pm 0{,}04$	$0{,}86 \pm 0{,}05$	$0{,}88 \pm 0{,}03$
LGBM	$0{,}84 \pm 0{,}02$	$0{,}84 \pm 0{,}04$	$0{,}86 \pm 0{,}04$	$0{,}85 \pm 0{,}02$
Naive Bayes	$0{,}83 \pm 0{,}03$	$0{,}83 \pm 0{,}04$	$0{,}85 \pm 0{,}04$	$0{,}83 \pm 0{,}03$
LSTM	$0{,}69 \pm 0{,}08$	$0{,}70 \pm 0{,}10$	$0{,}67 \pm 0{,}08$	$0{,}69 \pm 0{,}07$
KNN	$0{,}84 \pm 0{,}05$	$0{,}88 \pm 0{,}04$	$0{,}80 \pm 0{,}08$	$0{,}83 \pm 0{,}06$

A combinação entre aumento de dados e subamostragem resultou em desempenho superior ao observado com subamostragem isolada, indicando que o aumento dos dados ajuda a mitigar a perda de informação causada pela redução da classe majoritária. Porém, em comparação ao cenário com aumento dos dados sem balanceamento, houve uma leve redução nas métricas.

O Random Forest manteve bom desempenho com a combinação de aumento de dados e subamostragem, apresentando resultados sólidos em todas as métricas e variações moderadas entre execuções. Apesar de uma leve redução em relação ao cenário com aumento isolado, o modelo continuou estável e eficaz. O LGBM respondeu de forma positiva ao balanceamento, com uma melhora significativa nas métricas em comparação ao cenário original, além de estabilidade consistente, indicando que a técnica compensou parte de sua tendência à superclassificação.

O Naive Bayes teve desempenho estável e equilibrado, com métricas próximas às observadas no SMOTE e menor sensibilidade a variações entre execuções, sugerindo boa adaptação ao novo conjunto. Por outro lado, o LSTM foi o mais impactado negativamente, com queda expressiva em todas as métricas e aumento no desvio padrão, indicando prejuízo à sua capacidade de generalização, possivelmente pela perda de diversidade na subamostragem da classe majoritária.

Por fim, o KNN apresentou desempenho consistente, com resultados próximos ao SMOTE e ao aumento isolado, mas com leve crescimento na variabilidade. Ainda assim, o balanceamento contribuiu para decisões mais justas entre as classes, mantendo boa precisão e revocação.

6.6 Dados Aumentados com SMOTE

Nesta subseção, são apresentados os resultados e o desvio padrão com as técnicas de aumento dos dados e com o balanceamento por *SMOTE*. As métricas estão organizadas na Tabela 6.

Tabela 6 – Desempenho e o desvio padrão dos modelos com aumento dos dados e com a técnica de balanceamento de dados por *SMOTE*

Modelo	**Acurácia**	**Precisão**	**Revocação**	**F1-Score**
RF	0,95 ± 0,01	0,95 ± 0,02	0,94 ± 0,03	0,95 ± 0,01
LGBM	0,91 ± 0,01	0,93 ± 0,02	0,89 ± 0,01	0,91 ± 0,01
Naive Bayes	0,85 ± 0,02	0,83 ± 0,03	0,86 ± 0,04	0,85 ± 0,02
LSTM	0,83 ± 0,03	0,84 ± 0,06	0,84 ± 0,07	0,83 ± 0,03
KNN	0,87 ± 0,02	0,90 ± 0,03	0,85 ± 0,04	0,87 ± 0,02

A aplicação combinada das técnicas de aumento dos dados e balanceamento por *SMOTE* se destacou como o melhor cenário do estudo, proporcionando melhorias expressivas no desempenho dos modelos em relação aos cenários anteriores.

O Random Forest alcançou seu melhor desempenho neste cenário, com métricas elevadas e variação mínima entre execuções, evidenciando que a combinação de aumento de dados com SMOTE potencializou sua capacidade de generalização. O LGBM também obteve ganhos expressivos, superando os resultados dos demais cenários, com desempenho consistente em todas as métricas e excelente equilíbrio entre precisão e revocação.

