UNIVERSIDADE FEDERAL DE MATO GROSSO
CAMPUS UNIVERSITÁRIO DO ARAGUAIA
INSTITUTO DE CIÊNCIAS EXATAS E DA TERRA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

DETECÇÃO DE DEPRESSÃO EM GESTANTES ATRAVÉS DA VOZ COM DEEP LEARNING

JULIANO BACKES

BARRA DO GARÇAS – MT
2025

UNIVERSIDADE FEDERAL DE MATO GROSSO
CAMPUS UNIVERSITÁRIO DO ARAGUAIA
INSTITUTO DE CIÊNCIAS EXATAS E DA TERRA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

JULIANO BACKES

Monografia apresentada ao Instituto de Ciências Exatas e da Terra do Campus Universitário do Araguaia, da Universidade Federal de Mato Grosso, como requisito parcial para conclusão do Curso de Graduação em Ciência da Computação.

Orientador: Prof. Dr. Rafael Teixeira Sousa

CAMPUS ARAGUAIA
MAIO/2025

UNIVERSIDADE FEDERAL DE MATO GROSSO
CAMPUS UNIVERSITÁRIO DO ARAGUAIA
INSTITUTO DE CIÊNCIAS EXATAS E DA TERRA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

DETECÇÃO DE DEPRESSÃO EM GESTANTES ATRAVÉS DA VOZ COM DEEP LEARNING

Esta monografia foi avaliada e aprovada pela banca examinadora instituída pelo Colegiado do Curso de Bacharelado em Ciência da Computação do Campus Universitário do Araguaia, da Universidade Federal de Mato Grosso.

_____
Prof. Dr. Rafael Teixeira Sousa
ICET/UFMT
(Orientador)

_____
Prof. Dr. Anthony Ferreira La Marca
ICET/UFMT
(Membro)

_____
Prof. Dr. Thiago Pereira da Silva
ICET/UFMT
(Membro)

CAMPUS ARAGUAIA
MAIO/2025

AGRADECIMENTOS

Primeiramente agradeço os meus parentes, em especial aos meus pais Jacó Selvino Backes e Marilene Inês Backes que tornaram tudo isso possível, que não mediram esforços para me garantir uma boa educação. O apoio motivacional e emocional durante toda a minha jornada e em especial durante a execução desse trabalho foi imprescindível.

Também agradeço aos meus colegas de graduação, em especial ao Aristides Lemos Dourado, que sempre forneceu apoio e companheirismo.

Agradeço também à universidade como um todo, que me proporcionou vivenciar a experiência acadêmica, bem como a todos os ensinamentos transmitidos pelo corpo docente do curso no qual faço parte. Em especial, agradeço ao meu orientador Dr. Rafael Teixeira Sousa pela sua orientação, ensinamentos, paciência e disponibilidade.

Quero agradecer também ao Centro de Excelência em Inteligência Artificial (CEIA) por me proporcionar a experiência de iniciação científica.

Por fim, agradeço a minha namorada, Glades Kellen Machado, minha companheira de todas as horas, por todo o amor, paciência e apoio incondicional durante essa jornada. Sua compreensão nos momentos de ausência, incentivo nos momentos de dúvida e presença constante mesmo nos silêncios fizeram toda a diferença para que eu pudesse concluir mais essa etapa da minha vida.

RESUMO

A depressão é um transtorno mental que representa um sério desafio para a saúde pública, especialmente entre gestantes, com prevalência estimada entre 10% e 20% durante a gravidez. Nesse contexto, a detecção precoce é essencial para garantir intervenções adequadas. Este trabalho tem como objetivo investigar a viabilidade do uso de redes neurais para identificar sintomas depressivos em gestantes, por meio da análise da voz captada por smartphones. Foram utilizados 116 áudios de 93 pacientes, coletados em ambientes não controlados. Os áudios passaram por pré-processamento, segmentação e conversão em espectrogramas, os quais alimentaram uma rede neural convolucional. Para lidar com o pequeno volume de dados, foi aplicada a técnica de transferência de aprendizado, utilizando um modelo pré-treinado. A avaliação foi realizada com validação cruzada, utilizando a curva ROC e a AUC como métricas de desempenho. O modelo obteve uma AUC média de 0,63. Apesar de ainda insuficientes para aplicação clínica, os resultados indicam o potencial da abordagem para a triagem inicial de depressão perinatal com o uso de dispositivos acessíveis, como smartphones. Espera-se que, com mais dados e refinamento do modelo, a acurácia diagnóstica possa ser significativamente ampliada.

Palavras-chave: Detecção de depressão. Gestantes. Rede neural. Análise de voz.

ABSTRACT

Depression is a mental disorder that poses a serious public health challenge, especially among pregnant women, with an estimated prevalence of 10% to 20% during pregnancy. In this context, early detection is essential to ensure appropriate interventions. This study aims to investigate the feasibility of using neural networks to identify depressive symptoms in pregnant women through voice analysis captured by smartphones. A total of 116 audio recordings from 93 patients were used, collected in uncontrolled environments. The audio data underwent preprocessing, segmentation, and conversion into spectrograms, which were then fed into a convolutional neural network. To address the small dataset size, a transfer learning technique was applied using a pre-trained model. Evaluation was performed using cross-validation, with the ROC curve and AUC as performance metrics. The model achieved an average AUC of 0.63. Although still insufficient for clinical application, the results indicate the potential of this approach for initial screening of perinatal depression using accessible devices such as smartphones. It is expected that with more data and further model refinement, diagnostic accuracy can be significantly improved.

Keywords: Depression detection. Pregnant women. Neural network. Voice analysis.

SUMÁRIO

AGRADECIMENTOS.....3
RESUMO.....4
ABSTRACT.....5
SUMÁRIO.....6
1 INTRODUÇÃO.....7
2 FUNDAMENTAÇÃO TEÓRICA.....10
2.1 DEPRESSÃO EM GESTANTES.....10
2.2 ANÁLISE DE VOZ NA DETECÇÃO DE ESTADOS EMOCIONAIS.....10
2.3 ESPECTROGRAMAS.....12
2.3 REDE NEURAL.....13
3 METODOLOGIA.....16
3.1 CONJUNTO DE DADOS.....16
3.2 PRÉ-PROCESSAMENTO DOS DADOS.....19
3.3 DESENVOLVIMENTO DO MODELO.....21
4 RESULTADOS.....22
5 CONCLUSÃO.....26
REFERÊNCIAS BIBLIOGRÁFICAS.....28

1. INTRODUÇÃO

A gravidez é frequentemente vista como uma experiência gratificante e transformadora, marcada por intensas mudanças físicas, emocionais e sociais. No entanto, para certas mulheres, isso pode gerar estigma social e estresse, podendo originar desafios de saúde mental, como depressão e ansiedade (Waqas et al.,2022).

Os transtornos depressivos constituem um grave problema de saúde pública devido à sua alta prevalência, repercussões na saúde geral e impacto psicossocial. A depressão se apresenta como uma das principais causas de incapacitação no mundo, limitando o funcionamento físico, pessoal e social(Gonçalves et al., 2018).

Um dos problemas mais preocupantes que podem afetar mulheres grávidas é a depressão, uma condição mental que pode ter implicações importantes tanto para a mãe quanto para o desenvolvimento do bebê (Fernandes, 2013). Identificar a depressão em gestantes é essencial para fornecer o apoio necessário às mulheres que estão passando por essa situação.

