INSTITUTO FEDERAL
Pernambuco

CLASSIFICAÇÃO DE VOCALIZAÇÕES EM INDIVÍDUOS COM TEA: Avaliação de Modelos de Machine Learning

CLASSIFICATION OF VOCALIZATIONS IN INDIVIDUALS WITH ASD: Machine Learning Models Evaluation

Thiago Jorge Lins da Hora
tjlh@discente.ifpe.edu.br
Roberto Luiz Sena de Alencar
roberto.alencar@jaboatao.ifpe.edu.br

RESUMO

Este trabalho investiga a aplicação de técnicas de *machine learning* para a classificação de vocalizações não verbais de indivíduos minimamente verbais com Transtorno do Espectro Autista (TEA). Utilizando como base o conjunto de dados americano ReCANVo, foram exploradas seis categorias de vocalizações: *delight*, *dysregulated*, *frustrated*, *request*, *selftalk* e *social*. O processo envolveu a extração de características acústicas, o balanceamento de classes e o treinamento de múltiplos modelos. A avaliação experimental demonstrou que os modelos treinados exclusivamente com a base ReCANVo são ineficazes para generalizar ao contexto lusófono. Em um teste com 53 vocalizações de um indivíduo brasileiro, a acurácia de 75,47% obtida pelo melhor modelo (SVC) mostrou-se enganosa, concentrando-se em uma única classe majoritária e falhando completamente nas classes minoritárias. Tais resultados demonstram a ausência de generalização transcultural efetiva e evidenciam a necessidade crítica de dados alinhados ao contexto linguístico local. Para suprir essa lacuna, foi desenvolvido o aplicativo móvel *VocalizeAI*, que permitirá a criação de uma base de dados brasileira, fundamental para o avanço de tecnologias assistivas de comunicação para indivíduos com TEA.

Palavras-chave: Transtorno do Espectro Autista, vocalizações não verbais, *machine learning*, classificação de áudio, ReCANVo.

ABSTRACT

This study investigates the application of machine learning techniques for classifying nonverbal vocalizations of minimally verbal individuals with Autism Spectrum Disorder (ASD). Based on the American ReCANVo dataset, six categories of vocalizations were explored: *delight*, *dysregulated*, *frustrated*, *request*, *selftalk*, and *social*. The process involved acoustic feature extraction, class balancing, and training multiple models. The experimental evaluation demonstrated that models trained exclusively on the ReCANVo dataset are ineffective at generalizing to the Portuguese-speaking context. In a test

Instituto Federal de Pernambuco. Campus Jaboatão dos Guararapes. Curso Análise e Desenvolvimento de Sistemas. 26 de novembro de 2025.

with 53 vocalizations from a Brazilian individual, the 75.47% accuracy achieved by the best model (SVC) proved to be misleading, as it was concentrated on a single over-represented class while failing on the minority classes. These results demonstrate the lack of effective cross-cultural generalization and highlight the critical need for data aligned with the local linguistic context. To address this gap, the mobile application *VocalizeAI* was developed to enable the creation of a Brazilian dataset, which is essential for advancing assistive communication technologies for individuals with ASD.

Keywords: Autism Spectrum Disorder, nonverbal vocalizations, machine learning, audio classification, ReCANVo.

---

1 INTRODUÇÃO

O Transtorno do Espectro Autista (TEA) é uma condição do neurodesenvolvimento caracterizada por dificuldades no relacionamento social, repertório repetitivo e estereotipado de comportamentos, desafios na comunicação e insistência em rotinas não funcionais (AMERICAN PSYCHIATRIC ASSOCIATION, 2013). As manifestações do TEA variam significativamente entre indivíduos, abrangendo uma ampla gama de habilidades e desafios. Dentro deste espectro, um foco particular deste trabalho são os indivíduos minimamente verbais, definidos como aqueles que produzem entre 0 e 20 palavras ou aproximações de palavras faladas (NARAIN et al., 2020).

O uso de tecnologias assistivas tem se mostrado um facilitador essencial para a inclusão de pessoas com TEA em diversos contextos sociais (LAZAR; GOLDSTEIN; TAYLOR, 2015). Essas tecnologias, também conhecidas como tecnologias de apoio, englobam dispositivos, equipamentos, sistemas e estratégias projetadas para auxiliar pessoas com deficiência na superação de barreiras, promovendo sua independência, qualidade de vida e inclusão em diferentes áreas, como educação, trabalho e atividades cotidianas (BERSCH, 2008).

Superar barreiras comunicativas é um dos maiores desafios enfrentados por crianças no TEA, e as tecnologias assistivas têm se mostrado aliadas fundamentais nesse processo (DELL; NEWTON; PETROFF, 2008). Entre essas soluções, destaca-se a Comunicação Alternativa e Aumentativa (CAA), definida como um conjunto de métodos, recursos e tecnologias que têm como objetivo complementar (aumentativa) ou substituir (alternativa) a fala e a escrita em indivíduos com dificuldades significativas de comunicação (BEUKELMAN; MIRENDA, 2013). A CAA pode incluir desde estratégias sem auxílio, como gestos e expressões faciais, até sistemas com auxílio, como pranchas de símbolos, aplicativos digitais e dispositivos geradores de fala. Seu propósito central é garantir oportunidades de comunicação efetiva, promovendo autonomia, participação social e inclusão.

Ferramentas como os Comunicadores de Voz e as Tecnologias de Comunicação Aumentativa e Alternativa oferecem a essas crianças a possibilidade de expressarem suas vontades, emoções e pensamentos, muitas vezes pela primeira vez (COOK; POLGAR, 2014). Em contextos escolares, por exemplo, essas soluções viabilizam interações que antes eram impossíveis, promovendo inclusão e participação mais ativa.

Aplicativos voltados à Comunicação Social e ao Rastreamento de Emoções também desempenham um papel essencial, ao permitir que a criança não apenas compreenda, mas também consiga representar visualmente seus sentimentos. Tais recursos ampliam a capacidade de diálogo com o mundo ao redor, tornando as emoções mais acessíveis e inteligíveis, tanto para a própria criança quanto para aqueles que convivem com ela (KONECKI; LOVRENČIĆ; JERVIS, 2016).

As vocalizações de indivíduos minimamente verbais apresentam características acústicas únicas, que variam de forma significativa entre os sujeitos, tornando-se altamente específicas a cada pessoa. Essa singularidade dificulta a interpretação por parte de ouvintes não familiarizados, representando um desafio adicional à comunicação (JOHNSON; NARAIN et al., 2023). Tal particularidade configura uma barreira expressiva para a interação social efetiva, sobretudo em contextos nos quais o indivíduo se comunica com pessoas fora de seu círculo familiar ou de convivência próxima.

A problemática central desta pesquisa emerge da intersecção entre a necessidade crítica de compreender essas vocalizações e os desafios técnicos inerentes ao seu processamento automatizado. Do ponto de vista técnico, a classificação automática de vocalizações não verbais enfrenta diversos obstáculos: o ruído presente em dados coletados em ambientes naturais, o desbalanceamento de classes no conjunto de dados, a identificação de características acústicas verdadeiramente representativas das diferentes categorias comunicativas e a variabilidade individual significativa que caracteriza essas expressões (NARAIN et al., 2020).

Adicionalmente, existe uma lacuna crítica na disponibilidade de recursos de dados para diferentes contextos culturais e linguísticos. Até o presente momento, o banco de dados ReCANVo representa o único conjunto de dados disponível mundialmente com vocalizações não verbais de indivíduos minimamente verbais, tendo sido desenvolvido exclusivamente com base em dados coletados nos Estados Unidos (JOHNSON; NARAIN et al., 2023). Esta singularidade do ReCANVo, embora represente um avanço científico fundamental, simultaneamente cria uma limitação significativa para a aplicabilidade de tecnologias assistivas desenvolvidas a partir deste conjunto em outras culturas e contextos linguísticos.

