Aprendizado Profundo Multimodal Aplicado a Sinais de Fala para Classificação e Regressão de Níveis de Depressão

1st Carlos Henrique Brito Malta Leão
Departamento do Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil
chbmleao@ufmg.br

2nd George Teodoro (Orientador)
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte, Brasil
george@dcc.ufmg.br

***Abstract*—This work presents the development of a multimodal deep learning system aiming to predict depression from audio and textual data extracted from clinical interviews. The project began with the validation of facial emotion recognition models, which achieved accuracies of up to 84%. Subsequently, we developed models for depression prediction alongside topic-based data augmentation techniques. Our main contribution is a multimodal model that achieved an F1-Score of 83.04% in the binary classification task (depressed vs. non-depressed). For the more complex task of predicting PHQ-8 severity scores, the model obtained a Mean Absolute Error (MAE) of 4.99, a result that is competitive with the state-of-the-art. By effectively fusing multiple data modalities, this research validates a robust approach for the automatic detection of depression, offering a significant contribution to objective mental health diagnostics.**

***Keywords*: Depression prediction, multimodal deep learning, *dataset DAIC*, facial emotion recognition, voice analysis, text analysis, mental health diagnostics, automated depression detection, *Transformers*, *Convolutional Neural Network*.**

***Resumo*—Este trabalho apresenta o desenvolvimento de um sistema de aprendizado profundo multimodal para a predição de depressão a partir de dados de áudio e texto extraídos de entrevistas clínicas. O projeto iniciou-se com a validação de modelos de reconhecimento de emoções faciais, que alcançaram acurácias de até 84%. Posteriormente desenvolvemos modelos para a predição de depressão junto à técnicas de aumento de dados baseada em tópicos. Nossa principal contribuição é um modelo multimodal que alcançou um F1-Score de 83,04% na tarefa de classificação binária (depressivo vs. não depressivo). Para a tarefa mais complexa de predição dos escores de severidade do PHQ-8, o modelo obteve um Erro Absoluto Médio (MAE) de 4,99, resultado competitivo com o estado da arte. Ao fundir eficazmente múltiplas modalidades de dados, esta pesquisa valida uma abordagem robusta para a detecção automática de depressão, oferecendo uma contribuição significativa para o diagnóstico objetivo em saúde mental.**

***Palavras-chave*: Predição de depressão, aprendizado profundo multimodal, *dataset DAIC*, reconhecimento de emoções faciais, análise de voz, análise textual, diagnóstico em saúde mental, detecção automática de depressão, *Transformers*, *Convolutional Neural Network*.**

I. Introdução

Nos últimos anos, a saúde mental emergiu como um dos principais desafios de saúde pública global. A depressão, em particular, é um dos transtornos mais prevalentes, afetando cerca de 270 milhões de pessoas no mundo, conforme estimativas da Organização Mundial da Saúde (OMS) [1]. Esse transtorno é uma das principais causas de incapacidade mental, trazendo impactos psicológicos profundos, exigindo tratamentos complexos e, em casos extremos, levando ao suicídio. No entanto, um dos maiores obstáculos para o tratamento eficaz do Transtorno Depressivo Maior (TDM) é a avaliação imprecisa e subjetiva.

Os sintomas do TDM muitas vezes são sutis e variam de pessoa para pessoa. Alguns indivíduos podem relatar intensamente sintomas sem, de fato, apresentar um quadro depressivo significativo, enquanto pacientes com depressão grave podem minimizar seus sintomas durante o processo de triagem. A ausência de testes de diagnóstico objetivos para a depressão faz com que médicos dependam da observação clínica para diferenciar casos clínicos de depressão crônica. Estudos indicam que aproximadamente 70% dos pacientes com TDM procuram atendimento médico [2], onde frequentemente é utilizado o PHQ [3], que investiga sintomas como fadiga, insônia e alterações no apetite.

Esses desafios incentivaram a pesquisa em computação afetiva a explorar comportamentos observáveis para prever transtornos mentais, como depressão e transtorno de estresse pós-traumático [4]. Indicadores como expressões faciais e características da voz têm se mostrado promissores para identificar sinais de depressão [5], [6], [7].

Nesse cenário, o presente projeto como um todo visa desenvolver um sistema inovador para monitoramento contínuo e preciso da saúde mental, medindo e avaliando diversos parâmetros relevantes. O projeto é uma colaboração entre o DCC-UFMG e a Dyagnosys LTDA, no âmbito da Unidade Embrapii DCC-UFMG (UE DCC-UFMG), com

o objetivo de criar um sistema acessível e confiável para monitorar o bem-estar mental de pacientes.

O projeto apresenta dois focos principais. Durante o primeiro momento, trabalhamos com a construção de modelos de classificação de emoções a partir de imagens de expressões faciais. Em seguida, seguimos com o principal objetivo do projeto: o desenvolvimento de um modelo multimodal com o fim de auxiliar profissionais da saúde na identificação precoce de pacientes com depressão. A entrada do modelo são dados de múltiplas fontes combinadas, sendo áudio e transcrições de textos ditos por entrevistados. Esse modelo permitirá a construção de um sistema onde o profissional poderá utilizar os dados do paciente como entradas para o modelo, que realizará a análise automática, proporcionando uma classificação do estado mental do paciente.

II. Referencial teórico

Com o avanço das tecnologias de processamento de dados e inteligência artificial, novas abordagens vêm sendo desenvolvidas para identificar sinais de transtornos mentais de forma não invasiva e automatizada. Entre essas inovações, destacam-se a análise de voz, o reconhecimento de expressões faciais e as abordagens multimodais que integram múltiplos tipos de dados. Essas metodologias mostram-se promissoras na captação de informações relevantes para o diagnóstico e monitoramento da saúde mental, oferecendo alternativas que complementam os métodos tradicionais e possibilitam uma detecção precoce e um acompanhamento mais preciso de transtornos como depressão e ansiedade.

Nesta seção, serão abordados conceitos teóricos fundamentais sobre essas tecnologias, acompanhados por uma análise das soluções existentes na literatura e no mercado. A intenção é explorar como essas metodologias estão sendo aplicadas na identificação de problemas de saúde mental, destacando as contribuições dessas abordagens, além de apontar lacunas e desafios ainda presentes para promover avanços nesse campo.

A. Detecção de depressão utilizando análise de voz

A fala, especialmente as pistas paralinguísticas não-verbais, tem se mostrado fundamental na previsão de sofrimento psicológico, como depressão e risco de suicídio, por duas razões principais. Em primeiro lugar, profissionais de saúde consideram aspectos da fala, como prosódia reduzida, produção verbal limitada ou monótona e energia vocal, como indicadores importantes para o diagnóstico de sofrimento. Em segundo lugar, a gravação da fala, por ser uma abordagem não invasiva e discreta, torna-se uma excelente opção para automação em tarefas diagnósticas. Cummins *et al.* [8] conduziram uma revisão abrangente sobre o uso de biomarcadores vocais na avaliação de risco de depressão e suicídio, correlacionando características vocais com escores clínicos. Em outra pesquisa, a análise da fala para avaliação de sofrimento quantifica expressões emocionais, como raiva e excitação.

Estudos também investigam como o ruído e a reverberação afetam a previsão de depressão, utilizando coeficientes cepstrais, como MFCCs e DOCCs. Além disso, a pesquisa de [9] avalia como alterações nas densidades espectrais e na energia da fala podem prever a depressão, analisando a variabilidade acústica em relação à trajetória e ao volume do sinal vocal. Outros estudos exploram as características interculturais e interlinguísticas na fala deprimida através de biomarcadores vocais, enquanto alguns autores examinam mudanças neurocognitivas que influenciam a entrega do diálogo e a semântica, codificando aspectos semânticos com espaços de incorporação lexical e considerando o histórico clínico dos pacientes .

Esses avanços evidenciam a relevância da análise da fala, não apenas como ferramenta diagnóstica, mas também como um meio de compreender mais profundamente as nuances emocionais e psicológicas expressas na voz.

B. Detecção de doenças mentais utilizando expressões faciais

A relação entre o conteúdo verbal e os níveis de doenças mentais é bem estabelecida; no entanto, as características visuais desempenham um papel igualmente crucial na demonstração da ligação entre a depressão e as emoções faciais. Pesquisas indicam que indivíduos com depressão apresentam expressões faciais alteradas, como tremores nas sobrancelhas, sorrisos enfraquecidos, rostos franzidos, olhares intensos, movimentos labiais limitados e uma redução na frequência de piscar. Com o aumento de dados em vídeo e a maior disponibilidade de câmeras de alta qualidade em dispositivos vestíveis e sistemas de vigilância, a análise de emoções e sentimentos faciais tem se consolidado como uma tendência crescente na área de visão computacional.