O Naive Bayes manteve desempenho semelhante ao observado no SMOTE isolado, com resultados satisfatórios e leve melhora na estabilidade, mas sem avanços significativos. O LSTM, por sua vez, teve desempenho constante em relação aos demais cenários, com bons resultados gerais e redução moderada na variação entre execuções, ainda que menos expressiva do que em modelos mais simples.

Por fim, o KNN apresentou evolução estável, com melhora em todos os indicadores e boa consistência nos resultados. A combinação das duas técnicas contribuiu para decisões mais equilibradas, sem prejuízo à precisão, reforçando a eficácia da estratégia para esse tipo de modelo.

7 CONCLUSÃO

Este Trabalho de Conclusão de Curso teve como objetivo investigar o impacto de técnicas de ampliação de dados na detecção de manipulação de áudios de voz, utilizando amostras da base *FoR*, desenvolvida pela *Lassonde School of Engineering*. Com o avanço das tecnologias de clonagem de voz e síntese artificial, cresce a necessidade de desenvolver sistemas capazes de verificar a autenticidade de áudios. Para isso, buscou-se aprimorar o desempenho de modelos de aprendizado de máquina por meio da aplicação de métodos que aumentassem a diversidade dos dados de treinamento.

A pesquisa aplicou cinco técnicas específicas de ampliação de dados: adição de ruído, mudança de velocidade, mudança de tom, mascaramento tempo-frequência e equalização aleatória, com a finalidade de introduzir variações controladas nos sinais de áudio. Essas transformações simularam diferentes condições acústicas e fortaleceram a capacidade dos modelos de aprendizado de máquina em lidar com manipulações.

Além disso, foram aplicadas técnicas de balanceamento para corrigir o desequilíbrio entre as classes, uma vez que a base apresentava uma quantidade maior de áudios reais em relação aos áudios falsos. Para isso, utilizaram-se as técnicas de *SMOTE* e subamostragem. Essas estratégias evitaram que os modelos aprendessem padrões enviesados e contribuíram para melhorar sua capacidade de identificação de manipulações.

Os modelos de aprendizado de máquina testados abrangeram abordagens distintas, incluindo tanto algoritmos tradicionais quanto arquiteturas de redes neurais. Os modelos avaliados foram: *RF*, *LGBM*, *Naïve Bayes*, *KNN* e *LSTM*. Cada um desses modelos foi implementado e testado com o objetivo de identificar a abordagem mais eficaz para a detecção de manipulações em áudios.

Os resultados experimentais demonstraram que o modelo *RF* apresentou desempenho superior em relação às demais abordagens, quando combinado com a técnica de ampliação e com o balanceamento via *SMOTE*, alcançando acurácia de 95%, precisão de 95%, *recall* de 94% e *F1-score* de 95%. Em contrapartida, a arquitetura *LSTM* apresentou os piores resultados, possivelmente devido à complexidade e limitações do pré-processamento com *MFCCs*.

Por fim, este estudo contribui com a literatura e com a prática da ciência de dados aplicada à segurança da informação, ao fornecer estratégias de aumento e balanceamento de dados, contribuindo para o desenvolvimento de sistemas de detecção de áudios manipulados.

7.1 Trabalhos Futuros

Este trabalho abre caminho para várias possibilidades de pesquisa que podem ser exploradas no futuro. Algumas sugestões são:

- Testar parâmetros e hiperparâmetros para os modelos, buscando otimizar o desempenho.
- Criar uma base de validação específica para os modelos, melhorando a avaliação durante o treinamento.
- Utilizar técnicas de aprendizado auto-supervisionado, permitindo que o modelo aprenda com poucos dados rotulados.
- Experimentar outras técnicas de balanceamento de dados, visando reduzir desequilíbrios entre as classes.
- Explorar formas de otimizar o treinamento, buscando reduzir o tempo necessário para preparar os dados e ajustar os modelos.

REFERÊNCIAS

ALEX, A.; WANG, L.; GASTALDO, P.; CAVALLARO, A. Data augmentation for speech separation. **Speech Communication**, v. 152, p. 102949, 2023.

BATISTA GUSTAVO E. A. P. A., P. R. C. M. M. C. A study of the behavior of several methods for balancing machine learning training data. **ACM SIGKDD Explorations Newsletter**, ACM, v. 6, n. 1, p. 20–29, 2004.