Nas entrevistas psiquiátricas, a avaliação da voz e da fala dos pacientes serve como uma ferramenta importante para os médicos analisarem sintomas psicológicos. A voz fornece indícios sobre o funcionamento das cordas vocais, enquanto a fala reflete aspectos como a velocidade e a hesitação na comunicação verbal. É comum que, em casos de depressão, haja uma redução no ritmo da fala, com pausas mais frequentes durante as elocuções. Com o avanço das tecnologias computacionais, essas características vocais passaram a ser mensuradas de forma quantitativa, permitindo uma análise mais precisa da relação entre a voz e transtornos depressivos(Shin et al., 2023).

Estudos anteriores já haviam demonstrado que variáveis como a frequência fundamental (F0), que mede a dinâmica e a intensidade da voz, estão associadas à depressão. Outros parâmetros vocais, como o jitter vocal, o espectro do fluxo glótico e os coeficientes de cepstrum de frequência mel(MFCCs), também foram correlacionados à gravidade dos sintomas depressivos. Esses achados indicam que a voz tem potencial para atuar como um biomarcador eficaz para a depressão,

oferecendo uma maneira objetiva de avaliar a condição dos pacientes (Shin et al., 2023).

Com o crescente uso da inteligência artificial no campo médico, novas pesquisas têm explorado a possibilidade de prever a depressão com base nas características vocais. Por meio de algoritmos de aprendizado de máquina, tem sido possível identificar padrões na fala que podem ser utilizados para diagnosticar a depressão de forma precoce e precisa, abrindo caminhos para o desenvolvimento de novas ferramentas de suporte ao diagnóstico psiquiátrico (Shin et al., 2023).

Diante da dificuldade de acesso a diagnósticos especializados e da subjetividade das entrevistas clínicas, surge a necessidade de desenvolver ferramentas automatizadas que apoiem a detecção da depressão perinatal por meio de sinais vocais. A utilização de áudios gravados em smartphones como fonte de dados para essa detecção oferece uma vantagem significativa, uma vez que esses dispositivos são amplamente acessíveis e utilizados por muitas mulheres grávidas. Assim, a abordagem baseada em redes neurais torna-se uma ferramenta acessível e de fácil implementação para o rastreamento da saúde mental durante a gestação.

Redes neurais são modelos computacionais inspirados no funcionamento do cérebro humano, projetados para aprender e resolver tarefas complexas por meio de técnicas de aprendizado de máquina. Elas têm sido amplamente empregadas em diversas áreas, como reconhecimento de padrões, processamento de linguagem natural e visão computacional (CHAVES, 2023).

Esses modelos podem ser treinados para analisar características específicas da voz, como tom, ritmo, entonação e padrões de fala, permitindo a detecção de possíveis sinais de depressão. Para utilizar os áudios como entrada para modelos de aprendizado de máquina, é necessário realizar um pré-processamento dos dados brutos, incluindo a segmentação do áudio, a normalização dos dados em uma escala padrão e a extração de características relevantes (features).

Em um cenário com poucos dados, como é o nosso caso, o transfer learning (aprendizado por transferência) é uma técnica importante para evitar o sobreajuste (overfitting) comum quando se tenta treinar uma rede neural do zero com um conjunto de dados pequeno. Isso ocorre porque a rede pré-treinada já aprendeu a

generalizar informações de maneira eficaz, o que pode ajudar a evitar o ajuste excessivo aos dados de treinamento limitados. Transfer learning é uma abordagem fundamental no campo de redes neurais, que permite aproveitar o conhecimento prévio adquirido por uma rede neural treinada em uma tarefa para melhorar o desempenho em uma tarefa relacionada

Este trabalho tem como objetivo aplicar redes neurais profundas para identificar sinais de depressão em gestantes a partir da análise de características vocais extraídas de áudios gravados por smartphones. A identificação precoce da depressão durante a gestação é essencial para garantir o bem-estar da mãe e do bebê. No entanto, a escassez de métodos acessíveis e objetivos para essa detecção ainda representa um desafio. Nesse contexto, o uso de modelos de aprendizado de máquina aplicados à fala surge como uma alternativa promissora, especialmente em cenários com recursos limitados.

A detecção precoce e o tratamento adequado da depressão perinatal não apenas melhoram a qualidade de vida das gestantes, mas também têm implicações significativas para a saúde e o desenvolvimento da criança, tornando o uso de redes neurais profundas uma abordagem relevante e potencialmente transformadora no campo da saúde materna (HAHN et al., 2021).

A estrutura deste trabalho está organizada da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica, abordando os principais conceitos relacionados à depressão perinatal, à análise de sinais vocais como biomarcadores, às redes neurais convolucionais e ao uso de aprendizado por transferência. No Capítulo 3, são descritos os procedimentos metodológicos adotados, incluindo a coleta e o pré-processamento dos dados, a geração de espectrogramas e a aplicação da arquitetura ResNet-18. O Capítulo 4 expõe e discute os resultados obtidos a partir dos experimentos realizados, com base nas métricas de avaliação utilizadas. Por fim, o Capítulo 5 apresenta as considerações finais, destacando as contribuições do estudo, suas limitações e sugestões para trabalhos futuros.

2. FUNDAMENTAÇÃO TEÓRICA

2.1 Depressão em Gestantes

A depressão em gestantes é um transtorno mental caracterizado pela presença de sintomas depressivos durante a gravidez, sendo considerada uma condição de alta prevalência e impacto significativo na saúde pública. Estudos indicam que a prevalência da depressão gestacional varia entre 10% a 30%, com uma média aproximada de 20% em países em desenvolvimento, incluindo o Brasil, onde a incidência tende a ser maior no terceiro trimestre da gestação e em casos de gravidez de alto risco (PEREIRA; LOVISI, 2008). A depressão gestacional é frequentemente subdiagnosticada, apesar de sua relevância, e pode ser avaliada por meio de escalas específicas como a Edinburgh Postnatal Depression Scale (EPDS) (ALBERTINI, 2019).

As causas e fatores de risco para a depressão durante a gravidez são multifatoriais e envolvem aspectos biopsicossociais. Entre os principais fatores associados estão antecedentes psiquiátricos, dificuldades financeiras, baixa escolaridade, desemprego, falta de suporte social, uso de substâncias, violência doméstica, gravidez na adolescência e eventos estressores recentes (SÁNCHEZ et al., 2016). Esses fatores contribuem para o desenvolvimento ou agravamento dos sintomas depressivos, tornando essencial a identificação precoce para intervenções adequadas no pré-natal.

Os impactos da depressão em gestantes vão além do sofrimento materno, afetando também a saúde fetal e o desenvolvimento infantil. A depressão gestacional está associada a desfechos adversos como parto prematuro, baixo peso ao nascer e dificuldades no vínculo mãe-bebê, além de aumentar o risco de depressão pós-parto (MORAIS et al., 2018).

2.2 Análise de Voz na Detecção de Estados Emocionais

A voz humana constitui um importante canal de expressão emocional, refletindo estados internos muitas vezes não verbalizados de forma direta. Mudanças na entonação, ritmo, volume e qualidade vocal podem sinalizar alterações

emocionais significativas, como tristeza, ansiedade ou raiva. Em indivíduos com depressão, por exemplo, é comum observar uma voz monótona, sem variações expressivas, o que pode ser atribuído a alterações neurofisiológicas que afetam a coordenação motora e a precisão articulatória da fala (SCHERER, 2016; ESPY-WILSON, 2014). Essas características vocais são essenciais para a detecção precoce de transtornos emocionais, sobretudo em contextos clínicos.