No estudo [10], os autores aplicam abordagens de aprendizado de múltiplos núcleos convolucionais para o reconhecimento de emoções e análise de sentimentos em vídeos. Dalili *et al.* realizaram uma meta-análise abrangente sobre a correlação entre reconhecimento de emoções faciais e depressão [11]. Em [12], os pesquisadores utilizam técnicas baseadas em LSTM temporal para capturar informações contextuais em vídeos durante a análise de sentimentos. Valstar *et al.* apresentaram o conjunto de dados do Facial Expression Recognition and Analysis Challenge (FERA 2017) [13], projetado para estimar movimentos da cabeça e identificar unidades de ação, o que é essencial para quantificar expressões faciais em cenários complexos. Já Ebrahimi *et al.* introduziram o conjunto de dados do Emotion Recognition in the Wild (EmotiW) Challenge [14], utilizando uma arquitetura híbrida de rede neural convolucional e recorrente (CNN-RNN) para análise de expressões faciais. Esses conjuntos de dados têm sido cruciais para o avanço do estado da arte em pesquisas voltadas

ao reconhecimento de expressões faciais e à previsão de angústia.

C. Abordagens multimodais

Diversos trabalhos recentes apresentaram abordagens multimodais para detecção de problemas relacionados à saúde mental. Os autores de [15] realizaram uma revisão detalhada das técnicas de fusão voltadas para a detecção de depressão. Eles propõem, ainda, uma abordagem de fusão baseada em linguística computacional para a detecção multimodal desse transtorno. O trabalho [7], por sua vez, analisa os níveis de depressão utilizando o conjunto de dados do desafio DAIC, que consiste em entrevistas clínicas, aplicando técnicas de geração de características sensíveis ao contexto e redes neurais profundas que podem ser treinadas de maneira integrada. Adicionalmente, eles incorporam técnicas de aumento de dados fundamentadas na modelagem de tópicos em redes do tipo transformer.

Zhang *et al.* introduziram um conjunto de dados multimodal de emoções espontâneas destinado à análise do comportamento humano [16]. As emoções faciais são capturadas através de varredura dinâmica em 3D, gravações em vídeo de alta resolução e sensores de imagem infravermelha. Para complementar o contexto facial, também são monitorados a pressão arterial, a respiração e as taxas de pulso, permitindo uma avaliação mais completa do estado emocional dos indivíduos. Utilizando dados do desafio AVEC [17], são investigados parâmetros de áudio, vídeo e fisiológicos para revelar percepções sobre o estado emocional dos participantes. No artigo [18], os autores combinam pistas auditivas, visuais e textuais para extrair sentimentos de conteúdos multimídia, empregando técnicas de fusão em níveis de características e de decisão para realizar predições. Em [19], os autores utilizam dados paralinguísticos, pose da cabeça e fixações oculares para detectar a depressão de maneira multimodal. Com a aplicação de testes estatísticos nas características selecionadas, o motor de inferência classifica os sujeitos em categorias de deprimidos e saudáveis.

Além disso, diversos métodos têm sido propostos na literatura para a predição dos resultados do questionário PHQ-8, em que muitos destes utilizam o conjunto de dados DAIC, explorado na Subseção IV-A. Essas abordagens podem ser divididas em duas categorias principais: regressivas e binárias. As abordagens regressivas buscam prever o valor contínuo do questionário, variando de 0 a 24, o que reflete a intensidade dos sintomas de depressão em cada participante. Por outro lado, as abordagens binárias concentram-se em classificar os indivíduos em duas categorias: deprimidos ou não deprimidos, com base em um limiar da pontuação do PHQ-8 em que, geralmente, escores acima de 10 já indicam certo nível de depressão. Ambas as estratégias possuem vantagens e limitações, dado que as abordagens binárias costumam ser mais utilizadas em contextos clínicos para diagnóstico rápido, enquanto as regressivas podem oferecer uma avaliação mais detalhada do grau de depressão.

Em 2019, Ray, Anupama, *et al.* desenvolveram um modelo em [7] baseado em redes neurais profundas para prever os escores do questionário PHQ-8 utilizando o *dataset* DAIC. A arquitetura proposta alcançou um RMSE de 4.28, utilizando uma camada de atenção em cada modalidade (áudio, texto e vídeo) para identificar as características mais relevantes. Essas informações são processadas por redes feedforward específicas para cada modalidade e, em seguida, fundidas em uma camada BLSTM (Bidirectional Long Short-Term Memory) empilhada.

Após a fusão, uma nova camada de atenção é aplicada ao vetor concatenado, ajustando as características mais importantes. O resultado dessa atenção é combinado com o output do BLSTM e processado por um regressor final. O treinamento é realizado de forma end-to-end, garantindo a otimização conjunta de todas as camadas e fortalecendo a capacidade do modelo de capturar relações contextuais entre as modalidades. A Figura 1 apresenta uma representação da arquitetura do modelo.

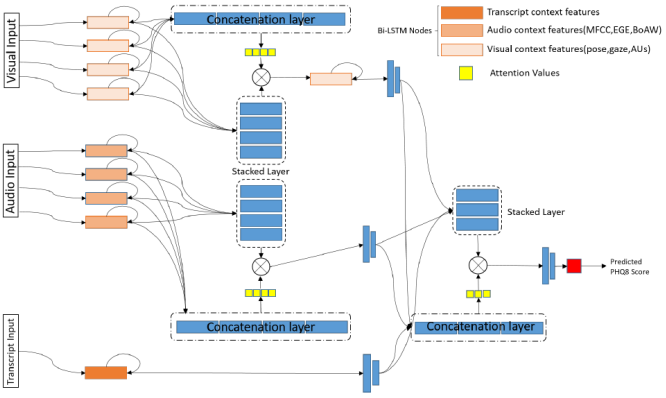


Figura 1. Arquitetura do modelo regressivo [7]

Por outro lado, Lam, Genevieve, Huang Dongyan, e Weisi Lin propuseram, em [20], uma abordagem binária para a detecção de depressão a partir também do conjunto de dados DAIC. O método lida com o baixo volume e desbalanceamento dos dados através de uma estratégia de aumento de dados baseada em modelagem de tópicos. Além disso, o estudo combina técnicas contextuais e baseadas em dados, utilizando um Transformer pré-treinado para modelagem de dados textuais, além de uma rede neural convolucional profunda (1D CNN) para extrair e modelar características acústicas relevantes.

Os resultados obtidos demonstraram desempenho de ponta, tanto para modalidades individuais quanto para o framework multimodal. Os modelos alcançaram F1-scores de 0.78 para texto e 0.67 para áudio, superando métodos anteriores. A combinação dessas modalidades em um framework multimodal elevou ainda mais o desempenho,

alcançando um F1-score de 0.87. A Figura 2 apresenta uma representação visual da arquitetura do modelo.

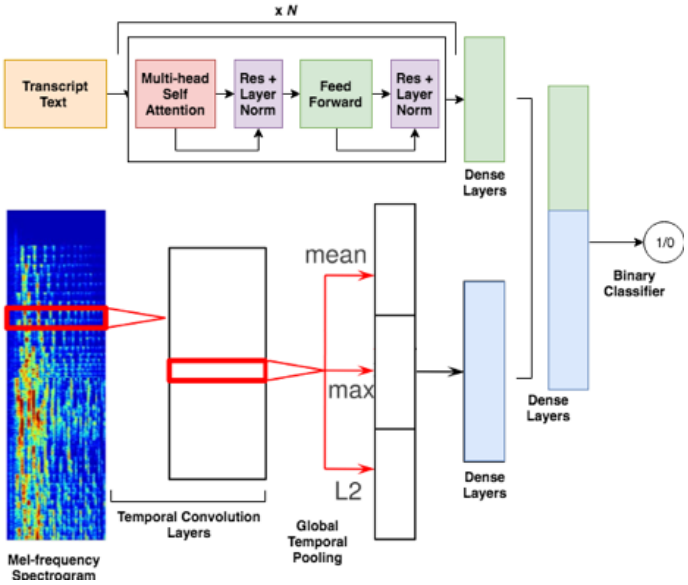


Figura 2. Arquitetura de um modelo de classificação binária [20]

III. Reconhecimento de sentimentos a partir de expressões faciais

Durante o primeiro momento do projeto, tratamos do reconhecimento de sentimentos a partir de imagens de expressões faciais, utilizando diversos conceitos de redes neurais que serão apresentados a seguir. Importante lembrar, apesar do reconhecimento de sentimentos ser uma tarefa natural para humanos, apresenta desafios significativos para os sistemas computacionais. A sutileza das expressões, a influência do contexto e a variabilidade interindividual são fatores que contribuem para essa complexidade [21], [22].

A. Abordagens implementadas

Foi realizado um abrangente estudo do estado da arte no contexto de reconhecimento de sentimentos a partir de imagens de expressões faciais. Nesse sentido, foram selecionadas três principais abordagens modernas para a solução do problema, que serão apresentadas e explicadas nas subseções a seguir.