BIANCO, M. J.; GERSTOFT, P.; TRAER, J.; OZANICH, E.; ROCH, M.; GANNOT, S.; DELEDALLE, C.-A. Machine learning in acoustics: Theory and applications. **The Journal of the Acoustical Society of America**, v. 146, n. 5, p. 3590–3628, 2019.

BIOLOGIACALLY INSPIRED LEARNING LAB. **Fake-or-Real (FoR) Dataset**. 2021. Acesso em: 07 abr. 2025. Disponível em: <https://bil.eecs.yorku.ca/datasets/#:~:text=The%20Fake-or-Real%20Dataset>.

BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais: Teoria e Aplicações**. [S.l.]: LTC, 2007.

CANTU, J. **Enhancing Speech Recognition Accuracy with Data Augmentation Techniques**. 2023. <https://medium.com/@jesus.cantu217/enhancing-speech-recognition-accuracy-with-data-augmentation-techniques-1debcb54628d>. Publicado em: 12 dez. 2023. Acesso em: 18 maio 2025.

CARNEGIE MELLON UNIVERSITY. **CMU Arctic Speech Database**. 2004. Acesso em: 07 abr. 2025. Disponível em: <http://festvox.org/cmu_arctic/>.

CASEY, E. **Handbook of Digital Forensics and Investigation**. [S.l.]: Academic Press, 2009. ISBN 9780123742674. ISBN 0123742676.

CATTIAU, J. **Recognizing impaired speech**. 2021. Google Blog, 8 nov. 2021. Acesso em: 13 maio 2025. Disponível em: <https://blog.google/outreach-initiatives/accessibility/impaired-speech-recognition/>.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. Smote: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, AAAI Press, v. 16, p. 321–357, 2002.

CHESNEY, R.; CITRON, D. K. Deepfakes: A looming challenge for privacy, democracy, and national security. **California Law Review**, v. 107, n. 6, p. 1753–1819, 2019.

DAS, R. K.; SHEN, T.; LEE, C. H. Detection of synthesized speech and deepfakes using raw audio features. In: **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**. [S.l.: s.n.], 2020.

FBI. **FBI Warns of Rise in Deepfake Content Used for Extortion and Fraud**. 2023. Acesso em: 13 maio 2025. Disponível em: <https://www.ic3.gov/PSA/2023/PSA230605>.

FERRI, C.; FLORES, A. J.; RODRíGUEZ, J. On the performance measures for binary classification. In: **Proceedings of the International Conference on Data Mining and Knowledge Engineering**. [S.l.]: ICDMKE, 2009. p. 1–12. Acesso em: 13 abr. 2025.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2nd. ed. New York, NY, USA: Springer, 2009. Disponível em: <https://web.stanford.edu/~hastie/ElemStatLearn/>.

HE, H.; GARCIA, E. A. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2008.

HEGDE, S.; ACHARY, K. K.; SHETTY, S. **Feature selection using Fisher's ratio technique for automatic speech recognition**. 2015. Acesso em: 07 mar. 2025. Disponível em: <https://arxiv.org/abs/1505.03239>.

ITO, K. **LJ Speech Dataset**. 2017. Acesso em: 07 abr. 2025. Disponível em: <https://keithito.com/LJ-Speech-Dataset/>.

KHANJANI, Z.; WATSON, G.; JANEJA, V. P. **How deep are the fakes? Focusing on audio deepfake: A survey**. 2021. CoRR, abs/2111.14203. Disponível em: <https://arxiv.org/abs/2111.14203>.

KO, T.; PEDDINTI, V.; POVEY, D.; KHUDANPUR, S. Audio augmentation for speech recognition. In: **Proceedings of Interspeech 2015**. [S.l.: s.n.], 2015. p. 3586–3589. Disponível em: <https://www.isca-speech.org/archive/interspeech_2015/ko15_interspeech.html>.

KORSHUNOV, P.; MARCEL, S. **DeepFakes: a New Threat to Face Recognition? Assessment and Detection**. 2018. <https://arxiv.org/abs/1812.08685>.

KUNDU, K. **Criminals Used AI To Clone Company Director's Voice And Steal $35 Million**. 2021. Screen Rant. Acesso em: 7 abr. 2025. Disponível em: <https://screenrant.com/ai-deepfake-cloned-voice-bank-scam-theft-millions/>.