Pesquisas recentes indicam que o reconhecimento de emoções por meio da voz pode ser mais eficaz do que por expressões faciais, especialmente em determinados grupos. Estudantes de medicina, por exemplo, demonstraram maior acurácia na identificação de emoções como raiva, medo e tristeza ao analisarem a prosódia vocal (SILVA et al., 2024). Além de transmitir emoções, a voz também influencia o estado emocional do próprio emissor. Experimentos mostraram que a manipulação digital da tonalidade vocal foi capaz de alterar o humor dos participantes (AUCOUTURIER et al., 2016). Essa relação bidirecional reforça a importância da voz como um indicador sensível das emoções, sugerindo seu potencial na criação de ferramentas diagnósticas baseadas em análise vocal.

No contexto da detecção de depressão em gestantes, o uso de técnicas de aprendizado profundo (deep learning) aplicadas à análise vocal representa um avanço significativo. Algoritmos de aprendizado profundo conseguem identificar padrões vocais sutis — muitas vezes imperceptíveis ao ouvido humano — que indicam sintomas depressivos, como menor variação tonal, pausas frequentes e redução do volume (BENTLEY, 2024). Dessa forma, a voz se torna uma fonte valiosa de dados para o monitoramento da saúde mental, possibilitando intervenções precoces que podem beneficiar tanto a mãe quanto o bebê. A integração entre análise vocal e inteligência artificial tem, portanto, grande potencial para contribuir com a saúde pública, especialmente entre populações vulneráveis como as gestantes.

A técnica de validação cruzada foi utilizada para avaliar a rede neural, sendo uma abordagem essencial em aprendizado de máquina para garantir a robustez e a precisão dos resultados. Essa técnica consiste em dividir o conjunto de dados em múltiplas partes, chamadas de folds, onde, em cada iteração, uma parte é usada

para teste e as outras para treinamento, de forma alternada. No caso específico deste trabalho, optou-se por uma divisão de 5 folds, permitindo que o modelo fosse testado cinco vezes, cada uma com um subconjunto diferente, assegurando assim uma avaliação mais confiável, especialmente em cenários com dados limitados. Além disso, a validação cruzada é crucial para evitar o sobreajuste, permitindo o ajuste eficiente dos hiperparâmetros e garantindo que o desempenho do modelo seja generalizável para dados novos.

2.3 Espectrogramas

O espectrograma é gerado a partir de um sinal de áudio através de um processo conhecido como Transformada de Fourier de Curto Prazo (STFT), que permite analisar como as frequências de um sinal variam ao longo do tempo. Inicialmente, o sinal de áudio contínuo é dividido em pequenos segmentos de tempo, chamados de janelas, que podem se sobrepor para uma melhor captura das variações no sinal. Em cada uma dessas janelas, aplica-se a Transformada Rápida de Fourier (FFT) para converter o sinal do domínio do tempo para o domínio da frequência. Esse processo resulta em um espectro de frequências para cada janela de tempo, e essas frequências são representadas em um gráfico, onde o eixo vertical indica a frequência, o eixo horizontal o tempo, e a intensidade das cores ou escalas de cinza representam a amplitude das frequências em cada instante (Wu et al., 2022).

A Transformada Rápida de Fourier (FFT) é um algoritmo eficiente que calcula a Transformada Discreta de Fourier (DFT) de uma sequência, permitindo a conversão de sinais temporais em seus componentes de frequência.A FFT melhora a eficiência computacional ao reduzir a complexidade de $O(N^2)$ para $O(NlogN)$, tornando-a ideal para o processamento de sinais de áudio em tempo real. O resultado da FFT é um conjunto de números complexos que representam a amplitude e a fase de cada frequência presente no sinal. Esses valores de amplitude são normalmente convertidos em um espectro de potência ou representados em decibeis, facilitando sua interpretação em gráficos como o espectrograma (Arias-Vergara et al., 202). O resultado deste processo pode ser visto na Figura 3,

onde se relaciona o tempo em x, com a frequência em y, tendo as amplitudes em dBs representadas pelas cores conforme a barra de cores à direita da figura.

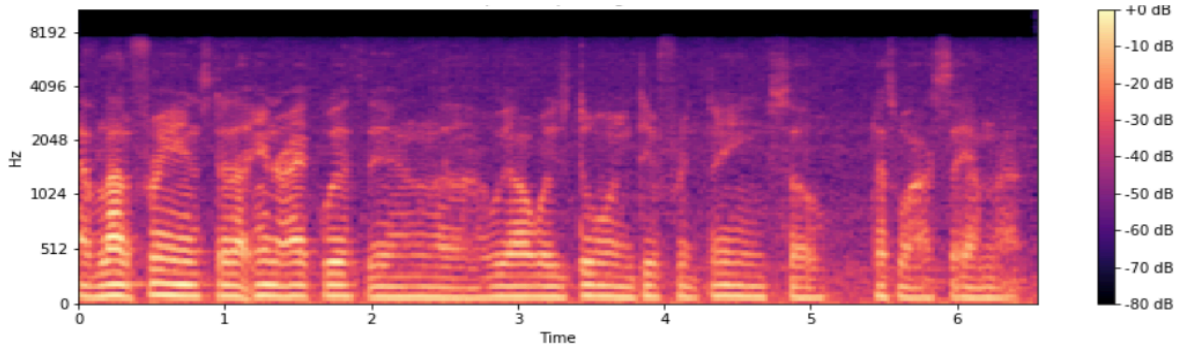

Figura 1: Espectrograma de Frequência Mel.

A partir dos espectrogramas, foi possível alimentar a rede neural, já que as redes neurais são eficazes na extração de características relevantes e padrões abstratos a partir de dados visuais. Assim, a padronização dos arquivos e a extração dos espectrogramas foram etapas fundamentais para garantir a qualidade e consistência na análise automatizada das emoções.

2.4 Rede Neural

Neste trabalho foi utilizada a ResNet-18, que foi desenvolvida por He et al. (2016) para resolver problemas relacionados ao treinamento de redes neurais profundas, como o desaparecimento e a explosão de gradientes. A principal inovação da ResNet é a introdução de conexões residuais, também chamadas de atalhos, que permitem que a entrada de uma camada seja somada diretamente à sua saída, promovendo uma melhor propagação do gradiente durante o treinamento. Com essa abordagem, a ResNet-18 é capaz de treinar redes mais profundas de forma eficiente, superando as limitações de redes tradicionais em diversas tarefas de visão computacional, como a classificação de imagens no conjunto de dados ImageNet.

A arquitetura da ResNet-18 é composta por 18 camadas, sendo distribuídas em camadas convolucionais, camadas de pooling (redução da dimensionalidade por amostragem espacial) e camadas totalmente conectadas, organizadas em blocos residuais. Cada bloco residual contém duas camadas convolucionais seguidas de

normalização e função de ativação ReLU (Rectified Linear Unit), e o atalho é aplicado diretamente à soma dessas camadas. Essa arquitetura permite que a rede aprenda mapeamentos de identidade, o que ajuda a manter a integridade dos dados à medida que são propagados através da rede. A ResNet-18 utiliza camadas de pooling max e global, que contribuem para a redução da dimensionalidade, preservando as informações mais relevantes das características extraídas (HE et al., 2016).