1) EmoNext: EmoNext é um modelo resultado do artigo [23], publicado em 2023. O estudo apresenta como ponto de partida a rede convolucional *ConvNext* [24], publicada um ano antes durante a *Conference on Computer Vision and Pattern Recognition (CVPR)*, uma das mais renomadas da área.

Nesse sentido, a arquitetura da rede EmoNext, apresentada na Figura 3, possui alguns blocos de ConvNext, baseados nos blocos residuais do artigo [25], porém com algumas atualizações importantes. Além disso, é utilizada a função de ativação *GeLU* ao invés de *ReLU* e uma camada de *Layer Normalization (LN)* ao invés de *Batch Normalization* utilizado anteriormente. Essas atualizações geraram bons resultados, o que fez com que cada bloco consiga extrair, dos dados de entrada, o máximo de informações relevantes para a classificação.

Além disso, essa rede também possui mais dois componentes também relevantes:

- **Spatial Transformer Networks (STN) [26]:** Uma rede convolucional capaz de realizar transformações espaciais dos dados de entrada, possibilitando um alinhamento e localização das faces.
- **Squeeze-and-Excitation (SE) [27]:** Outra rede de convolução que apresenta a capacidade de identificar as características mais importantes ao lidar com os dados de entrada. Dessa forma, a rede consegue designar mais peso para características mais relevantes, ignorando possíveis ruídos.

Cinco modelos foram implementados para essa rede (T, S, B, L e XL) em que se diferem somente em relação à quantidade de camadas de convolução. Porém, todos os modelos apresentam uma etapa de aumento de dados durante o processo de treinamento, como por exemplo, rotação e também o *crop* aleatório. Além disso, todos estes foram pré-treinados no dataset do *ImageNet*.

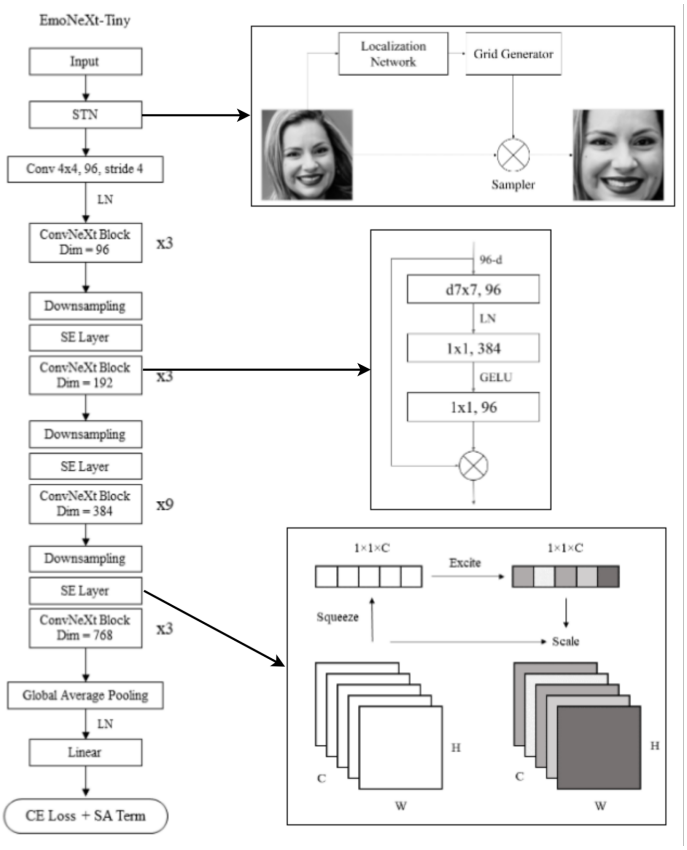


Figura 3. Arquitetura da rede EmoNext [24]

2) Vision Transformer (ViT): O Vision Transformer (ViT) é um modelo de classificação de imagens que intro-

duz a arquitetura de *Transformers*, amplamente utilizada em Processamento de Linguagem Natural, ao domínio da visão computacional. Proposto inicialmente por Dosovitskiy *et al.* [[28], o ViT foi o primeiro modelo a treinar com sucesso um codificador *Transformer* diretamente em imagens, alcançando resultados competitivos e, em muitos casos, superiores às arquiteturas tradicionais baseadas em redes convolucionais (CNNs) quando treinado em grandes datasets, como o *ImageNet*. O funcionamento do ViT pode ser dividido em quatro etapas principais, conforme ilustrado na Figura 4.

- **Divisão em Patches:** A imagem de entrada é dividida em pequenos patches de tamanho fixo (em [28], por exemplo, foi utilizado 16x16 pixels). Esses patches são transformados em vetores unidimensionais através de uma operação de linearização (*flattening*), que transforma a estrutura 2D de cada patch em uma sequência de valores.
- **Embedding e Posicionamento:** Cada vetor resultante passa por uma projeção linear, e a ele é adicionada uma codificação posicional que fornece informações sobre a localização do patch na imagem original. Além disso, um classification token extra é concatenado à sequência de embeddings antes de ser alimentado no codificador Transformer. Este token será utilizado para a tarefa de classificação.
- **Codificador Transformer:** A sequência de embeddings é processada por um codificador padrão, composto por camadas de atenção multi-cabeças (*Multi-Head Attention*) e redes feedforward (MLP). Cada camada também utiliza mecanismos de normalização e conexões residuais para melhorar o treinamento. O classification token, que é atualizado a cada camada, é utilizado como saída final para a classificação.
- **Classificação:** Por fim, um cabeçalho de rede totalmente conectada (MLP Head) é aplicado ao classification token para computar as probabilidades das classes da imagem, como "felicidade", "tristeza", "raiva", entre outros.

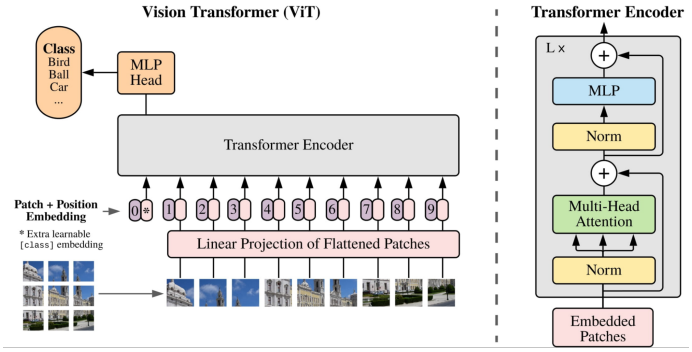


Figura 4. Arquitetura da rede ViT [28]

3) Attentive Pooling Vision Transformer (APViT): APViT é um modelo desenvolvido em [29] fortemente baseado na tecnologia de transformers para a extração de padrões em imagens. A sua arquitetura é demonstrada na Figura 5, em que é utilizada alguma rede de convolução que extrai as características iniciais da imagem de entrada, por exemplo, ResNet [25]. Além disso, para selecionar as regiões mais importantes e significativas da imagem de entrada, é utilizado um sistema baseado em atenção (Attn. Module) juntamente com pooling. Em seguida, essas regiões são convertidas em um vetor de características, chamados de *Patch Tokens*, usando uma rede neural simples e processada por blocos de atenção, os *APT Blocks*. Estes blocos são baseados em Transformers, de forma que filtram gradualmente quais regiões são mais significativas para a predição final. Por fim, a predição é gerada por uma rede neural básica, um *Multi-layer Perceptron*.

A literatura apresenta algumas redes pré-treinadas que utilizam diferentes conjuntos de dados. Nesse sentido, a rede convolucional IR50 [30] é usada para extrair as características das faces, enquanto a rede baseada em transformers ViT-small [29] é utilizada como base para o ATP Block. Além disso a rede convolucional MTCNN [31] é fundamental para localizar e alinhar as faces das imagens de entrada. Por fim, o treinamento da APViT também conta com um aumento de dados, como o crop aleatório e também rotações.

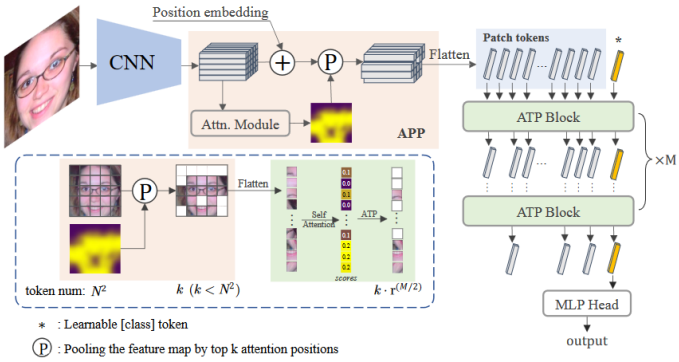


Figura 5. Arquitetura da rede APViT [29]

B. Bases de dados

Todos os métodos selecionados para o reconhecimento de emoções através de imagens são estritamente baseados no aprendizado supervisionado de padrões e características presentes nas imagens e classificações fornecidas. Nesse sentido, a qualidade final dos resultados obtidos dependem diretamente da qualidade do conjunto de dados, tanto das imagens como da representatividade das classificações que serão utilizadas para o treinamento e teste dos modelos.