Esta limitação foi empiricamente investigada através da formulação da seguinte Questão de Pesquisa: *Modelos de aprendizado de máquina treinados exclusivamente com datasets norte-americanos (ReCANVo) possuem capacidade de generalização suficiente para classificar vocalizações de indivíduos lusófonos com TEA?*

A comunicação é um direito fundamental que permeia todas as dimensões da experiência humana. Para indivíduos minimamente verbais (mv) com TEA, que produzem entre 0 e 20 palavras ou aproximações de palavras faladas (NARAIN et al., 2020), as vocalizações não verbais constituem o principal meio de expressão e interação com o mundo. Essas vocalizações, embora não sigam as convenções da linguagem verbal tradicional, podem carregar significados profundos sobre estados emocionais, necessidades e intenções comunicativas.

A relevância desta pesquisa fundamenta-se em múltiplas dimensões que convergem para um impacto significativo tanto na esfera científica quanto social. Do ponto de vista social, compreender e interpretar essas vocalizações é fundamental para melhorar a qualidade de vida dos indivíduos mv, pois permite que cuidadores, familiares

e profissionais de saúde respondam de forma mais adequada às suas necessidades (JOHNSON; O'BRIEN et al., 2022). A capacidade de interpretar corretamente uma vocalização de frustração, por exemplo, pode prevenir situações de estresse e promover intervenções mais eficazes.

As vocalizações não verbais podem expressar uma variedade de estados emocionais e intenções comunicativas que, quando adequadamente interpretadas, contribuem com percepções significativas sobre os processos cognitivos e emocionais fundamentais (JOHNSON; NARAIN et al., 2023). Tecnologicamente, a motivação reside na possibilidade de democratizar ferramentas assistivas através do desenvolvimento de sistemas automatizados baseados em inteligência artificial. Enquanto a interpretação dessas vocalizações tradicionalmente depende de familiaridade ou conhecimento contextual específico, limitando sua aplicabilidade a círculos próximos do indivíduo, a automação através de *machine learning* pode tornar essa capacidade interpretativa acessível a uma gama muito mais ampla de profissionais e contextos (NARAIN et al., 2020).

Assim, considerando o potencial das tecnologias baseadas em aprendizado de máquina para ampliar o acesso a ferramentas assistivas e a relevância da interpretação automatizada de vocalizações não verbais, este estudo tem como objetivo avaliar a robustez e a aplicabilidade do banco de dados ReCANVo na classificação de vocalizações de indivíduos minimamente verbais, com foco na representatividade e adequação dessas vocalizações no contexto da língua portuguesa. A análise será conduzida por meio da aplicação de diferentes algoritmos de *machine learning*, com o objetivo de estabelecer parâmetros iniciais de desempenho, além de identificar as potencialidades e limitações inerentes ao ReCANVo para seu uso em futuras tecnologias assistivas voltadas à comunicação não verbal.

Para suprir a lacuna crítica de dados locais identificada, este trabalho não se limitou à análise teórica, mas incluiu também o desenvolvimento e a implementação inicial do aplicativo móvel *VocalizeAI*. Esta ferramenta foi projetada para viabilizar a coleta distribuída e padronizada de vocalizações no contexto lusófono, constituindo o primeiro passo prático para a validação dos modelos em ambiente real.

2 REFERENCIAL TEÓRICO

Esta seção apresenta os fundamentos teóricos que sustentam o presente trabalho. A abordagem interdisciplinar adotada combina conceitos da psicologia e da ciência da computação para explorar como a tecnologia pode ser aplicada para compreender melhor a comunicação em indivíduos no Transtorno do Espectro Autista (TEA). Inicia-se com uma contextualização sobre o TEA, com foco especial nas vocalizações não verbais (VNVs) como uma forma de comunicação essencial. Em seguida, são detalhados os princípios do Aprendizado de Máquina (*Machine Learning* - ML) e os algoritmos específicos empregados para a classificação de áudio. Posteriormente, descrevem-se as técnicas de extração de características acústicas, que são cruciais para traduzir sinais sonoros em um formato que os modelos de ML possam processar. Por fim, abordam-se as estratégias para o tratamento de dados desbalanceados e os métodos para uma avaliação rigorosa do desempenho dos modelos, garantindo a robustez e a validade dos resultados obtidos.

2.1 Transtorno do Espectro Autista (TEA) e Comunicação

O Transtorno do Espectro Autista (TEA) é um transtorno do neurodesenvolvimento caracterizado por déficits persistentes na comunicação social e na interação social em múltiplos contextos, além de padrões restritos e repetitivos de comportamento, interesses ou atividades (AMERICAN PSYCHIATRIC ASSOCIATION, 2013). Indivíduos no espectro, especialmente aqueles considerados minimamente verbais (que produzem poucas ou nenhuma palavra funcional), frequentemente dependem de vocalizações não verbais (VNVs) para expressar necessidades, emoções e intenções. Para esta população, as vocalizações não verbais, em conjunto com outras expressões como gestos e expressões faciais, constituem modos de comunicação eficazes para expressar suas intenções, emoções e necessidades (JOHNSON; NARAIN et al., 2023). A interpretação precisa dessas VNVs por cuidadores e profissionais é crucial para o bem-estar e o desenvolvimento do indivíduo.

Para melhor caracterizar a heterogeneidade do espectro, o DSM-5 classifica o TEA em três níveis de suporte, baseados na intensidade do apoio necessário nas áreas de comunicação social e comportamentos restritos e repetitivos (AMERICAN PSYCHIATRIC ASSOCIATION, 2013). Essa classificação ajuda a delinear as diferentes necessidades de cada pessoa:

- Nível 1 - ”Exigindo apoio”: Indivíduos neste nível podem apresentar dificuldades em iniciar interações sociais e respostas atípicas ou sem sucesso às aberturas sociais de outros. Geralmente, possuem habilidades de linguagem verbal, mas enfrentam desafios na conversação e na compreensão de nuances sociais.
- Nível 2 - ”Exigindo apoio substancial”: Neste nível, os déficits na comunicação social verbal e não verbal são mais acentuados e aparentes, mesmo com apoios implementados. A pessoa pode ter uma fala com frases simples, um vocabulário limitado e uma interação social restrita a interesses específicos.
- Nível 3 - ”Exigindo apoio muito substancial”: Caracteriza-se por déficits graves nas habilidades de comunicação social verbal e não verbal que causam prejuízos severos no funcionamento. Indivíduos neste nível podem ser minimamente verbais ou não verbais, como mencionado anteriormente, utilizando vocalizações não verbais como forma primária de expressão. A iniciativa para interação social é muito limitada e a resposta às interações de outros é mínima.

A relevância do estudo do TEA é acentuada pelo contínuo aumento de sua prevalência, documentado por programas de vigilância ativa. De acordo com o mais recente relatório da Rede de Monitoramento de Autismo e Deficiências do Desenvolvimento (ADDM) dos Centros de Controle e Prevenção de Doenças (CDC) dos EUA, referente ao ano de 2022, a prevalência geral de TEA foi de 32,2 por 1.000 crianças de 8 anos, o que equivale a uma em cada 31 crianças (SHAW et al., 2025). Este dado representa um aumento substancial em relação a anos anteriores; a prevalência geral entre os sítios de pesquisa foi 22,2% maior em 2022 do que a registrada em 2020. O relatório também aponta que a prevalência do TEA continuou a variar amplamente entre as diferentes comunidades pesquisadas, de 9,7 por 1.000 crianças no Texas (Laredo) a 53,1

por 1.000 na Califórnia, sugerindo diferenças nas práticas de identificação e acesso a serviços.

2.2 Gradient Boosting

O *Gradient Boosting* é um técnica de aprendizado de conjunto (*ensemble learning*) proposta por Friedman (2001), que consiste em treinar uma sequência de modelos fracos (geralmente árvores de decisão), onde cada novo modelo é otimizado para corrigir os erros cometidos pelo conjunto anterior. A otimização é realizada por meio da descida de gradiente aplicada sobre uma função de perda arbitrária. O *gradient boosting* se tornou a base de algoritmos altamente eficientes e amplamente utilizados, como o XGBoost e o LightGBM.