Essa estrutura modular faz da ResNet-18 uma solução versátil, adequada para aplicações que exigem um equilíbrio entre precisão e eficiência computacional. Por ser menos profunda do que outras variantes da ResNet, como a ResNet-50, ela é mais leve e rápida de treinar, o que a torna ideal para uso em sistemas com restrições de hardware, como dispositivos embarcados e móveis. Sua simplicidade, aliada à eficácia dos blocos residuais, possibilita que seja amplamente utilizada em tarefas como classificação de imagens, detecção de objetos e segmentação, estabelecendo um novo padrão no campo de redes neurais profundas (HE et al., 2016).

O transfer learning é uma técnica amplamente utilizada em cenários onde a quantidade de dados disponíveis para treinamento é limitada. Em vez de treinar uma rede neural do zero, o transfer learning utiliza uma rede pré-treinada em um conjunto de dados robusto, como o ImageNet, para resolver novas tarefas. Essa abordagem evita problemas como o overfitting (sobreajuste), que pode ocorrer quando modelos complexos são treinados com poucos dados, levando a uma performance insatisfatória em novos dados. A ResNet-18, por exemplo, é frequentemente usada como base para o transfer learning em tarefas de classificação de imagens, devido à sua arquitetura balanceada, que permite que camadas profundas aprendam representações complexas sem comprometer a generalização do modelo (HE et al., 2016; YOSINSKI et al., 2014).

Além de prevenir o sobreajuste, o transfer learning oferece vantagens consideráveis em termos de eficiência computacional. Treinar redes neurais profundas do zero requer grandes volumes de dados e um tempo de processamento elevado. Ao utilizar uma rede pré-treinada como a ResNet-18, o tempo de

treinamento é significativamente reduzido, pois apenas as camadas superiores da rede precisam ser ajustadas para a nova tarefa, enquanto as camadas inferiores já possuem pesos otimizados que capturam características universais das imagens. Isso permite que a ResNet-18 mantenha um bom desempenho, mesmo em contextos onde a coleta de grandes volumes de dados é inviável (HE et al., 2016; DONAHUE et al., 2014).

A flexibilidade da ResNet-18 também a torna uma escolha sólida para diferentes aplicações além da classificação de imagens, como segmentação e detecção de objetos, especialmente em áreas como medicina e agricultura, onde os conjuntos de dados são frequentemente pequenos. Sua estrutura residual ajuda a mitigar o problema da degradação de redes profundas, tornando-a eficiente para ajuste fino do modelo com novos dados (fine-tuning), especialmente para novas tarefas com dados limitados. Assim, o uso de transfer learning com ResNet-18 não só garante um equilíbrio entre representatividade e complexidade, mas também proporciona uma solução prática e eficaz para a adaptação de redes profundas em diversos contextos (YOSINSKI et al., 2014; HUSSAIN et al., 2018).

3. METODOLOGIA

3.1 Conjunto de Dados

Os dados que fundamentam este estudo foram adquiridos em um ambulatório de gestação de alto risco localizado em um hospital público universitário. Visando um grupo demográfico específico, o estudo inscreveu gestantes acima de 18 anos, com consentimento informado, com pelo menos ensino fundamental e que possuíam smartphone. Estes critérios foram fundamentais para garantir que os participantes pudessem interagir com a tecnologia necessária para a recolha de dados e compreender a investigação em que participavam.

Os participantes forneceram consentimento por escrito após serem totalmente informados sobre os objetivos da pesquisa, a natureza do seu envolvimento e o tratamento dos seus dados. O desenho do estudo foi submetido e aprovado pelo conselho de ética local. As exclusões foram aplicadas criteriosamente para manter a integridade do estudo. Não foram incluídas mulheres que atualmente recebem tratamento para condições médicas graves que requerem hospitalização, aquelas com diagnóstico de deficiência intelectual, deficiências cognitivas graves ou qualquer forma de transtorno psicótico.

A coleta de dados capitalizou a onipresença dos smartphones, permitindo que as participantes se expressassem de forma natural em seus próprios contextos cotidianos. Este método propõe uma alternativa às entrevistas clínicas, proporcionando uma janela para os estados emocionais dos participantes através de comunicações diárias não estruturadas. Todas as gravações foram protegidas com o máximo respeito pelos protocolos de privacidade e proteção de dados.

Inicialmente, os participantes passaram por entrevista com psiquiatra credenciado e responderam a um questionário sociodemográfico. Após a entrevista inicial, quinzenalmente, houve sessão de triagem auto relatada para coleta de amostra de áudio e teste PHQ-4 (Patient Health Questionnaire-4). Para instigar uma resposta, pergunta-se ao paciente: “Como você está se sentindo hoje?”. Todos os dados foram coletados e gerenciados por meio de ferramentas eletrônicas de captura de dados REDCap hospedadas na Universidade de São Paulo (HARRIS,

2009; HARRIS, 2019). Essas questões são respondidas pela gestante, permitindo uma avaliação rápida e objetiva de seu estado emocional.

O PHQ-4 é um questionário ultra breve com quatro itens em uma escala de quatro pontos. Duas questões estão relacionadas à depressão e as outras duas à ansiedade(KROENKE, 2009). A utilização do PHQ-4 pretende ser rápida e fácil para evitar desistências, ao mesmo tempo que avalia tanto a depressão como a ansiedade.

O PHQ-4 tem se mostrado um instrumento eficaz na detecção precoce de sintomas depressivos durante a gestação, auxiliando os profissionais de saúde na identificação de mulheres que necessitam de suporte emocional e tratamento adequado. Suas questões são simples e fáceis de serem compreendidas, o que contribui para que a gestante possa responder de forma autônoma. Além disso, o teste PHQ-4 é de fácil aplicação e baixo custo, tornando-se uma ferramenta viável para ser utilizada em diferentes ambientes de saúde. Sua utilização regular no pré-natal pode contribuir para a melhoria da saúde mental das gestantes, garantindo um acompanhamento mais efetivo e precoce dos casos de depressão (Rodríguez-Muñoz et al., 2020).

Um total de 88 mulheres atenderam aos critérios de inclusão, concordaram em participar do estudo e realizaram pelo menos um dos exames. Não foi realizado nenhum acompanhamento quanto ao diagnóstico ou tratamento. Além dessa população, foram coletadas 110 amostras de áudio, sendo todas as amostras de fala conversacional livre e com resposta relativa de PHQ-4. Tudo isso foi coletado nos smartphones dos pacientes sem controle ambiental, portanto alguns apresentam ruído ambiente ou causado pelo próprio aparelho, volume não padronizado e formatos de arquivo diferentes.

As amostras de áudio utilizadas têm duração média de 37 segundos, com desvio padrão de 39 segundos, sendo o áudio mais curto de 1,62 segundos e o mais longo de 3,02 minutos (181,46 segundos).

Os escores totais do PHQ-4 foram selecionados como referência para a triagem devido à sua capacidade de identificar um ou ambos os transtornos mentais de interesse (depressão e ansiedade). Conforme apontado por Kroenke (KROENKE,

2009), pontuações crescentes foram associadas a múltiplos domínios de comprometimento funcional e carga de sintomas, podendo ser utilizadas como referência de triagem para orientar os pacientes para um diagnóstico posterior com um especialista.