Nesse contexto, foi realizada uma extensa exploração e varredura do Estado da Arte no âmbito do reconhecimento de expressões faciais, com o objetivo de encontrar potenciais conjuntos de dados para o treinamento, monitoramento e teste dos modelos. A Tabela I apresenta um resumo dos

datasets selecionados, apresentando a quantidade de imagens e suas dimensões. Todos os conjuntos apresentam as mesmas sete classes que representam as emoções básicas: Raiva, Nojo, Medo, Felicidade, Neutro, Tristeza e Surpresa, apenas o FerPlus apresenta a emoção Desprezo. Por fim, os datasets FERPlus e RAF-DB foram selecionados para treinar e avaliar os métodos. Alguns exemplos das imagens e classificações podem ser vistos nas Figuras 6 e 7.

Tabela I
CARACTERÍSTICAS DOS PRINCIPAIS DATASETS DE EMOÇÕES FACIAIS.

Dataset	Quantidade Imagens	Dimensão	Composição	Classes
FERPlus	≈ 35,000	48x48 pixels	Tons de cinza	8 emoções
RAF-DB	≈ 12,000	(100x100 até 500x500)	Coloridas	7 emoções
ExpW	≈ 90,000	Variadas	Coloridas	7 emoções
AffectNet	≈ 450,000	Variadas	Coloridas	7 emoções
KDEF	≈ 5,000	562x762 pixels	Coloridas	7 emoções

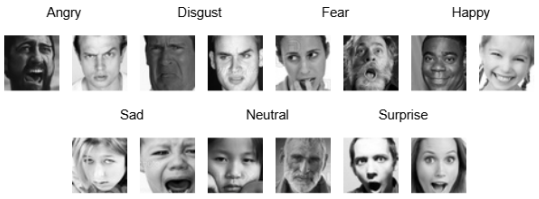


Figura 6. Imagens do dataset FerPlus

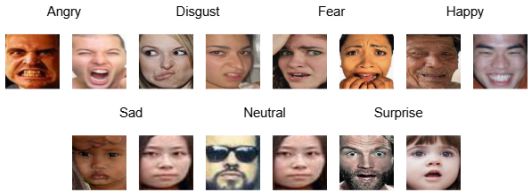


Figura 7. Imagens do dataset RAFDB

C. Resultados obtidos

A validação rigorosa é essencial para avaliar o desempenho do modelo e garantir a robustez e generalização para outros formatos de imagem. Nesse sentido, a escolha de métricas adequadas em conjunto com técnicas robustas de validação são essenciais para encontrar estimativas confiáveis do desempenho dos modelos e cenários reais.

A divisão dos dados em conjuntos de treinamento e teste é uma prática amplamente utilizada no âmbito de aprendizagem de máquina, o que torna possível avaliar a capacidade de generalização do modelo, além de evitar o *overfitting*. O conjunto de treinamento é utilizado durante o processo de aprendizagem do modelo, em que os parâmetros são atualizados constantemente. Já o conjunto de teste, composto somente por dados não utilizados durante o treinamento, é utilizado para avaliar o desempenho do modelo em dados ainda não explorados.

Para a avaliação de modelos existem diversas métricas comumente utilizadas. A acurácia, por exemplo, indica a porcentagem de classificações corretas em relação ao total de amostras. Por outro lado, outras métricas existem, como a precisão (proporção de positivos verdadeiros em relação ao total de positivos classificados) e a revocação (a proporção de positivos verdadeiros em relação ao total de positivos reais). Nesse sentido, a pontuação F1 realiza uma média harmônica entre a precisão e a revocação, o que fornece uma visão mais completa do desempenho do modelo, considerando falsos positivos e negativos.

Ambos *datasets*, foram divididos em conjuntos de treinamento e testes, já pré-definidos pelo conjunto de dados e frequentemente usados na literatura, o que torna os modelos gerados comparáveis. Dessa forma, o conjunto de treinamento foi utilizado para o processo de aprendizagem do modelo, enquanto o de testes para a avaliação do modelo já treinado. Os resultados gerais obtidos, em termos de acurácia geral, com os três modelos selecionados podem ser visualizados na Figura 8.

Além disso, como pode ser visto, os métodos EmoNext [23] e ViT [28] obtiveram resultados similares para ambos os datasets, variando entre 81% até 84%. Por outro lado, o APViT [29] obteve resultados insatisfatórios, por volta de 45% a 65%. As Figuras 9 e 10 apresentam as matrizes de confusão obtidas para os dois melhores métodos.

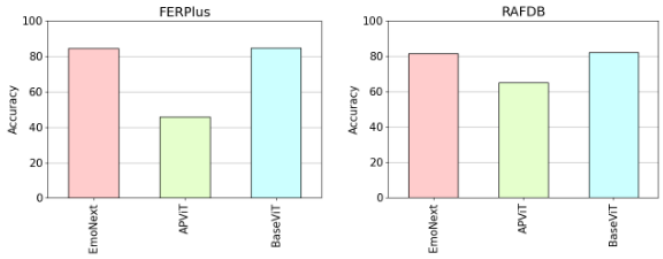
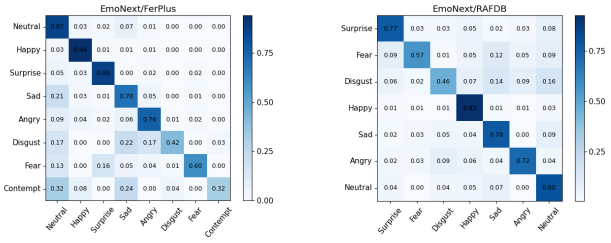


Figura 8. Acurácia dos datasets e modelos desenvolvidos

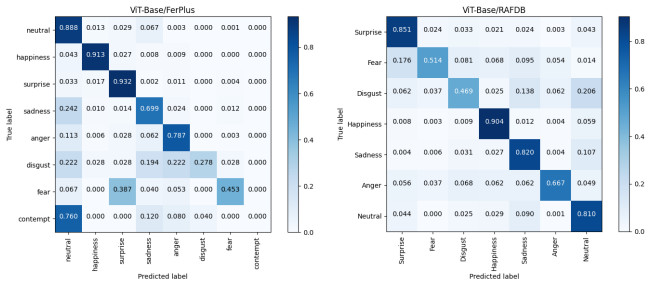


(a) FerPlus (b) RafDB

Figura 9. Matrizes de confusão do modelo EmoNext para ambos os datasets.

IV. MODELO MULTIMODAL DE DETECÇÃO DEPRESSÃO

Durante o segundo momento do projeto, nosso foco foi o desenvolvimento de um modelo multimodal para a predição de níveis de depressão, utilizando uma análise integrada de sinais de imagem, voz e transcrições de fala. O objetivo central foi de propor técnicas inovadoras que



(a) FerPlus (b) RafDB

Figura 10. Matrizes de confusão do modelo ViT para ambos os datasets.

ampliem a capacidade das redes neurais profundas em capturar e modelar relações não lineares entre esses sinais, incorporando conhecimentos contextuais pré-existentes.

Um dos primeiros e principais desafios enfrentados foi a escassez de dados e o desequilíbrio significativo de classes. Na grande maioria dos *datasets* que abordam o tema de depressão, as amostras de indivíduos não deprimidos superam em quantidade as de indivíduos deprimidos consideravelmente. Esse fator limita o desempenho e a generalização de sistemas baseados em dados.

Nesse sentido, priorizamos, primeiramente, o levantamento e a análise de possíveis conjuntos de dados, além do estudo de trabalhos prévios que utilizaram os mesmos *datasets* ou semelhantes. Como parte desse esforço, a Seção IV-A explora todo o processo de criação e desenvolvimento dos modelos de detecção de depressão, desde a escolha do *dataset*, aumento de dados, modelos de classificação binária e um modelo de regressão de níveis.