2.3 Otimização de Hiperparâmetros

Para garantir que os modelos de *machine learning* atinjam seu desempenho máximo e que sua avaliação seja robusta e imparcial, foram empregadas técnicas sistemáticas de otimização e validação. Esta seção detalha os conceitos de hiperparâmetros, o método de busca em grade (*Grid Search*) com validação cruzada para sua otimização e as métricas utilizadas para quantificar o desempenho dos classificadores.

2.3.1 Hiperparâmetros e Otimização com GridSearchCV

Diferente dos parâmetros de um modelo (como pesos aprendidos), os **hiperparâmetros** são configurações externas definidas antes do treinamento e influenciam a complexidade e o comportamento do modelo. Eles são essenciais para o desempenho ideal de algoritmos como SVM (com hiperparâmetros *C* e *gamma*) e Random Forest (com *n_estimators* e *max_depth*). A importância e métodos sistemáticos de otimização de hiperparâmetros estão bem fundamentados na literatura (BISCHL et al., 2021).

Para encontrar as melhores combinações, utilizamos o **GridSearchCV**, uma busca exaustiva em um *grid* de valores, juntamente com validação cruzada interna para avaliação, conforme implementado em Scikit-learn (PEDREGOSA et al., 2011).

2.3.2 Validação Cruzada Estratificada

A fim de evitar avaliações enviesadas (muito otimistas) e depender de uma única divisão treino-teste, aplicou-se a **validação cruzada estratificada k-fold**. O conjunto é dividido em k *folds*, preservando-se a proporção das classes (estratificação), com cada *fold* usado uma vez como validação. O desempenho médio sobre os k *folds* oferece uma estimativa mais estável e menos suscetível a *overfitting*. A eficácia dessa abordagem para seleção de modelos é bem discutida por Arlot e Celisse (2010).

2.4 Aprendizado de Máquina para Classificação de Áudio

O aprendizado de máquina (*Machine Learning* - ML) é um campo da inteligência artificial que se concentra no desenvolvimento de algoritmos que permitem aos compu-

tadores aprender a partir de dados. No contexto da classificação de áudio, o objetivo é treinar um modelo para reconhecer e categorizar sons com base em suas características acústicas. Para isso, são utilizados diversos algoritmos, cada um com suas particularidades.

- **Random Forest**: Proposto por Breiman (2001), é um algoritmo de conjunto (*ensemble*) que constrói múltiplas árvores de decisão durante o treinamento e produz a classe que é o modo das classes (classificação) das árvores individuais.
- **Support Vector Machine (SVM)**: Desenvolvido por Burges (1998), o SVM busca encontrar um hiperplano ótimo que maximize a margem de separação entre diferentes classes em um espaço de características.
- **XGBoost**: Desenvolvido por Chen e Guestrin (2016), o XGBoost (*Extreme Gradient Boosting*) é uma implementação otimizada e escalável do algoritmo de *gradient boosting*. O *gradient boosting* consiste em treinar modelos fracos (geralmente árvores de decisão) de forma sequencial, onde cada novo modelo é ajustado para corrigir os erros do conjunto anterior por meio da descida de gradiente sobre a função de perda.
- **LightGBM**: É uma implementação de *gradient boosting* altamente eficiente e escalável desenvolvida pela Microsoft (KE et al., 2017). Assim como o XGBoost, utiliza o princípio do *gradient boosting*, mas introduz duas técnicas centrais para ganho de eficiência:
  - **Gradient-based One-Side Sampling (GOSS)**: Método que reduz o número de instâncias utilizadas em cada iteração, preservando aquelas com maiores resíduos (gradientes), que são as mais informativas para o treinamento. Dessa forma, acelera o processo sem perder muita precisão.
  - **Exclusive Feature Bundling (EFB)**: Técnica que combina características mutuamente exclusivas (isto é, que raramente assumem valores diferentes de zero simultaneamente) em uma única característica. Isso reduz a dimensionalidade efetiva dos dados e diminui o custo computacional.

2.5 Extração de Características Acústicas

Sinais de áudio brutos são complexos e não podem ser diretamente utilizados pela maioria dos modelos de aprendizado de máquina. Por isso, é necessário transformá-los em representações numéricas que capturem as propriedades relevantes do som. Para essa tarefa, foi implementada uma função de extração de características baseada na biblioteca `Librosa` (MCFEE et al., 2015), com suporte adicional da `Pydub` para padronização dos arquivos de entrada.

O procedimento adotado segue as seguintes etapas:

1. **Pré-processamento do áudio**:
   - O arquivo é carregado com a biblioteca `Pydub`, convertido para **mono** e padronizado com taxa de amostragem de **16 kHz**, garantindo consistência entre diferentes gravações.

- Caso o sinal seja muito curto, é aplicado *padding* para assegurar um comprimento mínimo compatível com a janela de análise.

2. **Extração de características espectrais e temporais**: A função extrai múltiplas representações complementares do sinal, calculando tanto a **média** quanto o **desvio padrão** de cada uma, o que permite capturar não apenas valores centrais, mas também a variação das medidas ao longo do tempo. As características consideradas são:

   - **Coeficientes Cepstrais de Frequência Mel (MFCCs)**: 40 coeficientes que descrevem a forma do espectro em uma escala perceptualmente relevante, aproximando a percepção auditiva humana (MCFEE et al., 2015).
   - **Espectrograma Mel**: é uma representação visual da distribuição de energia do sinal do domínio tempo-frequência em uma escala perceptual, no qual, captura informações complementares aos MFCCs (MCFEE et al., 2015).
   - **Cromaticidade (Chroma Features)**: distribuição da energia espectral em 12 classes correspondentes às notas musicais, relacionada à harmonia e melodia (MCFEE et al., 2015).
   - **Contraste Espectral**: diferença de energia entre picos e vales do espectro, útil para diferenciar timbres e texturas sonoras (MCFEE et al., 2015).
   - **Taxa de Cruzamento por Zero (Zero Crossing Rate, ZCR)**: frequência de mudanças de sinal positivo para negativo, associada à presença de ruídos e sons de alta frequência (GOUYON; PACHET; DELERUE, 2002).
   - **Roll-off Espectral**: frequência abaixo da qual se encontra uma porcentagem da energia espectral total, frequentemente associada à noção de "brilho" do som (MCFEE et al., 2015).
   - **Centróide Espectral**: ponto de "gravidade" do espectro, também relacionado ao brilho e percepção de intensidade dos agudos (MCFEE et al., 2015).

2.6 Classificação de Vocalizações Não Verbais em Indivíduos com TEA

As vocalizações não verbais representam uma via expressiva relevante para a comunicação de indivíduos com Transtorno do Espectro Autista (TEA), podendo refletir diferentes estados emocionais e intenções comunicativas. Neste trabalho, utiliza-se aprendizado de máquina para classificar essas vocalizações a partir da base de dados ReCANVo, que categoriza os sons em seis grupos distintos: *delight*, *dysregulated*, *frustrated*, *request*, *selftalk* e *social*. Cada categoria apresenta características acústicas próprias, associadas a emoções ou funções específicas. Vocalizações de *frustration*, por exemplo, indicam insatisfação e são marcadas por tons mais intensos; *delight* envolve sons agudos e rápidos relacionados à alegria; enquanto *dysregulation* reflete estados de desconforto ou superestimulação. O *selftalk* é caracterizado por sons exploratórios e repetitivos, sem função comunicativa clara, ao passo que *request* representa tentativas de solicitar algo, e *social* inclui vocalizações com intenção de interação. A compreensão e a classificação dessas categorias são fundamentais para o desenvolvimento de tecnologias assistivas e intervenções clínicas mais eficazes (JOHNSON; NARAIN et al., 2023).

2.7 Balanceamento de Classes com SMOTE

Em problemas de classificação do mundo real, é comum que os dados sejam desbalanceados, ou seja, algumas classes possuem muito mais amostras do que outras. A técnica **SMOTE (Synthetic Minority Over-sampling Technique)** (CHAWLA et al., 2002) aborda esse problema criando amostras sintéticas da classe minoritária, ajudando a evitar que o modelo se torne enviesado para a classe majoritária.