LI, X.; ZHANG, Y.; ZHUANG, X.; LIU, D. **Frame-level SpecAugment for Deep Convolutional Neural Networks in Hybrid ASR Systems**. 2020. <https://arxiv.org/abs/2012.04094>. Acesso em: 18 maio 2025.

LIBROSA. **Librosa: Python package for music and audio analysis**. 2022. <https://librosa.org>. Acesso em: 7 jun. 2025.

MICHELSANTI, D. *et al.* An overview of deep-learning-based audio-visual speech enhancement and separation. **IEEE Transactions on Cognitive and Developmental Systems**, v. 23, n. 2, p. 345–364, 2021.

MÜLLER, T.; ROESNER, F.; KOHNO, T. "they hear you when you talk to them": Security and privacy of voice input. **Communications of the ACM**, v. 64, n. 3, p. 70–78, 2021.

NAKASHIMA *et al.* Avaliação de modelos para detecção de ataques de replay usando diferentes bases de dados. In: **Anais do 14. Simpósio Brasileiro de Tecnologia da Informação e Linguagem Humana (STIL)**. São Paulo: SBC, 2024. Acesso em: 8 mar. 2025. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/31109>.

OLIVEIRA, V. P. **Técnicas de Manipulação do Sinal de Áudio na Produção Musical**. 2020. Trabalho de Conclusão de Curso (Graduação em Música) – Universidade Federal de Santa Maria, Santa Maria, RS. Disponível em: <https://repositorio.ufsm.br/bitstream/handle/1/29827/Vinicius%20Oliveira%20-%20TCC%20.pdf?isAllowed=y&sequence=1>. Acesso em: 13 maio 2025.

PARK, D. S.; CHAN, W.; ZHANG, Y.; CHIU, C.-C.; ZOPH, B.; CUBUK, E. D.; LE, Q. V. Specaugment: A simple data augmentation method for automatic speech recognition. In: **INTERSPEECH 2019 – Conference of the International Speech Communication Association**. [S.l.: s.n.], 2019. p. 2613–2617. Disponível em: <https://arxiv.org/abs/1904.08779>. Acesso em: 16 maio 2025.

PEREIRA, J. G. S.; ARAUJO, E. N. L.; LINS, I. D.; CAVALCANTE, A. M. B. Análise comparativa de mfccs para reconhecimento de voz em diferentes ambientes. In: **Anais do XX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais - SBrT 2017**. São Pedro, SP: [s.n.], 2017.

PEREZ-LOPEZ, E. F. A.; SERRA, X. A hybrid parametric-deep learning approach for sound event localization and detection. In: **4th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2019)**. [s.n.], 2019. Disponível em: <https://doi.org/10.33682/v1za-0k45>.

RATNASARI, A. P. Performance of random oversampling, random undersampling, and smote-nc methods in handling imbalanced class in classification models. **International Journal of Scientific Research and Management (IJSRM)**, v. 12, n. 04, p. 494–501, 2024. Acesso em: 18 maio 2025. Disponível em: <https://ijsrm.net/index.php/ijsrm/article/view/5280>.

SALAMON, J.; BELLO, J. P. Deep convolutional neural networks and data augmentation for environmental sound classification. **IEEE Signal Processing Letters**, v. 24, n. 3, p. 279–283, 2017.

TANDOC, E. C.; LIM, Z. W.; LING, R. Defining “fake news”: A typology of scholarly definitions. **Digital Journalism**, v. 6, n. 2, p. 137–153, 2018.

VOXFORGE. **VoxForge Speech Corpus**. 2020. Acesso em: 07 abr. 2025. Disponível em: <http://www.voxforge.org>.

WAINER, J. **Comparison of 14 different families of classification algorithms on 115 binary datasets**. 2016. ArXiv preprint. Acesso em: 18 maio 2025. Disponível em: <https://arxiv.org/abs/1606.00930>.

YU, D.; LI, J. **Deep Learning for Audio**. [S.l.]: Springer, 2021. ISBN 978-3-030-47495-2.

ZHANG, Y.; SONG, X.; LIU, L. Text-to-speech synthesis: A review. **IEEE Transactions on Audio, Speech, and Language Processing**, 2021.