A. Dataset DAIC

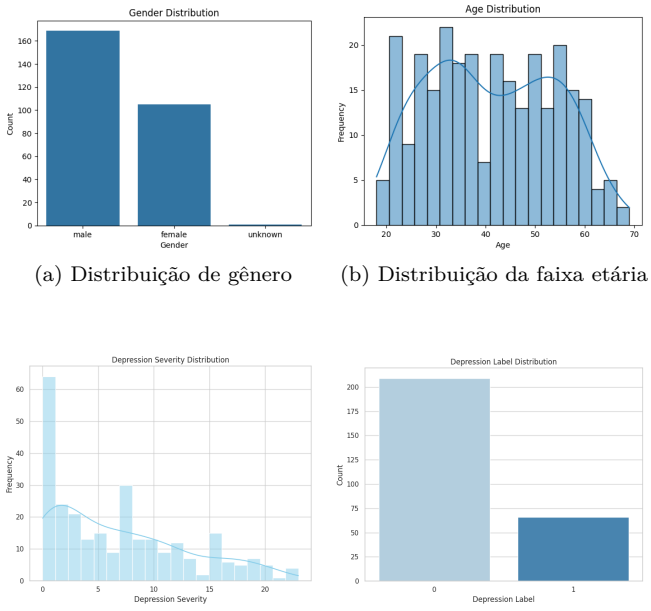
O dataset Distress Analysis Interview Corpus (DAIC) [32] contém entrevistas clínicas desenvolvidas para auxiliar no diagnóstico de condições de sofrimento psicológico, como ansiedade, depressão e estresse pós-traumático. As entrevistas foram conduzidas por humanos, agentes controlados por humanos e agentes autônomos, com diversos participantes com ou sem distúrbios psicológicos, com duração entre 7 e 33 minutos. Os dados coletados incluem gravações de áudio e vídeo e respostas a questionários extensivos. Além disso, as entrevistas foram transcritas e anotadas para uma variedade de características, tanto verbais quanto não verbais.

Os entrevistados responderam ao Patient Health Questionnaire-8 (PHQ-8) [33], um instrumento amplamente utilizado para avaliar sintomas de depressão. O PHQ-8 é derivado do PHQ-9, excluindo a última questão relacionada à presença de pensamentos suicidas, o que torna especialmente adequado para pesquisas onde questões mais sensíveis devem ser evitadas. O questionário avalia a frequência de oito sintomas principais da depressão nas últimas duas semanas, incluindo humor deprimido, perda de interesse ou prazer, dificuldades de sono, fadiga, alterações no apetite, baixa autoestima, dificuldades de concentração e sensação de lentidão ou agitação.

As respostas são classificadas em uma escala Likert de 0 a 3, que varia entre "de forma alguma" e "quase todos os dias", com uma pontuação total que pode variar de 0 a 24. Pontuações mais altas indicam maior gravidade dos sintomas depressivos. A utilização do PHQ-8 no DAIC permite não apenas uma avaliação quantitativa da gravidade da depressão, mas também fornece um ponto de referência clínico para correlacionar os dados multimodais (áudio, vídeo e respostas verbais) com os níveis de sofrimento psicológico dos participantes.

O *dataset* apresenta 189 entrevistados, distribuídos entre homens e mulheres de diferentes faixas etárias, variando de 18 a 69 anos, conforme ilustrado nos gráficos das Figuras 11a e 11b. No entanto, observa-se um desbalanceamento na classificação entre pacientes com e sem depressão em que, aproximadamente, apenas 25% dos participantes são diagnosticados com depressão. Além disso, verifica-se que, à medida que os níveis de gravidade da depressão aumentam, o número de indivíduos correspondentes diminui significativamente. As Figuras 11c e 11d demonstram essas informações de forma visual.

O *dataset* DAIC apresenta 3 divisões oficiais: treinamento (57%, 107 entrevistas); desenvolvimento ou validação (25%, 35 entrevistas); e testes (25%, 47 entrevistas). Os dados de áudios e transcrições das entrevistas de todos os 3 conjuntos de dados foram utilizados durante o treinamento dos modelo de detecção de depressão.



(a) Distribuição de gênero (b) Distribuição da faixa etária

(c) Distribuição da gravidade de depressão (d) Distribuição entre depressivos e não depressivos

Figura 11. Distribuições e características dos participantes do *dataset* DAIC.

B. Aumento e tratamento dos dados

Como discutido anteriormente na Subseção IV-A, o treinamento do modelo com o conjunto de dados DAIC apresenta desafios significativos, principalmente devido à sua quantidade limitada de amostras (189 entrevistas) e ao desbalanceamento de classes (25% depressivos e 75% não depressivos). Além do desequilíbrio binário, observa-se também uma distribuição assimétrica entre os diferentes níveis de severidade da doença: escores baixos no PHQ-8 (indicando ausência ou leve presença de sintomas) são consideravelmente mais frequentes do que escores altos (indicativos de depressão severa).

Para mitigar esses problemas, empregamos uma estratégia de aumento de dados inspirada no estudo de Lam et al. [20]. Inicialmente, foram definidos manualmente sete tópicos principais associados às dimensões avaliadas pelo PHQ-8: (i) interesses pessoais; (ii) qualidade do sono; (iii) sentimentos depressivos; (iv) sensação de fracasso; (v) aspectos de personalidade; (vi) diagnóstico de transtornos mentais; e (vii) relações parentais. A partir desses tópicos, foi possível identificar e extrair trechos relevantes das transcrições das entrevistas, nos quais os participantes abordavam diretamente ou indiretamente esses temas.

Cada trecho extraído foi categorizado com base no tópico correspondente, sendo armazenado juntamente com sua transcrição textual e respectivo áudio. A partir desse banco estruturado de segmentos temáticos, realizamos a geração de novas amostras por meio da recombinação de trechos pertencentes a diferentes tópicos. Esse processo permite criar combinações variadas e coerentes de dados para treinamento.

Importante destacar que a geração de novas amostras foi conduzida de forma controlada, priorizando os participantes com diagnóstico de depressão. Essa abordagem visa compensar o desbalanceamento original do conjunto de dados, contribuindo para um treinamento mais equilibrado do modelo.

Por fim, a estratégia de aumento de dados foi aplicada aos três subconjuntos do *dataset*: (i) treinamento, com 427 instâncias (195 não depressivos e 232 depressivos); (ii) validação, com 150 instâncias (54 não depressivos e 96 depressivos); e (iii) teste, com 187 instâncias (75 não depressivos e 112 depressivos). Essa redistribuição contribuiu para um aumento significativo no número de instâncias, além de proporcionar uma maior representatividade dos participantes com diagnóstico de depressão, promovendo um maior equilíbrio entre as classes.

Além do aumento de dados, também foram realizados alguns pré-processamentos. Um destes foi a conversão dos sinais de áudio para espectrogramas na escala de Mel. O espectrograma é uma representação visual do conteúdo espectral de um sinal ao longo do tempo, ou seja, mostra como as frequências presentes no sinal variam temporalmente. A escala de Mel, por sua vez, é inspirada na percepção auditiva humana, atribuindo maior resolução a frequências mais baixas e menor resolução a frequências mais altas — refletindo, assim, a forma como o ouvido humano percebe sons em diferentes faixas de frequência. A Figura 12 exemplifica um espectrograma na escala de Mel de forma visual.

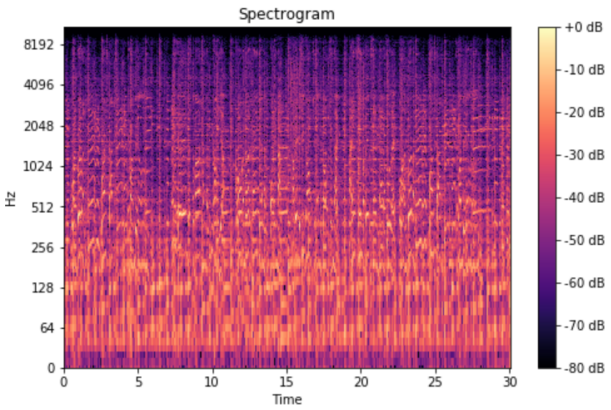


Figura 12. Espectograma na escala de Mel

C. Classificação binária de depressão

Inicialmente, optamos por abordar a tarefa de classificação binária da depressão, ou seja, detectar se um determinado participante apresenta ou não sintomas depressivos. Para isso, utilizamos a categorização original do *dataset*, na qual escores do PHQ-8 superiores a 10 indicam presença de sintomas de depressão significativos.

Por se tratar de um problema com apenas duas classes, esse modelo apresenta uma estrutura de treinamento mais direta e, em geral, proporciona resultados mais estáveis e precisos em comparação com abordagens regressivas, que buscam estimar a pontuação exata do questionário. No entanto, essa simplicidade também limita a riqueza da informação obtida, uma vez que não é possível estimar o grau de severidade dos sintomas apresentados por cada participante.

A seguir, apresentamos os quatro modelos desenvolvidos para lidar com a tarefa de classificação binária.

1) Modelo de áudio: Foi desenvolvido um modelo baseado em redes neurais convolucionais (CNNs) [34] que recebe como entrada espectrogramas de Mel extraídos das falas dos participantes durante as entrevistas. Os espectrogramas de Mel são representações visuais da distribuição das frequências ao longo do tempo, projetadas em uma escala perceptualmente motivada que simula a forma como o ouvido humano percebe variações sonoras — enfatizando frequências mais baixas, onde a audição humana é mais sensível.