A Figura 2 mostra a distribuição das pontuações totais do PHQ-4 calculadas no momento da gravação. É possível observar maior concentração de escores baixos, cenário condizente com a prevalência na população em questão (BENNETT, 2004; WOODY, 2017). Com base na categorização recomendada da pontuação PHQ-4, todos os pacientes foram estratificados em amostras positivas e negativas com base na pontuação total com um ponto de corte >=6, resultando em 80 (73%) amostras negativas e 30 (27%) amostras positivas. Com base nos itens específicos de depressão e ansiedade do PHQ-4, das amostras positivas, 37 (34%) sugeriram ansiedade, 27 (24%) sugeriram depressão, onde 22 (20%) sugeriram ambos os transtornos e 20 (18%) um ou outro.

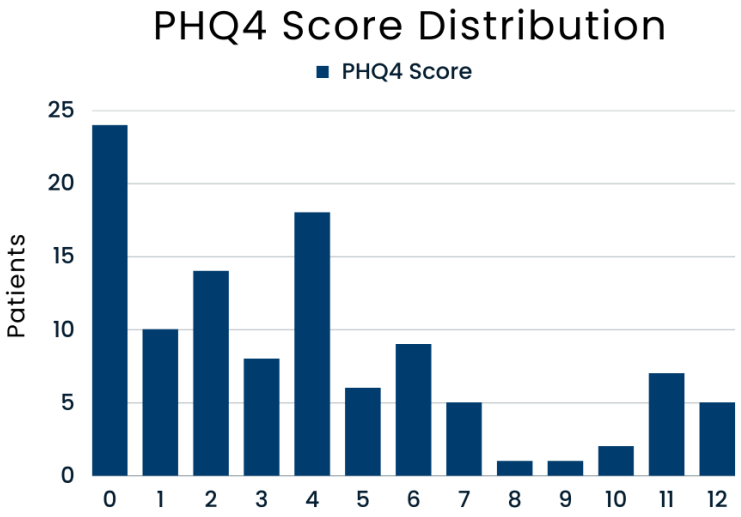

Figura 2: Distribuição de pontuação do PHQ-4 entre os pacientes.

A Figura 3 apresenta a distribuição da idade das pacientes, onde a faixa etária média das pacientes é de 31.3 anos, com a idade mínima registrada de 17 anos e a idade máxima de 45 anos.

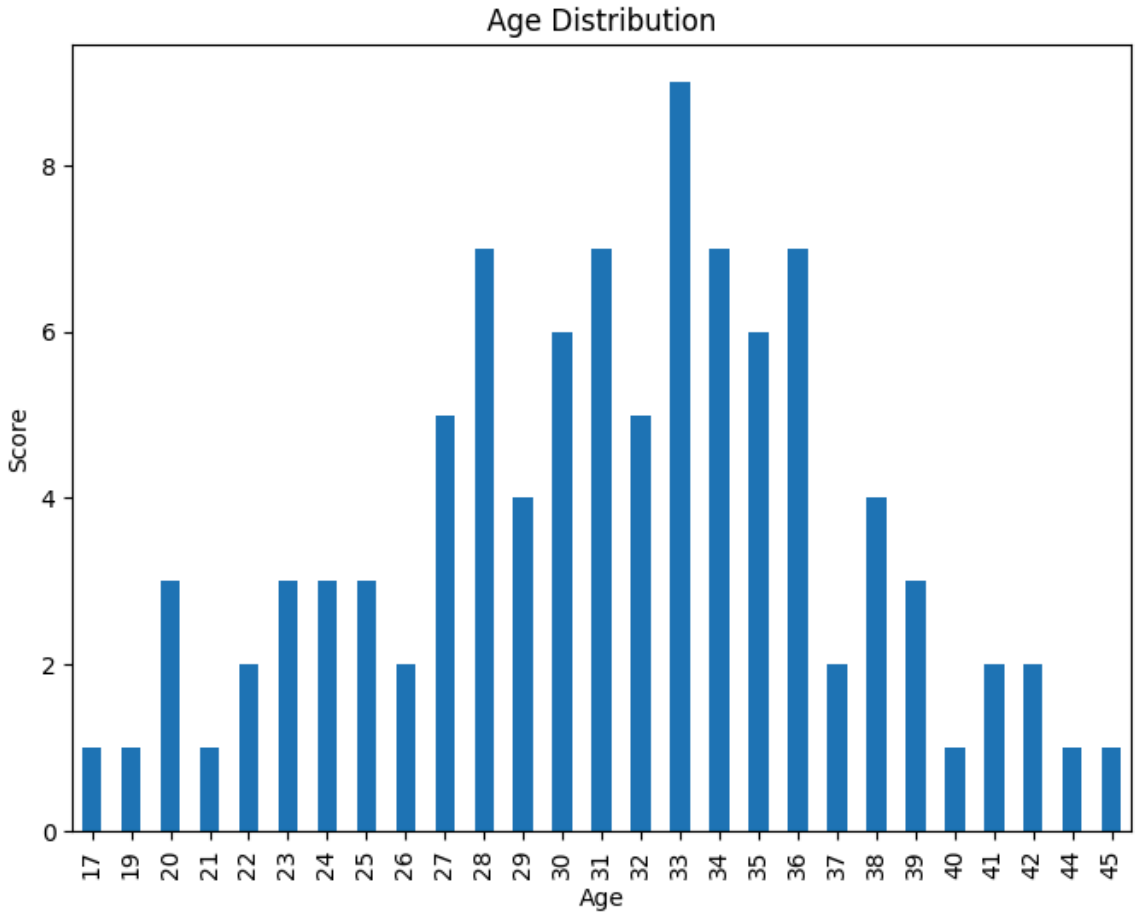

Figura 3: Distribuição de idades dos pacientes.

Dessa forma, o conjunto de dados coletado foi cuidadosamente estruturado para garantir confiabilidade, representatividade e conformidade ética, permitindo a aplicação eficaz de modelos de aprendizado profundo na análise vocal.

3.2 Pré-processamento dos Dados

Para garantir a consistência na análise dos dados, todos os áudios utilizados neste estudo foram padronizados para o formato WAV (Waveform Audio Format). Essa padronização é essencial para evitar inconsistências nos processos subsequentes, garantindo que todas as amostras tenham as mesmas propriedades, como taxa de amostragem e codificação de áudio.

A técnica de fragmentação de áudio é amplamente utilizada no reconhecimento de emoções por meio de sinais de fala, ajudando a melhorar a análise e a extração de características relevantes, como tons de voz, frequência e ritmo. Estudos mostram que essa técnica é fundamental para separar partes de interesse em gravações de áudio, permitindo uma análise mais detalhada de componentes emocionais, como tristeza, raiva ou alegria. Por exemplo, técnicas como a extração de Coeficientes Cepstrais na Escala Mel (MFCCs, do inglês Mel-Frequency Cepstral Coefficients) e outras características temporais são essenciais para melhorar a precisão na identificação de emoções em sistemas automatizados (ALLUHAIDAN et al., 2023; PEDDINTI; POVEY; KHUDANPUR, 2015).

A fragmentação de áudio permite que sistemas de reconhecimento de emoção identifiquem nuances emocionais que podem não ser detectadas em uma análise mais global do sinal. Com a aplicação de técnicas de aprendizado de máquina e inteligência artificial, esse processo é refinado, permitindo a detecção de emoções com altos níveis de precisão (KWON, 2019; CHEN et al., 2018).

No presente trabalho, os áudios foram fragmentados aleatoriamente em segmentos de 3, 4 ou 5 segundos de duração, uma estratégia que permite capturar diferentes nuances emocionais de maneira mais precisa(Barhoumi e Ayed, 2024). Essa etapa se faz crucial, oferecendo uma abordagem promissora para a detecção de emoções, aumentando a precisão e a eficiência do modelo.