2.8 Métricas de Avaliação

A avaliação de um modelo de classificação não deve se basear apenas na acurácia, especialmente em contextos com classes desbalanceadas. Para uma análise completa, é fundamental utilizar uma matriz de confusão, que tabula os resultados em quatro categorias: Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN) (FAWCETT, 2006). A partir delas, derivam-se as seguintes métricas:

- **Acurácia (Accuracy)**: Representa a proporção de predições corretas sobre o total de amostras. Embora intuitiva, pode ser enganosa em dados desbalanceados.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}$$

- **Precisão (Precision)**: Mede a proporção de predições positivas que estavam de fato corretas. É uma métrica importante quando o custo de um falso positivo é alto.

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

- **Recall (Sensibilidade ou Revocação)**: Mede a proporção de amostras positivas que foram corretamente identificadas pelo modelo. É crucial quando o custo de um falso negativo é alto.

$$\text{Recall} = \frac{VP}{VP + FN}$$

- **F1-Score**: É a média harmônica entre a precisão e o *recall*. Fornece um valor único que equilibra ambas as métricas, sendo particularmente útil para comparar modelos em cenários com classes desbalanceadas.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} = \frac{2 \cdot VP}{2 \cdot VP + FP + FN}$$

3 TRABALHOS RELACIONADOS

A análise de vocalizações não-verbais (NVVs), como risos, suspiros, choros e outros sons expressivos, representa uma área de pesquisa crucial e desafiadora na interseção entre processamento de áudio, computação afetiva e acessibilidade. Diferentemente da fala, as NVVs possuem características acústicas únicas que os modelos de

linguagem e fala tradicionais frequentemente não conseguem capturar de forma eficaz. Para superar essa lacuna, a literatura recente tem explorado diversas frentes. Estas incluem desde a adaptação de Grandes Modelos de Fala (LSMs) pré-treinados, aplicando técnicas de ajuste fino eficiente em parâmetros (PEFT), até o desenvolvimento de novos modelos de fundação especializados em áudio não-verbal. Adicionalmente, investiga-se o impacto de novas metodologias de coleta de dados, como a detecção egocêntrica, e estratégias de treinamento robustas, como a aprendizagem contrastiva, para lidar com a escassez e o desbalanceamento de dados comuns neste domínio.

Em um esforço para aprimorar a análise de interações clinicamente relevantes, Feng et al. (2025) investigou o uso de detecção de áudio egocêntrica para a classificação de falantes em interações diádicas entre crianças e adultos, especialmente no contexto de avaliações do Transtorno do Espectro Autista (TEA). Os autores propuseram uma abordagem que utiliza sensores de áudio vestíveis para capturar a fala na perspectiva de primeira pessoa, superando as limitações dos microfones de campo distante. O estudo demonstrou que pré-treinar um modelo wav2vec 2.0 com um grande conjunto de dados de fala egocêntrica (Ego4D) melhora significativamente a precisão da classificação. Os resultados evidenciaram que essa estratégia, especialmente quando combinada com uma entrada de canal duplo (áudio de ambos os participantes), alcançou um aumento de 0.08 no macro F1 em comparação com métodos de gravação convencionais, ressaltando o potencial da detecção egocêntrica e do pré-treinamento especializado para a análise automática da comunicação social.

Abordando a exclusão de indivíduos não-verbais em sistemas de identificação de pessoas (PID) baseados em voz, Tran e Tsai (2024) conduziu uma das primeiras investigações sobre o uso de vocalizações não-verbais para a identificação de indivíduos não-falantes e com fala mínima (NMS). Os autores introduziram os conceitos de NMS-PID (identificação exclusiva de usuários NMS) e S-NMS-PID (um sistema unificado para usuários NMS e falantes), utilizando o *dataset* ReCANVo. A metodologia central envolveu uma estratégia de treinamento em duas etapas, empregando Aprendizagem Contrastiva Supervisionada (SCL) para a aprendizagem de representações, seguida de um ajuste fino. Os resultados demonstraram a alta viabilidade da abordagem, com acurácias de até 92.61% para o sistema S-NMS-PID. Notavelmente, o modelo proposto de Rede Neural Convolucional Recorrente (CRNN), apesar de menor, superou modelos mais profundos como VGG16 e ResNet50 quando treinado com SCL, mostrando-se especialmente eficaz na melhoria da acurácia para classes com poucos dados (minoritárias) e na robustez a ruídos.

Um estudo recente de Di Automatica E Informatica et al. (2025) investigou a adaptabilidade de grandes modelos de fala (*Large Speech Models* – LSMs) a tarefas de vocalizações não-verbais (*Non-Verbal Vocalizations* – NVVs), como risos, suspiros e gemidos. Foram avaliados modelos como Wav2Vec 2.0, HuBERT, WavLM e Whisper, com destaque para o desempenho consistente do Whisper em múltiplos conjuntos de dados de NVVs. O trabalho explorou técnicas de ajuste fino eficiente em parâmetros (*Parameter-Efficient Fine-Tuning* – PEFT), como LoRA, Adapters e Prompt Tuning, sendo que o LoRA demonstrou os melhores resultados em geral. A análise camada a camada revelou que a informação não-verbal está majoritariamente codificada nas camadas finais do Whisper, e que adaptar camadas menos informativas inicialmente pode resultar em melhorias mais significativas. Esses achados oferecem contribuições

relevantes sobre a utilização de modelos de fala pré-treinados para a classificação de sinais vocais não-verbais, especialmente em contextos de acessibilidade e interação afetiva homem-máquina.

Para resolver as deficiências dos modelos de fundação de fala e áudio no processamento de sons humanos não-verbais, Koudounas et al. (2025) propuseram o voc2vec, um modelo de fundação projetado especificamente para vocalizações não-verbais. Baseado na arquitetura wav2vec 2.0, o modelo foi pré-treinado usando aprendizagem autossupervisionada em uma coleção de 10 *datasets* de áudio não-verbal de código aberto, totalizando aproximadamente 125 horas. A estratégia mais eficaz consistiu em inicializar o modelo com pesos pré-treinados em fala (LibriSpeech) e, em seguida, continuar o treinamento no corpus de vocalizações. Nos experimentos, o voc2vec superou consistentemente os modelos de fundação de fala e áudio existentes, bem como *baselines* robustos como OpenSmile e emotion2vec, em seis tarefas de classificação de vocalizações. O estudo conclui que a pré-especialização em um corpus de dados de domínio específico é crucial para capturar as nuances das vocalizações não-verbais, apresentando o voc2vec como o primeiro modelo de representação universal para esta finalidade.

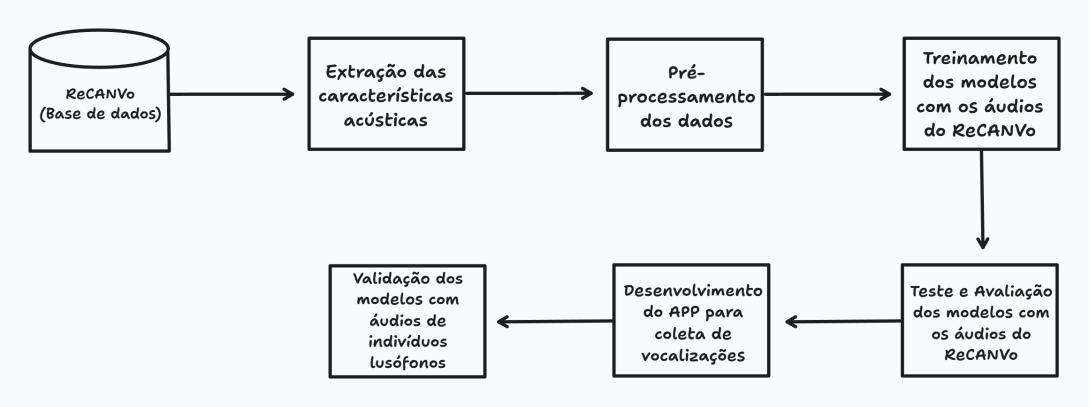
4 METODOLOGIA

A condução deste estudo segue um processo estruturado, alinhado às práticas consolidadas em projetos de aprendizado de máquina, que tradicionalmente envolvem o detalhamento do problema, a obtenção e preparação dos dados, a exploração e modelagem e a avaliação dos resultados, conforme descrito por Domingos (2012). O projeto se define como uma **pesquisa aplicada**, pois seu propósito central é gerar conhecimento para a solução de um problema prático e imediato: a classificação automatizada de vocalizações não verbais de indivíduos com TEA para o desenvolvimento de tecnologias assistivas, conforme sugere Gil (2002). A abordagem adotada é eminentemente **quantitativa**, uma vez que se baseia na coleta de dados numéricos (características acústicas) e na aplicação de métricas estatísticas (acurácia, precisão, *recall*, F1-score) para avaliar objetivamente o desempenho dos modelos de classificação e testar as hipóteses formuladas. Esta abordagem, segundo Creswell (2014), é ideal para examinar a relação entre variáveis de forma empírica.