Embora os espectrogramas sejam, tecnicamente, representações bidimensionais (tempo *versus* frequência), o modelo faz uso de uma arquitetura de **CNN 1D**, realizando convoluções ao longo da dimensão temporal. Isso se justifica pelo fato de o objetivo principal ser a extração de

padrões sequenciais, como ritmo, entonação e variações temporais da voz, tornando a convolução unidimensional mais apropriada para esse tipo de dado.

A arquitetura do modelo aplica camadas convolucionais sucessivas ao espectrograma, seguidas de normalização em lote (batch normalization), funções de ativação ReLU e camadas de *dropout* para regularização. Após as etapas convolucionais, três tipos de operações de agregação global são aplicadas: *average pooling*, *max pooling* e normalização L2 da soma dos filtros. Essas saídas são então concatenadas e passadas por camadas densas antes da camada final de predição.

A configuração da camada de saída utiliza a função de ativação `softmax` enquanto a função de perda apropriase de `categorical_crossentropy`. O treinamento é realizado com o otimizador Adam, utilizando *early stopping* no dataset de validação para evitar um possível *overfitting*. Além disso, foi realizada uma busca de hiperparâmetros, para encontrar as combinações que geram os modelos mais precisos e robustos.

2) Modelo de transcrições: Para interpretar as transcrições textuais das entrevistas, foi empregado um modelo baseado na arquitetura *Transformer*, mais especificamente o `RoBERTa-base`, uma variante otimizada do BERT. Os *Transformers* foram introduzidos por Vaswani *et al.* [35], e revolucionaram o processamento de linguagem natural ao empregar mecanismos de autoatenção (*self-attention*) que permitem capturar relações contextuais entre palavras em uma sequência, independentemente da distância entre elas.

Neste trabalho, o modelo `RoBERTa-base` foi utilizado com uma camada final de classificação composta por uma função `softmax`, responsável por prever a probabilidade de cada classe — "depressivo"ou "não depressivo". A função de perda adotada foi a `categorical_crossentropy`, adequada à tarefa de classificação multiclasse (neste caso, binária). A entrada do modelo consiste nas transcrições de trechos selecionados das entrevistas, tokenizadas com o `AutoTokenizer` da biblioteca Hugging Face, com truncamento e preenchimento até o limite máximo de 512 tokens.

Durante o treinamento, foi realizada uma busca de hiperparâmetros envolvendo diferentes tamanhos de batch, taxa de aprendizado e regularização com `weight decay`. Para cada combinação, o modelo foi treinado utilizando o algoritmo `AdamW` por até 100 épocas, com *early stopping* ativado com paciência de 10 épocas para evitar sobreajuste.

3) Modelo multimodal - modular: Com o objetivo de explorar a complementaridade entre as modalidades de áudio e texto, desenvolvemos um modelo multimodal de arquitetura modular. Nessa abordagem, os modelos unimodais previamente treinados — baseados em espectrogramas de áudio e transcrições textuais — são utilizados como extratores de características (features). Especificamente, foram extraídas as ativações das últimas camadas antes da predição final de cada modelo, representando de forma densa e semântica os sinais de voz e o conteúdo verbal de cada participante.

Essas representações foram posteriormente concatenadas, formando um vetor multimodal unificado que é utilizado como entrada para uma rede neural do tipo *feedforward*. Diferentemente de abordagens end-to-end, nesse modelo os parâmetros dos extratores de features permanecem congelados: apenas os pesos da rede densa multimodal são ajustados durante o treinamento. Essa separação modular permite melhor controle sobre cada componente do sistema, além de facilitar a análise da contribuição de cada modalidade.

A rede *feedforward* foi projetada com múltiplas camadas densas, seguidas por normalização em lote (*batch normalization*), funções de ativação não-lineares (como ReLU, Leaky ReLU, ELU e Swish), e camadas de *dropout* para regularização. A camada de saída utiliza a ativação `softmax`, adequada à tarefa de classificação binária, sendo otimizada pela função de perda `categorical_crossentropy`.

Para maximizar o desempenho, foi realizada uma busca extensiva de hiperparâmetros (*grid search*) envolvendo os seguintes parâmetros: número de camadas densas, tamanho das camadas, taxa de dropout, função de ativação, taxa de aprendizado, tamanho de batch e número de épocas. O treinamento foi conduzido utilizando o otimizador Adam, com *early stopping* baseado na acurácia de validação para evitar sobreajuste.

4) Modelo multimodal - end-to-end: Para explorar todo o potencial de aprendizado conjunto entre as modalidades de texto e áudio, desenvolvemos uma arquitetura multimodal do tipo *end-to-end*. Diferentemente do modelo modular, nesta abordagem todas as partes do modelo — incluindo os extratores de características de áudio e texto — são treinadas simultaneamente. Isso permite que os parâmetros de cada componente se adaptem conjuntamente ao objetivo da tarefa, promovendo uma fusão mais sinérgica entre os sinais verbais e não verbais.

O componente textual é baseado no modelo `BERT`, utilizado como codificador semântico. O módulo de áudio, por sua vez, é composto por uma rede convolucional 1D profunda, capaz de processar diretamente espectrogramas temporais. As representações aprendidas por ambas as modalidades são então concatenadas e processadas por uma rede densa de fusão (*fusion MLP*), composta por múltiplas camadas ocultas com normalização, ativação (Swish) e *dropout*. A saída da rede utiliza uma camada `softmax`. Perceptivelmente, a estrutura desse modelo é muito semelhante ao modelo multimodal modular, a principal diferença é que todas as camadas são treinadas simultaneamente, sem a necessidade de congelamento de pesos.

5) Resultados: A validação rigorosa é essencial para avaliar o desempenho do modelo. A métrica de precisão por exemplo, indica a proporção de positivos verdadeiros em relação ao total de positivos classificados, enquanto a revocação representa a proporção de positivos verdadeiros

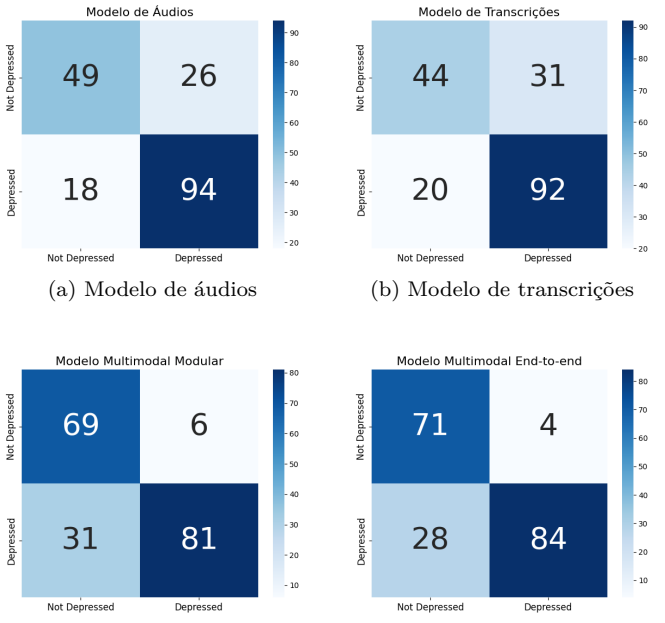
em relação ao total de positivos reais. Nesse sentido a pontuação F1 realiza uma junção das duas métricas ao calcular a média harmônica entre precisão e revocação. Por sua robustez e significância, a métrica F1 foi a principal técnica de avaliação utilizada durante o estudo de classificações.

A Tabela II apresenta os resultados obtidos pelos diferentes modelos na tarefa de classificação binária de depressão. Os modelos unimodais — baseados apenas em áudios ou transcrições — obtiveram desempenhos razoáveis, com destaque para o modelo de áudios, que superou o de transcrições em todas as métricas avaliadas. O modelo de áudios alcançou **76,47%** de acurácia e **76,21%** de F1-Score, enquanto o modelo baseado em transcrições atingiu **72,73%** de acurácia e **72,29%** de F1-Score.

A fusão das duas modalidades resultou em ganhos significativos de desempenho. O modelo multimodal modular, que concatena as representações aprendidas de cada modalidade em uma rede `feedforward`, apresentou uma melhoria considerável, atingindo **80,21%** de acurácia e **80,38%** de F1-Score. Essa melhora demonstra a complementaridade entre os sinais sonoros da fala e as informações linguísticas contidas nas transcrições.

Tabela II
DESEMPENHO DOS MODELOS NAS TAREFAS DE CLASSIFICAÇÃO BINÁRIA

Modelo	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Modelo de áudios	76.47	76.25	76.47	76.21
Modelo de transcrições	72.73	72.37	72.73	72.29
Multimodal (modular)	80.21	83.44	80.21	80.38
Multimodal (end-to-end)	**82.89**	**85.93**	**82.89**	**83.04**



(a) Modelo de áudios

(b) Modelo de transcrições

(c) Modelo multimodal (modular)

(d) Modelo multimodal (end-to-end)

Figura 13. Matrizes de confusão dos modelos de classificação binária de depressão.