Após a padronização e fragmentação dos áudios, foi realizado um processamento específico para a extração dos espectrogramas de cada áudio. O espectrograma é uma representação visual que transforma o sinal sonoro em uma imagem bidimensional, na qual o eixo horizontal representa o tempo, o eixo vertical representa a frequência, e a intensidade das cores indica a amplitude do sinal em cada frequência(Khodzhaev, 2024). Essa técnica é amplamente utilizada, pois facilita a identificação de padrões temporais e espectrais que podem ser complexos demais para serem percebidos diretamente no sinal de áudio bruto.

3.3 Desenvolvimento do Modelo

A arquitetura ResNet-18 foi utilizada como base para a técnica de transfer learning, que por ter sido previamente treinada em grandes bases de dados de imagens, permite reutilizar seus filtros convolucionais para extrair características relevantes dos espectrogramas, reduzindo a necessidade de grandes volumes de dados. Em nosso contexto, os espectrogramas gerados a partir dos áudios são tratados como imagens e utilizados como entrada para a ResNet-18. Com isso, as camadas iniciais da rede podem ser reutilizadas sem necessidade de treinamento do zero, sendo necessário apenas realizar o fine-tuning para que a rede aprenda a distinguir padrões específicos associados à depressão.

A curva ROC e a AUC são métricas ideais nesse contexto, pois avaliam a capacidade do modelo de distinguir entre classes (no caso, presença ou ausência de sinais de depressão), independentemente de limiares específicos. A AUC, em particular, fornece um valor entre 0 e 1 que resume essa performance onde quanto mais próximo de 1, melhor o desempenho do modelo. Para avaliar o desempenho da rede nesse cenário, empregamos validação cruzada, dividindo o conjunto de dados em 5 folds, treinando o modelo em algumas e testando em outras, de forma rotativa. Isso permitiu obter uma estimativa mais robusta e generalizável da performance do modelo.

4. Resultados

Para os experimentos realizados, a configuração final adotada consistiu na utilização de validação cruzada com 5 folds estratificados, assegurando o balanceamento das classes em cada partição dos dados. O treinamento do modelo foi conduzido ao longo de 15 épocas, com uma taxa de aprendizado ajustada para 0,0001. O otimizador selecionado foi o Adam, amplamente reconhecido por sua eficácia em tarefas de otimização de redes neurais profundas. Como função de perda, foi empregada a Entropia Cruzada, abordagem comumente utilizada em problemas de classificação multiclasse, por sua capacidade de mensurar com precisão a divergência entre as distribuições previstas e reais.

Essas escolhas de parâmetros foram fundamentadas na necessidade de maximizar o desempenho do modelo e minimizar erros de generalização. A combinação de validação cruzada e ajustes de hiperparâmetros permitiu uma avaliação rigorosa da rede neural antes da sua aplicação em dados desconhecidos, garantindo que os resultados finais sejam confiáveis e reproduzíveis em diferentes contextos.

O desempenho da rede foi avaliado utilizando como principais métricas a Curva ROC e a AUC (Área Sob a Curva). A Curva ROC ilustra a taxa de verdadeiros positivos (Sensibilidade) no eixo y, em relação à taxa de falsos positivos (Especificidade) no eixo x, para diferentes pontos de corte. Já a AUC mensura a capacidade do modelo em classificar corretamente exemplos positivos em comparação a exemplos negativos.

Na **Figura 4**, é apresentada a Curva ROC gerada para cada fold de validação, juntamente com a AUC correspondente calculada em cada fold.

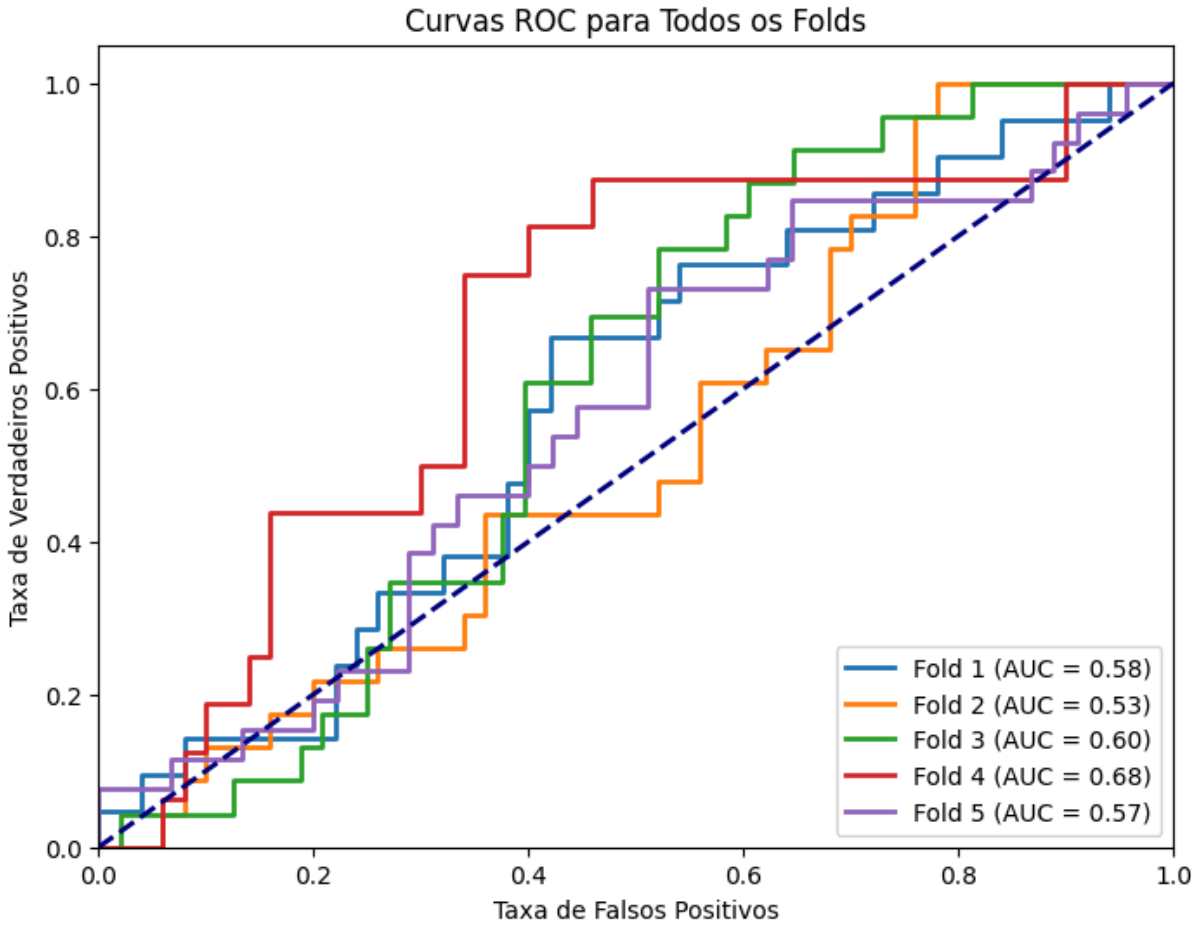

Figura 4: Curva ROC calculada para cada fold.

Embora fragmentos de áudios tenham sido utilizados para treinar e validar a rede, essas métricas foram calculadas considerando uma única resposta por indivíduo presente no conjunto de dados. A **Figura 5** mostra a Curva ROC considerando a média geral obtida em todos os folds de validação.