Quanto aos objetivos, a pesquisa possui um caráter **exploratório**, principalmente em sua fase inicial. O propósito é investigar um fenômeno pouco explorado no contexto linguístico do português brasileiro, a aplicabilidade de modelos treinados com a base ReCANVo. Essa exploração inicial visa aprofundar a familiaridade com o problema e construir hipóteses para investigações futuras, em linha com a definição de Marconi e Lakatos (2003). Por fim, os procedimentos adotados conferem à pesquisa uma natureza **experimental**. Há uma manipulação controlada de variáveis, como a escolha de diferentes algoritmos de aprendizado de máquina, a otimização de seus hiperparâmetros e a aplicação de técnicas de pré-processamento, para observar e analisar os efeitos dessas manipulações nos resultados da classificação. Essa busca por relações de causa e efeito, mesmo utilizando um conjunto de dados secundário, é uma característica central da pesquisa experimental (BARBETTA; REIS; BORNIA, 2008).

O fluxo de trabalho detalhado, desde a aquisição dos dados até a avaliação dos modelos e o desenvolvimento do aplicativo, é apresentado visualmente no fluxograma da Figura 1.

Figura 1 – Fluxograma detalhando a metodologia do trabalho.



Fonte: Elaborada pelos autores (2025)

4.1 Fonte de Dados Inicial: O Banco de Dados ReCANVo

O ponto de partida para a avaliação dos modelos de classificação é o banco de dados ReCANVo (JOHNSON; NARAIN et al., 2023). Este conjunto de dados público é atualmente o mais abrangente para VNVs de indivíduos com necessidades complexas de comunicação. Ele contém mais de 7.000 vocalizações coletadas em contextos naturais de 8 indivíduos com diagnósticos diversos, incluindo TEA, Paralisia Cerebral e síndromes genéticas raras. As gravações foram realizadas com gravadores portáteis (Sony PCM-M10 e Zoom H4n Pro), e os rótulos foram atribuídos em tempo real por cuidadores e pesquisadores familiarizados com os indivíduos, garantindo precisão nos rótulos. As seis categorias de rótulos utilizadas, *delight, dysregulated, frustrated, request, selftalk, social*, fornecem uma base funcional para a classificação das intenções comunicativas.

4.2 Extração de Características e Pré-processamento

O processamento dos sinais de áudio foi implementado em linguagem Python, utilizando a biblioteca `Librosa` (versão 0.10). Seguindo a fundamentação teórica apresentada na Seção 2.5, o *pipeline* de extração iniciou-se com a padronização da taxa de amostragem para 16 kHz e conversão para canal mono.

Para a construção dos vetores de características, foram extraídos os atributos espectrais e temporais previamente definidos: MFCCs (40 coeficientes), Mel Espectrogramas, Cromaticidade, Contraste Espectral, ZCR, *Roll-off* e Centróide Espectral. De cada uma dessas representações, foram calculadas a média e o desvio padrão ao longo do tempo, resultando em um vetor de características de tamanho fixo para cada arquivo de áudio, independentemente de sua duração original.

Posteriormente, aplicou-se a normalização dos dados através do `StandardScaler` da biblioteca `scikit-learn`, ajustando as características para média zero e desvio padrão unitário. O tratamento do desbalanceamento de classes, discutido na Seção 2.7, foi realizado estritamente no conjunto de treinamento de cada passo da validação cruzada. Utilizou-se o algoritmo SMOTE com a estratégia de reamostragem automática para igualar a frequência das classes minoritárias à da classe majoritária, garantindo que o conjunto de teste permanecesse inalterado e livre de vazamento de dados (*data leakage*).

Para fins de reprodutibilidade e transparência, o código-fonte completo dos experimentos encontra-se disponível no notebook do projeto (Google Colab).

4.3 Configuração e Treinamento dos Modelos

Foram treinados e avaliados os quatro classificadores detalhados no Referencial Teórico: SVC, Random Forest, XGBoost e LightGBM. A implementação utilizou as bibliotecas `scikit-learn`, `xgboost` e `lightgbm`, respectivamente.

Para garantir a robustez dos resultados e a seleção ideal dos modelos, adotou-se uma estratégia de busca em grade (`GridSearchCV`) aninhada a uma validação cruzada estratificada de 5 pontas ($k = 5$). O espaço de busca de hiperparâmetros explorou:

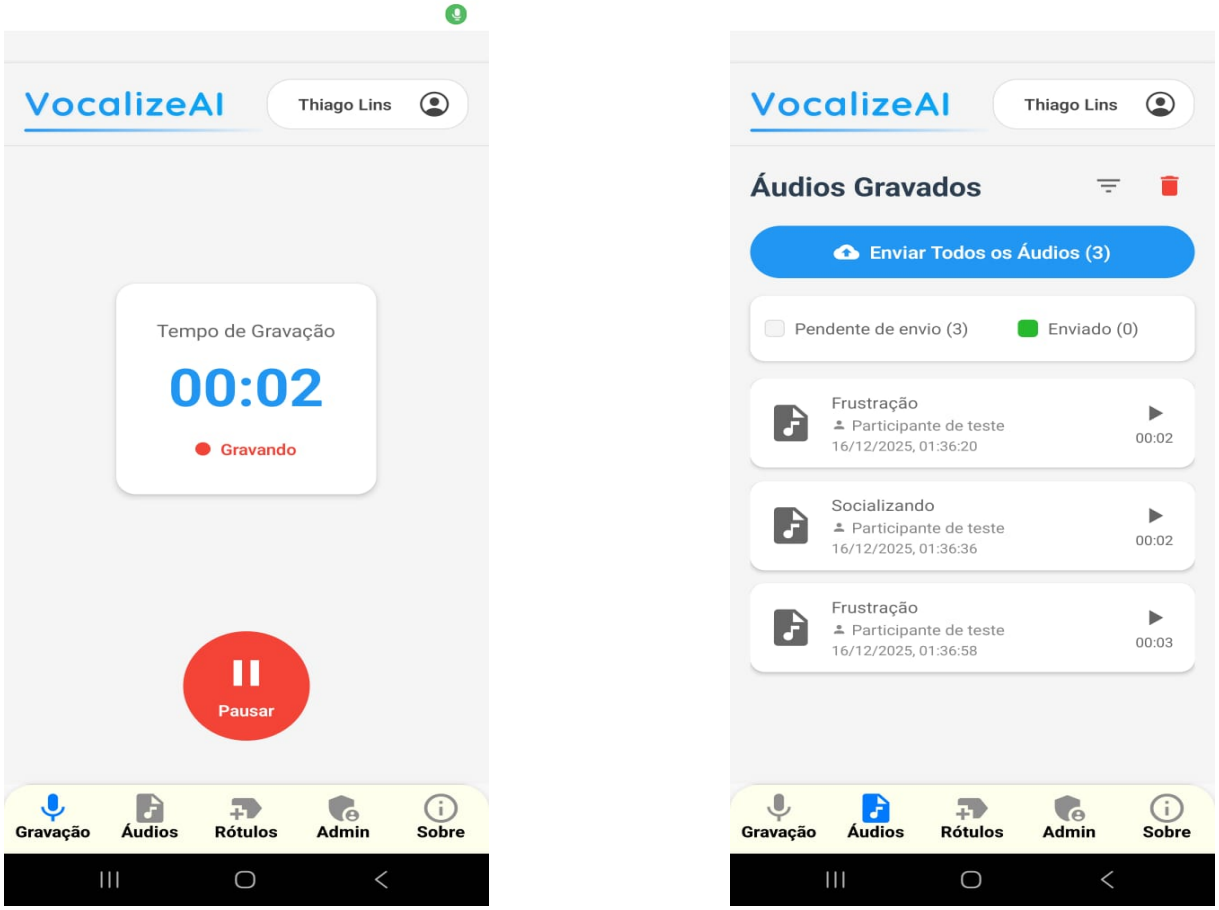
- **SVC:** Variações nos *kernels* (RBF, linear), parâmetro de regularização C e coeficiente do kernel γ.
- **Random Forest:** Número de estimadores (100 a 500) e profundidade máxima das árvores.
- **XGBoost e LightGBM:** Taxas de aprendizado, número de estimadores e profundidade máxima.