O melhor desempenho foi obtido pelo modelo multimodal *end-to-end*, que realiza o processamento conjunto das duas modalidades em um único pipeline treinado de forma integrada. Esse modelo atingiu **82,89%** de acurácia e **83,04%** de F1-Score, destacando-se também em precisão (**85,93%**) e recall (**82,89%**). Tais resultados evidenciam que a integração completa das modalidades durante o treinamento permite uma representação mais rica e discriminativa para a detecção de sinais de depressão. A Figura 13 apresenta as matrizes de confusão correspondentes a cada modelo, permitindo uma análise mais detalhada das classificações corretas e incorretas.

D. Regressão de níveis de depressão

Após a realização da tarefa de classificação binária para depressão, partimos para uma tarefa potencialmente mais complexa: a regressão de níveis de depressão. Em outras palavras, desenvolvemos um modelo regressivo com o intuito de realizar a predição das pontuações do questionário PHQ-8.

As pontuações geradas pelo PHQ-8 variam de 0 até 24, em que 0 representa um indivíduo sem sintomas depressivos significativos, enquanto o 24 é um alerta para depressão severa. Nesse sentido, tratamos de um problema intrinsecamente mais complexo e com potenciais resultados mais instáveis, quando comparado com o modelo de classificação binária. Porém, essa complexidade adicional proporciona também informações mais ricas e pertinentes, já que é possível estimar o grau de severidade dos sintomas apresentados por cada entrevistado.

Semelhante ao problema de classificação binária, iniciamos o treinamento dos quatro modelos: (i) modelo de áudio; (ii) modelo de transcrições; (iii) modelo multimodal modular; e (iv) modelo multimodal end-to-end. Porém, os primeiros três modelos não convergiram e não apresentaram resultados significativos para o estudo. Por outro lado, o quarto modelo apresentou resultados pertinentes, descritos nas subseções a seguir.

1) Modelo multimodal - end-to-end: Para a tarefa de regressão, mantivemos a arquitetura multimodal *end-to-end* semelhante à adotada na tarefa de classificação binária, aproveitando sua capacidade de aprendizado conjunto entre os sinais textuais e acústicos. A principal diferença está na adaptação da camada de saída e da função de perda, que agora são configuradas para uma tarefa de regressão contínua.

O modelo combina representações de transcrições textuais, extraídas por um codificador baseado no `BERT`, com características acústicas processadas por uma rede convolucional profunda 1D. Essas representações são integradas por meio de uma rede densa de fusão (*fusion MLP*) composta por múltiplas camadas ocultas com normalização, ativações não lineares e *dropout*, promovendo uma integração eficaz entre as modalidades.

Na saída, a rede retorna um valor escalar não negativo correspondente à pontuação prevista do PHQ-8,

e a otimização é feita com a função de perda do erro quadrático médio (`MSELoss`). Durante o treinamento, foram utilizadas estratégias como *early stopping* e ajuste dinâmico da taxa de aprendizado por meio do algoritmo `ReduceLROnPlateau`, que contribuem para uma melhor generalização do modelo.

2) Resultados: Os resultados obtidos com o modelo multimodal *end-to-end* para a tarefa de regressão dos níveis de depressão são apresentados na Tabela III.

Tabela III
DESEMPENHO DO MODELO MULTIMODAL *END-TO-END* NA TAREFA DE REGRESSÃO DOS ESCORES DO PHQ-8.

Métrica	**MSE**	**RMSE**	**MAE**	$\mathbf{R^2}$
Valor	34.41	5.87	4.99	0.2464

Apesar da robustez da arquitetura e das estratégias de regularização, os resultados da tarefa de regressão demonstram os desafios de se prever com exatidão os níveis de depressão. O Erro Absoluto Médio (MAE) de **4.99** obtido, embora indique uma margem de erro clinicamente relevante na escala de 0 a 24 do PHQ-8, se mostra competitivo e não distante do estado da arte. Como referência, o trabalho de Ray *et al.* [36], vencedor do desafio AVEC 2019, alcançou um MAE de **4.28** utilizando uma rede multimodal complexa com dados de áudio, vídeo e texto. Considerando que nosso modelo utiliza uma abordagem bimodal (áudio e texto), a proximidade no desempenho é um forte indicativo de sua eficácia. Adicionalmente, o valor de R^2 de **0.2464** confirma que o modelo possui uma capacidade preditiva modesta, mas significativa, explicando aproximadamente 24.6% da variabilidade nos escores.

Essa performance encorajadora, especialmente quando contextualizada com os principais trabalhos da área, posiciona nosso modelo como uma base sólida e promissora. A tarefa de regressão direta permanece intrinsecamente complexa, mesmo para as arquiteturas mais avançadas, o que justifica a contínua exploração de melhorias. Portanto, estratégias futuras como a incorporação da modalidade de vídeo ou a reformulação do problema como uma **classificação ordinal** (prevendo faixas de severidade), são caminhos naturais e promissores para desenvolver um sistema ainda mais robusto e clinicamente útil.

V. CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho se propôs a enfrentar o desafio da avaliação subjetiva da saúde mental, especificamente do Transtorno Depressivo Maior (TDM), por meio do desenvolvimento de sistemas computacionais objetivos e baseados em dados. O projeto foi estruturado em duas fases principais: uma exploração inicial focada no reconhecimento de emoções por meio de expressões faciais e uma fase central dedicada à criação de um modelo multimodal para a predição de depressão a partir de dados de áudio e texto.

Na primeira fase, demonstramos a viabilidade de utilizar arquiteturas modernas, como *EmoNext* e *Vision Transformer (ViT)*, para classificar emoções com alta acurácia, alcançando resultados superiores a 80% em *datasets* de referência. Essa etapa foi fundamental para consolidar o conhecimento em técnicas de visão computacional e aprendizagem profunda que serviram de alicerce para os desenvolvimentos subsequentes.

O foco principal do projeto, no entanto, foi o desenvolvimento de um sistema multimodal para detecção de depressão utilizando o conjunto de dados DAIC. Cientes das limitações de dados, como o desbalanceamento de classes e o número reduzido de amostras, implementamos uma estratégia robusta de aumento de dados que se mostrou crucial para o sucesso do treinamento. Ao recombinar segmentos temáticos das entrevistas, conseguimos não apenas aumentar o volume de dados, mas também equilibrar a distribuição entre as classes, criando um ambiente mais propício para a aprendizagem dos modelos.

Os resultados da tarefa de classificação binária (depressivo vs. não depressivo) foram extremamente promissores. Confirmamos a hipótese de que a fusão de modalidades é superior à análise unimodal. O modelo multimodal *end-to-end*, que aprende representações conjuntas de áudio e texto, alcançou o melhor desempenho, com um **F1-Score de 83,04%**. Este resultado evidencia que a combinação de pistas vocais (prosódia, ritmo) e linguísticas (conteúdo semântico) oferece um sinal mais rico e discriminativo para a identificação da depressão.

Em contrapartida, a predição do nível de severidade da depressão, por meio da regressão dos escores do PHQ-8, revelou-se uma tarefa consideravelmente mais complexa, porém com resultados encorajadores. O modelo multimodal *end-to-end* alcançou um erro médio absoluto (MAE) de **4.99**, um desempenho notavelmente competitivo que se aproxima do estado da arte. Além disso, o coeficiente de determinação ($R^2 \approx 0.25$) demonstra que nosso modelo foi capaz de explicar cerca de um quarto da variabilidade dos dados, um avanço significativo em relação a uma predição aleatória e um indicativo de que padrões relevantes foram capturados.

Em suma, o projeto demonstrou com sucesso a eficácia de uma abordagem multimodal para a classificação binária da depressão e, adicionalmente, produziu um modelo de regressão com desempenho competitivo, estabelecendo uma base sólida para a futura quantificação da severidade. Tais achados confirmam o potencial da fusão de modalidades para criar ferramentas de avaliação da saúde mental mais objetivas. Contudo, a precisão necessária para a quantificação exata da severidade em um contexto clínico permanece um desafio em aberto, motivando os próximos passos desta pesquisa.

A. Trabalhos Futuros

Dada a complexidade da tarefa de regressão e as limitações observadas, diversos caminhos promissores podem ser explorados em estudos futuros:

- **Incorporação da Modalidade Visual:** Um passo natural e promissor é integrar a análise de expressões

faciais, explorada na primeira fase do projeto, ao modelo de predição de depressão. Um sistema trimodal (áudio, texto e vídeo) tem o potencial de capturar sinais não-verbais sutis, como microexpressões, posição da cabeça e mudanças no contato visual, que podem enriquecer a representação do estado afetivo do paciente e melhorar a acurácia do modelo.