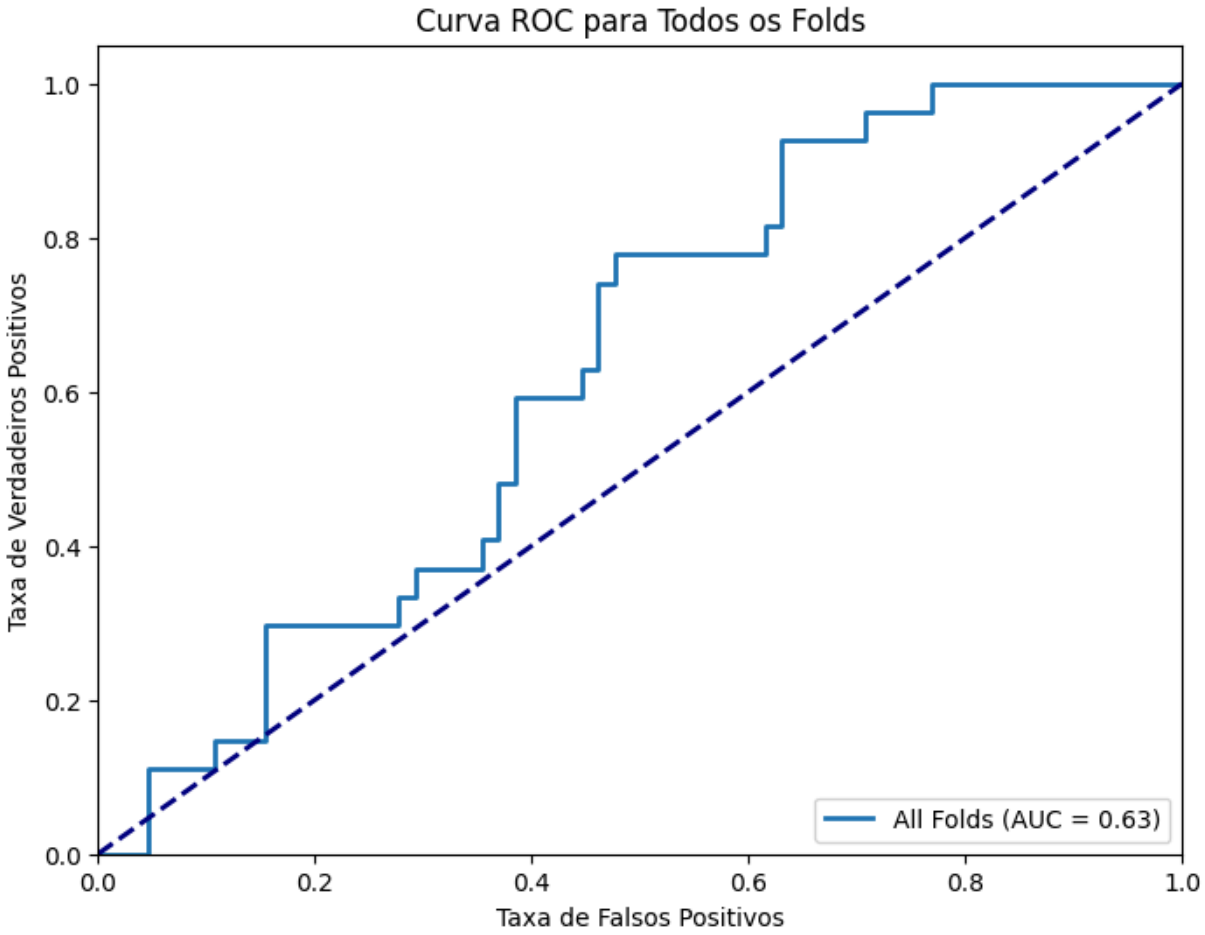

Figura 5: Curva ROC com a média entre todos os folds.

Os resultados da análise indicaram que a AUC do modelo foi de 0,63, o que demonstra um desempenho razoável, pois está acima de 0,5, que seria o valor de uma classificação puramente aleatória. Embora este resultado ainda não seja considerado ideal para um modelo de alta precisão, ele sugere que o modelo consegue captar alguma estrutura relevante dos dados, visto que essa métrica de AUC em torno de 0,63 é comum em trabalhos de detecção de depressão, onde o conjunto de dados é limitado.

Os resultados obtidos neste estudo sugerem que a voz pode ser uma ferramenta útil na identificação de sintomas depressivos em gestantes. Apesar de o modelo ter alcançado um desempenho ainda moderado, ele aponta para um caminho promissor, especialmente por se tratar de uma abordagem acessível e não

invasiva. A literatura mostra que características como ritmo, entonação e frequência da voz costumam mudar em estados depressivos, o que reforça o potencial da voz como indicador de saúde mental. No entanto, essas variações nem sempre são iguais entre diferentes pessoas, o que exige mais estudos e validações.

O uso de smartphones para coleta de dados é um ponto positivo importante, pois facilita o acesso mesmo em regiões com poucos recursos. Por outro lado, desafios como possíveis vieses nos dados, necessidade de validação clínica rigorosa e questões éticas sobre privacidade e uso responsável dos resultados ainda precisam ser enfrentados.

Para transformar esse modelo em uma aplicação confiável, será necessário aumentar a base de dados, testar com diferentes populações e integrar a ferramenta com redes de apoio, como profissionais de saúde e serviços públicos. Também é essencial garantir que, após uma detecção de risco, a gestante receba o suporte necessário. A proposta tem potencial para beneficiar mulheres de baixa renda e pode ser integrada a programas de saúde materna, desde que com o devido acompanhamento e cuidado ético.

5. Conclusão

Este trabalho teve como objetivo investigar a viabilidade da detecção de sintomas depressivos em gestantes por meio da análise da voz, utilizando técnicas de deep learning aplicadas a espectrogramas derivados de áudios coletados em ambiente não controlado. Os principais achados indicam que, mesmo com um conjunto de dados limitado e ruidoso, foi possível obter um desempenho razoável do modelo, com uma AUC média de 0,63.

Esse resultado, embora modesto, aponta para um potencial inicial da abordagem, sugerindo que há padrões vocais relevantes associados à depressão que podem ser captados por redes neurais convolucionais, sobretudo quando há suporte de estratégias como o aprendizado por transferência. No entanto, uma AUC de 0,63 ainda não é suficiente para aplicação clínica imediata, mas sinaliza uma direção para pesquisas futuras, especialmente em contextos de triagem ou monitoramento remoto de saúde mental.

A utilização de espectrogramas de Frequência Mel e a fragmentação dos áudios em segmentos de curta duração permitiram extrair informações acústicas mais granulares, fundamentais para o reconhecimento de sinais depressivos sutis. Além disso, o uso da validação cruzada garantiu uma avaliação mais robusta e generalizável, aspecto essencial diante da alta variabilidade e do volume restrito de dados disponíveis.

A principal contribuição deste trabalho reside na demonstração empírica de que é possível explorar a análise vocal como ferramenta complementar na detecção de sintomas depressivos em gestantes, mesmo em condições realistas e desafiadoras. Trata-se de uma aplicação de técnicas de inteligência artificial no campo da saúde materna, com potencial impacto tanto em políticas públicas quanto em soluções tecnológicas voltadas à atenção psicossocial.