O desempenho final foi mensurado através das métricas de Acurácia, Precisão, *Recall* e F1-Score (macro), permitindo a comparação direta da capacidade de generalização de cada algoritmo frente ao desbalanceamento original dos dados.

4.4 Ferramenta de Coleta: Aplicativo VocalizeAI

Conforme ilustrado na etapa final do fluxograma metodológico (Figura 1), foi desenvolvido um aplicativo móvel para operacionalizar a coleta de dados no contexto brasileiro. A ferramenta, denominada *VocalizeAI*, foi projetada para permitir que pais e cuidadores gravem vocalizações em ambiente natural e atribuam rótulos em tempo real, seguindo as mesmas seis categorias funcionais do *dataset* ReCANVo.

O aplicativo atua como uma ponte para a validação dos modelos, garantindo que os áudios sejam capturados com taxa de amostragem consistente e metadados organizados. A interface foi desenhada para ser intuitiva e permitir o registro rápido da intenção comunicativa, conforme demonstrado na Figura 2.



(a) Tela de gravação

(b) Tela dos áudios gravados

Figura 2 – Interface do aplicativo VocalizeAI desenvolvido para coleta de dados.

Fonte: Elaborada pelos autores (2025)

5 EXPERIMENTOS E ANÁLISE DE RESULTADOS

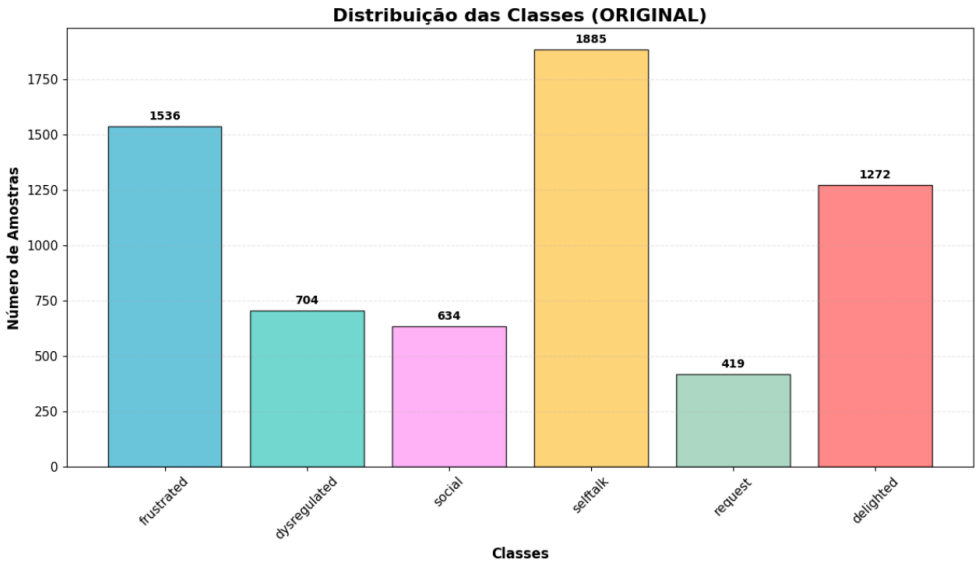
Para responder à questão de pesquisa formulada e avaliar a robustez dos classificadores, o estudo foi estruturado em duas etapas experimentais principais. A primeira consistiu no treinamento e na avaliação de quatro modelos de *machine learning* , utilizando exclusivamente o *dataset* ReCANVo. A segunda etapa envolveu a aplicação desses modelos, já treinados, na classificação de um novo conjunto de vocalizações coletadas de um indivíduo lusófono, a fim de testar a capacidade de generalização dos modelos para um contexto linguístico e cultural distinto.

5.1 Treinamento e Validação no Dataset ReCANVo

Nesta fase inicial, os modelos Support Vector Machine (SVC), Random Forest, XGBoost e LightGBM (LGBM) foram treinados com os dados do ReCANVo. O particionamento dos dados foi realizado utilizando a estratégia de *train-test split* (80% para treinamento e 20% para teste), garantindo a avaliação em amostras não vistas durante o ajuste dos modelos.

Uma análise exploratória do *dataset* revelou um desbalanceamento significativo entre as seis classes de vocalizações, conforme detalhado: *delighted* (1272 amostras), *dysregulated* (704), *frustrated* (1536), *request* (419), *selftalk* (1885) e *social* (634). Para mitigar o risco de que os modelos se tornassem enviesados em favor das classes majoritárias (*selftalk* e *frustrated*), foi aplicada a técnica de sobreamostragem SMOTE (*Synthetic Minority Over-sampling Technique*) durante a etapa de treinamento. Essa abordagem equilibra a distribuição das classes gerando amostras sintéticas para as categorias minoritárias, como *request* e *social*, garantindo que todas as classes tivessem representatividade equivalente no processo de aprendizado.

Figura 3 – Gráfico com a distribuição das classes antes do SMOTE.



Fonte: Elaborada pelos autores (2025)

Após o treinamento com os dados balanceados e a otimização de hiperparâmetros, os modelos foram avaliados em um conjunto de teste separado. A Tabela 1 apresenta um comparativo do desempenho dos quatro algoritmos, com as métricas de precisão, *recall* e F1-Score para cada uma das seis classes. Além disso, foi calculada a média macro dos escores, permitindo uma visão mais justa em cenários desbalanceados. Os resultados mostram que os modelos baseados em *Gradient Boosting* (XGBoost e LGBM) e o SVC obtiveram o desempenho mais robusto e equilibrado entre as classes, com destaque para a classificação de *dysregulated* e *frustrated*. A classe *request*, por ser a mais minoritária originalmente, apresentou os maiores desafios de classificação para todos os modelos, mesmo após a aplicação do SMOTE.

5.2 Validação com Vocalizações de Indivíduo Lusófono

Para a etapa de validação externa, foram coletadas 53 vocalizações de um indivíduo lusófono utilizando o aplicativo *VocalizeAI*. O uso da ferramenta garantiu a padronização inicial da taxa de amostragem e facilitou a associação correta dos rótulos

Tabela 1 – Comparação de Desempenho dos modelos utilizando a base do ReCANVo.

Classes	SVC			Random Forest			XGBoost			LGBM		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Delighted	0.56	0.58	0.57	0.55	0.53	0.54	0.63	0.61	0.62	0.61	0.59	0.60
Dysregulated	0.75	0.67	0.70	0.73	0.67	0.70	0.83	0.68	0.75	0.81	0.67	0.73
Frustrated	0.67	0.73	0.70	0.69	0.78	0.73	0.74	0.80	0.77	0.75	0.80	0.77
Request	0.52	0.42	0.46	0.56	0.40	0.47	0.62	0.43	0.51	0.58	0.39	0.47
Selftalk	0.65	0.68	0.68	0.66	0.68	0.67	0.69	0.78	0.73	0.68	0.77	0.72
Social	0.64	0.51	0.57	0.61	0.56	0.58	0.66	0.57	0.61	0.66	0.57	0.61
Macro Average	0.63	0.60	0.61	0.63	0.60	0.62	0.69	0.65	0.66	0.68	0.63	0.65

Fonte: Elaborada pelos autores (2025)

pelo cuidador no momento da ocorrência da vocalização. Essas amostras foram então submetidas aos quatro modelos previamente treinados com o *dataset* ReCANVo, para avaliar seu desempenho preditivo em dados fora do domínio original. A Tabela 2 resume a acurácia geral de cada modelo nesta tarefa.