- **Validação em Cenários Clínicos Reais:** Para avançar em direção ao objetivo final de criar um sistema de apoio ao diagnóstico, é crucial validar os modelos desenvolvidos com dados coletados em ambientes clínicos reais. A colaboração com a Dyagnosys LTDA será fundamental para testar a robustez e a generalização do sistema com pacientes reais, comparando suas predições com as avaliações de profissionais de saúde.
- **Classificação ordinal:** Reformular a tarefa de regressão como um problema de classificação ordinal pode ajudar a capturar a natureza ordenada, porém discreta, dos escores do PHQ-8. Essa abordagem permitiria preservar o grau de severidade dos sintomas, ao mesmo tempo em que facilita o processo de modelagem.
- **Fusão mais profunda entre modalidades:** Explorar métodos de fusão multimodal mais avançados, como atenção cruzada ou redes co-atencionais, pode melhorar a integração entre os sinais textuais e acústicos, levando a representações mais informativas.

Com o avanço dessas abordagens, espera-se contribuir para a construção de ferramentas clínicas mais confiáveis, acessíveis e objetivas no suporte ao diagnóstico e acompanhamento de transtornos mentais, especialmente a depressão — um dos maiores desafios de saúde pública da atualidade.

Disponibilidade de Dados e Código

O conjunto de dados DAIC, utilizado neste estudo, está disponível para fins de pesquisa acadêmica mediante solicitação em https://dcapswoz.ict.usc.edu/. O código-fonte foi desenvolvido como parte de um projeto de colaboração entre a *Embrapii* e a *Dyagnosys* e, por estar sujeito a um acordo de confidencialidade, é de natureza proprietária.

Referências

[1] World Health Organization. *Depressive disorder (depression)*. 31 March 2023. Acessado em 7 de junho de 2025. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/depression>.

[2] DALY, K.; OLUKOYA, O. Depression detection in read and spontaneous speech: A multimodal approach for lesser-resourced languages. *Biomedical Signal Processing and Control*, v. 108, p. 107959, 2025. ISSN 1746-8094. https://github.com/56kd/MulitmodalDepressionDetection.

[3] KROENKE, K. et al. The phq-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, v. 114, n. 1-3, p. 163–173, abr. 2009. Epub 2008 Aug 27.

[4] VALSTAR, M. et al. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In: *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. New York, NY, USA: Association for Computing Machinery, 2016. (AVEC '16), p. 3–10. ISBN 9781450345163. Disponível em: <https://doi.org/10.1145/2988257.2988258>.

[5] SCHERER, S. et al. Automatic audiovisual behavior descriptors for psychological disorder analysis. *Image and Vision Computing*, v. 32, n. 10, p. 648–658, 2014. ISSN 0262-8856. Best of Automatic Face and Gesture Recognition 2013. https://github.com/yourproject (if relevant).

[6] COHN, J. F. et al. Detecting depression from facial actions and vocal prosody. In: *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. [S.l.: s.n.], 2009. p. 1–7.

[7] RAY, A. et al. Multi-level attention network using text, audio and video for depression prediction. In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. New York, NY, USA: Association for Computing Machinery, 2019. (AVEC '19), p. 81–88. ISBN 9781450369138. Disponível em: <https://doi.org/10.1145/3347320.3357697>.

[8] CUMMINS, N. et al. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, v. 71, p. 10–49, 2015. ISSN 0167-6393. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167639315000369>.

[9] CUMMINS, N. et al. Analysis of acoustic space variability in speech affected by depression. *Speech Communication*, v. 75, p. 27–49, 2015. ISSN 0167-6393. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167639315000989>.

[10] PORIA, S. et al. Convolutional mkl based multimodal emotion recognition and sentiment analysis. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. [S.l.: s.n.], 2016. p. 439–448.

[11] DALILI, M. N. et al. Meta-analysis of emotion recognition deficits in major depressive disorder. *Psychological Medicine*, v. 45, n. 6, p. 1135–1144, 2015.

[12] PORIA, S. et al. Context-dependent sentiment analysis in user-generated videos. In: BARZILAY, R.; KAN, M.-Y. (Ed.). *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017. p. 873–883. Disponível em: <https://aclanthology.org/P17-1081/>.

[13] VALSTAR, M. et al. Fera 2017 - addressing head pose in the third facial expression recognition and analysis challenge. In: *12th IEEE International Conference on Automatic Face Gesture Recognition 2017*. United States: IEEE, 2017. p. 839–847. ISBN 978-1-5090-4024-7. 12th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2017, FG ; Conference date: 30-05-2017 Through 03-06-2017. Disponível em: <http://www.fg2017.org/>.

[14] KAHOU, S. E. et al. Recurrent neural networks for emotion recognition in video. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. New York, NY, USA: Association for Computing Machinery, 2015. (ICMI '15), p. 467–474. ISBN 9781450339124. Disponível em: <https://doi.org/10.1145/2818346.2830596>.

[15] DENG, W. et al. Smart contract vulnerability detection based on deep learning and multimodal decision fusion. *Sensors*, v. 23, n. 16, 2023. ISSN 1424-8220. Disponível em: <https://www.mdpi.com/1424-8220/23/16/7246>.

[16] ZIELONKA, W.; BOLKART, T.; THIES, J. Towards metrical reconstruction ofnbsp;human faces. In: *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*. Berlin, Heidelberg: Springer-Verlag, 2022. p. 250–269. ISBN 978-3-031-19777-2. Disponível em: <https://doi.org/10.1007/978-3-031-19778-9$_1$5>.

[17] RINGEVAL, F. et al. Av+ec 2015: The first affect recognition challenge bridging across audio, video, and physiological data. In: *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. New York, NY, USA: Association for Computing Machinery, 2015. (AVEC '15), p. 3–8. ISBN 9781450337434. Disponível em: <https://doi.org/10.1145/2808196.2811642>.

[18] PORIA, S. et al. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*,

v. 174, p. 50–59, 2016. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231215011297>.
[19] ALGHOWINEM, S. et al. Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Trans. Affect. Comput.*, IEEE Computer Society Press, Washington, DC, USA, v. 9, n. 4, p. 478–490, out. 2018. ISSN 1949-3045. Disponível em: <https://doi.org/10.1109/TAFFC.2016.2634527>.
[20] LAM, G.; DONGYAN, H.; LIN, W. Context-aware deep learning for multi-modal depression detection. In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. [S.l.: s.n.], 2019. p. 3946–3950.
[21] YALçIN, N.; ALISAWI, M. Introducing a novel dataset for facial emotion recognition and demonstrating significant enhancements in deep learning performance through pre-processing techniques. *Heliyon*, v. 10, n. 20, p. e38913, 2024. ISSN 2405-8440. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2405844024149444>.
[22] BISWAS, S.; SIL, J. An efficient expression recognition method using contourlet transform. In: *Proceedings of the 2nd International Conference on Perception and Machine Intelligence (PerMIn '15)*. New York, NY, USA: Association for Computing Machinery, 2015. p. 167–174.
[23] BOUDOURI, Y. E.; BOHI, A. Emonext: an adapted convnext for facial emotion recognition. In: *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*. [S.l.: s.n.], 2023. p. 1–6.
[24] LIU, Z. et al. A convnet for the 2020s. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2022. p. 11966–11976.
[25] HE, K. et al. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016. p. 770–778.
[26] JADERBERG, M. et al. *Spatial Transformer Networks*. 2016. Disponível em: <https://arxiv.org/abs/1506.02025>.
[27] HU, J.; SHEN, L.; SUN, G. Squeeze-and-excitation networks. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2018. p. 7132–7141.
[28] DOSOVITSKIY, A. et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. Disponível em: <https://arxiv.org/abs/2010.11929>.
[29] XUE, F. et al. Vision transformer with attentive pooling for robust facial expression recognition. *IEEE Transactions on Affective Computing*, Institute of Electrical and Electronics Engineers (IEEE), v. 14, n. 4, p. 3244–3256, out. 2023. ISSN 2371-9850. Disponível em: <http://dx.doi.org/10.1109/TAFFC.2022.3226473>.
[30] DENG, J. et al. Arcface: Additive angular margin loss for deep face recognition. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2019. p. 4685–4694.
[31] ZHANG, K. et al. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, v. 23, n. 10, p. 1499–1503, Oct 2016. ISSN 1070-9908.
[32] GRATCH, J. et al. The distress analysis interview corpus of human and computer interviews. In: CALZOLARI, N. et al. (Ed.). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014. p. 3123–3128. Disponível em: <https://aclanthology.org/L14-1421/>.
[33] KROENKE, K.; SPITZER, R. L. The phq-9: A new depression diagnostic and severity measure. *Psychiatric Annals*, v. 32, n. 9, p. 509–515, 2002.
[34] LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, p. 436–44, 05 2015.
[35] VASWANI, A. et al. *Attention Is All You Need*. 2023. Disponível em: <https://arxiv.org/abs/1706.03762>.
[36] RAY, A. et al. *Multi-level Attention network using text, audio and video for Depression Prediction*. 2019. Disponível em: <https://arxiv.org/abs/1909.01417>.