Contudo, a pesquisa apresenta limitações técnicas importantes. O número reduzido de participantes e amostras de áudio restringe o poder preditivo e a capacidade de generalização do modelo. A coleta dos áudios em ambientes não controlados resultou em ruídos e variações de qualidade que podem ter impactado

negativamente o desempenho da rede. Trabalhos futuros devem considerar a ampliação do banco de dados, a padronização do ambiente de coleta e a comparação entre diferentes modelos de deep learning para otimização da acurácia diagnóstica.

Por fim, este trabalho reforça a importância de integrar tecnologia e saúde mental, oferecendo caminhos para que vozes vulneráveis não passem despercebidas. A voz pode tornar-se uma aliada na escuta silenciosa do sofrimento psicológico, especialmente na maternidade, onde o cuidado com a saúde mental é tão necessário quanto o cuidado físico.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLUHAIDAN, Ala Saleh; SAIDANI, Oumaima; JAHANGIR, Rashid; NAUMAN, Muhammad Asif; NEFFATI, Omnia Saidani. Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network. Applied Sciences, v. 13, n. 8, p. 4750, 2023. Disponível em: https://doi.org/10.3390/app13084750. Acesso em: 13 out. 2024.

ARIAS-VERGARA, T.; KLUMPP, P.; VASQUEZ-CORREA, J. C.; et al. Multi-channel spectrograms for speech processing applications using deep learning methods. Pattern Analysis and Applications, [S.l.], v. 24, p. 423–431, 2021. Disponível em: https://doi.org/10.1007/s10044-020-00921-5. Acesso em: 22 maio 2025.

AUCOUTURIER, J.-J. et al. Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. Proceedings of the National Academy of Sciences, Washington, DC, v. 113, n. 4, p. 948–953, 2016.

BARHOUMI, Chawki; AYED, Yassine Ben. Real-time speech emotion recognition using deep learning and data augmentation. Artificial Intelligence Review, [S.l.], v. 58, p. 1–25, 2024. Springer. Disponível em: https://link.springer.com/article/10.1007/s10462-024-11065-x. Acesso em: 22 maio 2025.

BENNETT, Heather A. et al. Prevalence of depression during pregnancy: systematic review. Obstetrics & Gynecology, Philadelphia, v. 103, n. 4, p. 698–709, 2004.

CHAVES, THIAGO ZIMMERMANN LOUREIRO. Uso de Visão Computacional com Redes Neurais Convolucionais para Classificação de Fungos, p. 22, 2023.

CHEN, M.; HE, X.; YANG, J.; ZHANG, H. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Processing Letters, v. 25, n. 10, p. 1440-1444, 2018.

DETECTION of Minor and Major Depression through Voice as a Biomarker Using Machine Learning. MDPI, 2023. Disponível em: https://www.mdpi.com/journal/psychiatry. Acesso em: 13 out. 2024.

DONAHUE, Jeff et al. Decaf: A deep convolutional activation feature for generic visual recognition. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 2014, Beijing. Anais [...]. [S.l.]: PMLR, 2014. p. 647–655.

ESPY-WILSON, C. Alterações na voz podem ser indicativo de depressão, diz pesquisa. UAI, 2014.

GONÇALVES, Angela Maria Corrêa et al. Prevalência de depressão e fatores associados em mulheres atendidas pela Estratégia de Saúde da Família. Jornal Brasileiro de Psiquiatria, v. 67, p. 101-109, 2018.

HARRIS, Paul A. et al. The REDCap consortium: building an international community of software platform partners. Journal of Biomedical Informatics, [S.l.], v. 95, p. 103208, 2019.

HAHN, Lisa; EICKHOFF, Simon B.; HABEL, Ute; STICKELER, Elmar; SCHNAKENBERG, Patricia; GOECKE, Tamme W.; STICKEL, Susanne; FRANZ, Matthias; DUKART, Juergen; CHECHKO, Natalia. Early identification of postpartum depression using demographic, clinical, and digital phenotyping. Translational Psychiatry, Londres, v. 11, n. 121, 2021. Disponível em: https://doi.org/10.1038/s41398-021-01234-0. Acesso em: 13 out. 2024.

HASSANZAD, M.; HAJIAN-TILAKI, K. Methods of determining optimal cut-point of diagnostic biomarkers with application of clinical data in ROC analysis: an update review. BMC medical research methodology, v. 24, n. 1, p. 84, 2024.

Hussain, Mahbub; Bird, Jordan; Faria, Diego. (2018). A Study on CNN Transfer Learning for Image Classification.

KHODZHAEV, Zulfidin. A Practical Guide to Spectrogram Analysis for Audio Signal Processing. University of Texas at Austin, 2024. Disponível em: https://arxiv.org/abs/2403.09321. Acesso em: 13 out. 2024.

KROENKE, Kurt et al. An ultra-brief screening scale for anxiety and depression: the PHQ–4. Psychosomatics, [S.l.], v. 50, n. 6, p. 613–621, 2009.

KWON, Seung. A CNN-assisted enhanced audio signal processing for speech emotion recognition. Sensors, v. 20, n. 1, p. 183, 2019.

MORAIS, L. H. et al. Ansiedade e depressão em gestantes. Revista Saúde e Desenvolvimento, v. 10, n. 6, p. 123-134, 2018.

PEREIRA, P. K.; LOVISI, G. M. Prevalência da depressão gestacional e fatores associados. Revista de Psiquiatria Clínica, v. 35, n. 4, p. 144-153, 2008.

PEDDINTI, Vijayaditya; POVEY, Daniel; KHUDANPUR, Sanjeev. A time delay neural network architecture for efficient modeling of long temporal contexts. In: Sixteenth Annual Conference of the International Speech Communication Association, 2015.

RODRÍGUEZ-MUÑOZ, M. F.; RUIZ-SEGOVIA, N.; SOTO-BALBUENA, C.; LE, H. N.; OLIVARES-CRESPO, M. E.; IZQUIERDO-MÉNDEZ, N. The Psychometric Properties of the Patient Health Questionnaire-4 for Pregnant Women. International Journal of Environmental Research and Public Health, v. 17, n. 20, p. 7583, 2020. DOI: 10.3390/ijerph17207583. Disponível em: https://www.mdpi.com/1660-4601/17/20/7583. Acesso em: 13 out. 2024.

SÁNCHEZ, M. A. et al. Depressão na gravidez: prevalência e fatores associados. Revista Colombiana de Psiquiatria, v. 45, n. 2, p. 73-79, 2016.

SCHERER, S. É possível diagnosticar depressão por meio da voz? BBC, 2016.

SILVA, et al. Reconhecimento de emoções pela voz e expressão facial por estudantes de medicina. Revista da Sociedade Brasileira de Fonoaudiologia, 2024.

Waqas A, Koukab A, Meraj H, Dua T, Chowdhary N, Fatima B, Rahman A. Screening programs for common maternal mental health disorders among perinatal women: report of the systematic review of evidence. BMC Psychiatry. 2022 Jan 24;22(1):54. doi: 10.1186/s12888-022-03694-9. PMID: 35073867; PMCID: PMC8787899.

WU, Chen; ELANGAGE, Janaka. Very Long-Length FFT Using Multi-Resolution Piecewise-Constant Windows for Hardware-Accelerated Time–Frequency Distribution Calculations. Sensors, v. 22, n. 23, p. 9192, 2022. DOI: 10.3390/s22239192.

YOSINSKI, Jason et al. How transferable are features in deep neural networks? In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2014, Montreal. Anais [...]. [S.l.]: NIPS, 2014. p. 3320–3328.

K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.