Tabela 2 – Resultados da classificação das 53 vocalizações lusófonas coletadas.

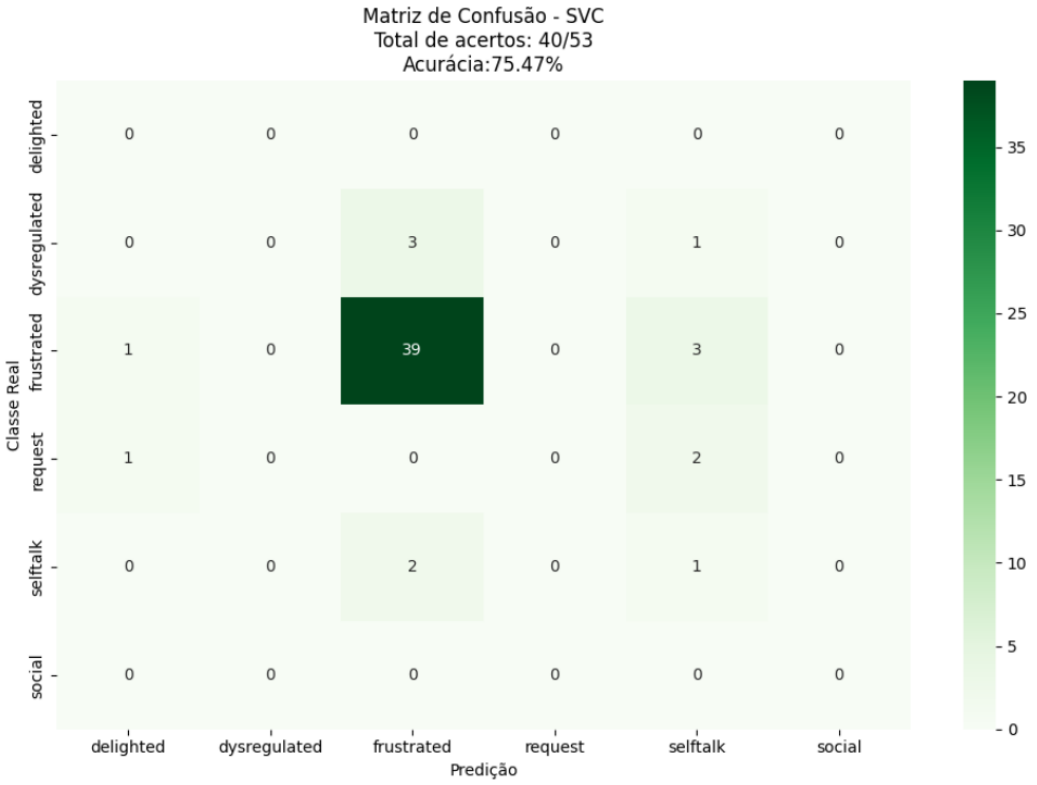
Modelo	Acurácia (%)	Predições Corretas
SVC	75.47	40 / 53
LGBM	60.38	32 / 53
XGBoost	56.66	30 / 53
Random Forest (RF)	45.28	24 / 53

Fonte: Elaborada pelos autores (2025)

Apesar da acurácia de 75,47% obtida pelo modelo SVC, uma análise aprofundada dos resultados revela que esse valor é enganoso e mascara uma falha severa de generalização. A distribuição das amostras era de 43 para *frustrated*, 4 para *dysregulated*, 3 para *request* e 3 para *selftalk*. O modelo errou a totalidade das 4 amostras de *dysregulated* e das 3 amostras de *request*, acertando apenas 1 de 3 em *selftalk*. Este resultado demonstra que o modelo não aprendeu a distinguir as categorias, mas sim a enviesar sua resposta para a classe majoritária da amostra. Tal comportamento é compreensível, dado que o conjunto de validação estava extremamente desbalanceado e o indivíduo não compartilhava o mesmo idioma dos dados de treinamento, o que introduz variações acústicas não capturadas pelo modelo original.

A matriz de confusão para o teste do modelo SVC, ilustrada na Figura 4, evidencia que o bom desempenho geral é quase inteiramente atribuível à sua capacidade de identificar a classe *frustrated*, que compunha a grande maioria do conjunto de teste (81%). O modelo falhou em generalizar para as outras categorias. Esse resultado demonstra que a métrica de acurácia, sozinha, é insuficiente para avaliar modelos em cenários desbalanceados, reforçando a necessidade de métricas como *balanced accuracy* e macro-F1 e, principalmente, de dados locais para treinamento.

Figura 4 – Matriz de confusão do modelo SVC na classificação das 53 vocalizações coletadas.



Fonte: Elaborada pelos autores (2025)

6 CONCLUSÃO

Os resultados experimentais obtidos nesta pesquisa demonstram de forma contundente que a eficácia dos modelos de aprendizado de máquina para classificação de vocalizações não verbais de indivíduos com TEA é altamente dependente do contexto cultural e linguístico dos dados de treinamento. Em resposta à questão de pesquisa formulada, conclui-se que os modelos treinados exclusivamente com a base americana ReCANVo não possuem capacidade de generalização suficiente para o contexto lusófono.

Embora o modelo SVC tenha alcançado uma acurácia de 75,47%, a análise detalhada revelou que este desempenho foi enganoso. A performance foi impulsionada pela classe super-representada na amostra, enquanto o modelo falhou completamente em classificar as amostras das classes minoritárias. Esta falha em generalizar evidencia a limitação crítica do dataset ReCANVo para aplicações no contexto brasileiro, corroborando a premissa de que modelos de classificação de VNVs só apresentam desempenho robusto quando treinados com dados cultural e linguisticamente alinhados.

Portanto, conclui-se que, para o desenvolvimento de tecnologias assistivas eficazes e confiáveis, é imprescindível a criação de bases de dados locais. A aplicação móvel *VocalizeAI*, desenvolvida no âmbito deste trabalho, representa um passo fundamental nessa direção. O próximo passo é focar na coleta de mais dados para enriquecer a base lusófona com crianças brasileiras, visando corrigir o desbalanceamento observado e treinar modelos verdadeiramente representativos. Este esforço não apenas permitirá o aprimoramento dos modelos de classificação, mas também abrirá caminhos para pesquisas futuras e para o desenvolvimento de ferramentas que verdadeiramente auxiliem na comunicação e na qualidade de vida de indivíduos com TEA.

6.1 Limitações da Pesquisa

O presente estudo, apesar de seus achados, possui limitações que devem ser consideradas e que abrem caminhos para futuras investigações:

- **Amostragem de Validação Reduzida:** O conjunto de dados coletado para validação consistiu em apenas 53 vocalizações, um número insuficiente para tirar conclusões estatisticamente robustas sobre o desempenho geral dos modelos no contexto lusófono.
- **Fonte de Dados Única e Viés na Coleta:** As 53 vocalizações foram coletadas de um único indivíduo, o que impede a generalização dos resultados para a população mais ampla de pessoas com TEA no Brasil. Adicionalmente, por se tratar de uma fase inicial de coleta, não houve controle rigoroso sobre o balanceamento das classes, resultando em uma amostra enviesada (81% *frustrated*) que inflou artificialmente a métrica de acurácia.
- **Dependência de um Único Dataset de Referência:** A pesquisa se baseia no ReCANVo, o único *dataset* público de grande escala disponível atualmente. A ausência de outros bancos de dados para comparação ou enriquecimento do treinamento limita a diversidade de cenários acústicos e culturais que podem ser modelados.

- **Limitação Metodológica:** Não foi realizada validação estatística formal (por exemplo, testes de significância) para a comparação entre modelos no cenário lusófono, devido ao baixo número de amostras coletadas.

6.2 Trabalhos Futuros

Com base nos resultados e nas limitações identificadas, a principal direção para trabalhos futuros é a expansão da coleta de dados por meio da distribuição ampla do aplicativo *VocalizeAI*. O objetivo é construir um *dataset* robusto e representativo de vocalizações não verbais de indivíduos com TEA no Brasil, mitigando o viés observado nos modelos treinados exclusivamente com dados norte-americanos.

Além disso, destacam-se as seguintes direções:

- Aplicação de técnicas de **aprendizado por transferência (*transfer learning*)**, aproveitando representações acústicas robustas extraídas de modelos pré-treinados, como Wav2Vec 2.0 ou Whisper.
- Exploração de **técnicas de aumento de dados** (*data augmentation*), como variações de *pitch* e *time stretching*, para mitigar a escassez inicial de dados da base brasileira.
- Utilização dos modelos já treinados como ferramentas de **pré-rotulagem semi-supervisionada**, acelerando o processo de construção da nova base de dados.

REFERÊNCIAS

AMERICAN PSYCHIATRIC ASSOCIATION. **Diagnostic and statistical manual of mental disorders (DSM-5)**. [S.l.]: American Psychiatric Publishing, 2013.

ARLOT, Sylvain; CELISSE, Alain. A survey of cross-validation procedures for model selection. **Statistics Surveys**, Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., e the Statist. Soc. Canada, v. 4, none, p. 40–79, 2010. DOI: 10.1214/09-SS054. Disponível em: <https://doi.org/10.1214/09-SS054>.

BARBETTA, P. A.; REIS, M. M.; BORNIA, A. C. Estatística para cursos de engenharia e informática. Atlas, 2008.

BERSCH, Rita. Introdução à tecnologia assistiva. **Porto Alegre: CEDI**, v. 21, p. 1–20, 2008.

BEUKELMAN, David R.; MIRENDA, Pat. **Augmentative and Alternative Communication: Supporting Children & Adults with Complex Communication Needs**. 4. ed. Baltimore: Paul H. Brookes Publishing, 2013.

BISCHL, Bernd et al. Hyperparameter Optimization: foundations, algorithms, best practices and open challenges. **arXiv (Cornell University)**, jan. 2021. DOI: 10.48550/arxiv.2107.05847. Disponível em: <https://arxiv.org/abs/2107.05847>.

BREIMAN, Leo. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, jan. 2001. DOI: 10.1023/a:1010933404324. Disponível em: <https://doi.org/10.1023/a:1010933404324>.

BURGES, Christopher J.C. A tutorial on support vector machines for pattern recognition. **Data Mining and Knowledge Discovery**, v. 2, n. 2, p. 121–167, jan. 1998. DOI: 10.1023/a:1009715923555. Disponível em: <https://doi.org/10.1023/a:1009715923555>.

CHAWLA, Nitesh V. et al. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002. DOI: 10.1613/jair.953.

CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. In: PROCEEDINGS of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA: Association for Computing Machinery, 2016. (KDD '16), p. 785–794. ISBN 9781450342322. DOI: 10.1145/2939672.2939785. Disponível em: <https://doi.org/10.1145/2939672.2939785>.

COOK, Albert M; POLGAR, Janice Miller. **Assistive technologies-e-book: principles and practice**. [S.l.]: Elsevier Health Sciences, 2014.

CRESWELL, J. W. **Research design: Qualitative, quantitative, and mixed methods approaches**. [S.l.]: Sage publications, 2014.

DELL, A.G.; NEWTON, D.A.; PETROFF, J.G. **Assistive Technology in the Classroom: Enhancing the School Experiences of Students with Disabilities**. [S.l.]: Pearson Merrill Prentice Hall, 2008. ISBN 9780131191648. Disponível em: <https://books.google.com.br/books?id=Fly1OwAACAAJ>.

DI AUTOMATICA E INFORMATICA, Dipartimento et al. **Exploring the Adaptability of Large Speech Models to Non-Verbal Vocalization task**. [S.l.: s.n.], jan. 2025. Disponível em: <https://hdl.handle.net/11583/3002059>.

DOMINGOS, Pedro. A few useful things to know about machine learning. **Communications of the ACM**, v. 55, n. 10, p. 78–87, 25 set. 2012. DOI: 10.1145/2347736.2347755. Disponível em: <https://doi.org/10.1145/2347736.2347755>.

FAWCETT, Tom. An Introduction to ROC Analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006. DOI: 10.1016/j.patrec.2005.10.010.

FENG, Tiantian et al. **Egocentric Speaker Classification in Child-Adult Dyadic Interactions: From Sensing to Computational Modeling**. [S.l.: s.n.], 2025. arXiv: 2409.09340 [cs.SD]. Disponível em: <https://arxiv.org/abs/2409.09340>.

FRIEDMAN, Jerome H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, JSTOR, p. 1189–1232, 2001.

GIL, A.C. **Como elaborar projetos de pesquisa**. [S.l.]: Atlas, 2002. ISBN 9788522431694. Disponível em: <https://books.google.com.br/books?id=X4uvAAAACAAJ>.

GOUYON, Fabien; PACHET, Francois; DELERUE, Olivier. On the Use of Zero-Crossing Rate for an Application of Classification of Percussive Sounds, ago. 2002.

JOHNSON, Kristina T.; NARAIN, Jaya et al. ReCANVo: A database of real-world communicative and affective nonverbal vocalizations. **Scientific Data**, Springer Nature, v. 10, p. 523, 2023.

JOHNSON, Kristina T.; O'BRIEN, Amanda M. et al. Affective Ratings of Nonverbal Vocalizations Produced by Minimally-Speaking Individuals: What Do Naive Listeners Perceive? In: IEEE. 10TH International Conference on Affective Computing and Intelligent Interaction (ACII). [S.l.: s.n.], 2022. P. 1–8.

KE, Guolin et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In_____. **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>.

KONECKI, Mario; LOVRENČIĆ, Sandra; JERVIS, Keith. The use of assistive technology in education of programming. In: 2016 ICBTS International Academic Research Conference Proceedings, Boston, USA. [S.l.: s.n.], 2016. P. 1–11.

KOUDOUNAS, Alkis et al. **voc2vec: A Foundation Model for Non-Verbal Vocalization**. [S.l.: s.n.], fev. 2025. DOI: 10.48550/arXiv.2502.16298.

LAZAR, Jonathan; GOLDSTEIN, Daniel F; TAYLOR, Anne. **Ensuring digital accessibility through process and policy**. [S.l.]: Morgan kaufmann, 2015.

MARCONI, M. de A.; LAKATOS, E. M. **Fundamentos de metodologia científica**. [S.l.]: Atlas, 2003.

MCFEE, Brian et al. librosa: Audio and Music Signal Analysis in Python. **Proceedings of the Python in Science Conferences**, p. 18–24, jan. 2015. DOI: 10.25080/majora-7b98e3ed-003. Disponível em: <https://doi.org/10.25080/majora-7b98e3ed-003>.

NARAIN, Jaya et al. Nonverbal Vocalizations as Speech: Characterizing Natural-Environment Audio from Nonverbal Individuals with Autism. **Laughter Workshop Proceedings**, MIT Media Lab, 2020.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

SHAW, Kelly A. et al. Prevalence and early identification of autism spectrum disorder among children aged 4 and 8 years — Autism and Developmental Disabilities Monitoring Network, 16 sites, United States, 2022. **MMWR Surveillance Summaries**, v. 74, n. 2, p. 1–22, abr. 2025. DOI: 10.15585/mmwr.ss7402a1. Disponível em: <https://doi.org/10.15585/mmwr.ss7402a1>.

TRAN, van-Thuan; TSAI, Wei-Ho. Identification of Non-Speaking and Minimal-Speaking Individuals Using Nonverbal Vocalizations. **IEEE Access**, v. 12, p. 68954–68967, 2024. DOI: 10.1109/ACCESS.2024.3398584.