ANÁLISE COMPARATIVA DE ALGORITMOS DE RECONHECIMENTO FACIAL: FACENET E DEEPFACE NA DETECÇÃO DE EMOÇÕES EM APLICAÇÕES CLÍNICAS

COMPARATIVE ANALYSIS OF FACIAL RECOGNITION ALGORITHMS: FACENET AND DEEPFACE IN EMOTION DETECTION IN CLINICAL APPLICATIONS

Gabriel Gonçalves Pereira [1]
Luiz Fernando Nunes [2]

RESUMO

Este estudo comparou o desempenho dos modelos de reconhecimento facial FaceNet e DeepFace na identificação de emoções e comportamentos afetivos em contextos clínicos. A pesquisa foi conduzida utilizando três bases de dados distintas: FER-2013 para o reconhecimento de emoções básicas, LFW para avaliar a robustez em imagens não tratadas e DAiSEE para detectar estados afetivos em vídeos, como distração e confusão, simulando ambientes de análise comportamental. Os modelos foram avaliados com as métricas de Acurácia, F1-Score, MSE, MAE e MAPE, cada uma selecionada para medir aspectos específicos em diferentes condições de imagem. Os resultados indicaram que o FaceNet superou o DeepFace em cenários com variações adversas de qualidade de imagem e no reconhecimento de mudanças emocionais sutis ao longo de vídeos. O estudo contribui para a aplicação de IA no apoio a diagnósticos clínicos, destacando o FaceNet como modelo mais robusto para contextos de saúde mental, sugerindo melhorias futuras em bases de dados e técnicas de pré-processamento.

Palavras-chave: DeepFace; FaceNet; reconhecimento facial; reconhecimento de emoções.

ABSTRACT

This study compared the performance of the facial recognition models FaceNet and DeepFace in identifying emotions and affective behaviors in clinical settings. The research was conducted using three distinct datasets: FER-2013 for recognizing basic emotions, LFW to assess robustness in raw images, and DAiSEE to detect affective states in videos, such as distraction and confusion, simulating behavioral analysis environments. The models were evaluated with the metrics of Accuracy, F1-Score, MSE, MAE, and MAPE, each selected to measure specific aspects in different image

[1] Discente do Centro Universitário Filadélfia de Londrina - UniFil
[2] Docente do Centro Universitário Filadélfia de Londrina - UniFil

conditions. The results indicated that FaceNet outperformed DeepFace in scenarios with adverse image quality variations and in recognizing subtle emotional changes throughout videos. The study contributes to the application of AI in supporting clinical diagnoses, highlighting FaceNet as the most robust model for mental health contexts, suggesting future improvements in databases and preprocessing techniques.

Keywords: DeepFace; FaceNet; facial recognition; emotion recognition.

1 INTRODUÇÃO

O reconhecimento facial, nos últimos anos, emergiu como uma das tecnologias mais promissoras no campo da Inteligência Artificial (IA), com aplicações em em diversas áreas, como segurança, autenticação biométrica e análise emocional. Segundo os estudos de Mellouk; Handouzi (2023) e Guo *et al*. (2024), o reconhecimento facial em aplicações clínicas têm ganhado destaque, especialmente no que se refere à detecção de emoções e comportamentos, o que abre novas possibilidades para o auxílio em diagnósticos psicológicos e psiquiátricos. Nessas aplicações clínicas o reconhecimento facial tem potencial para auxiliar no diagnóstico e acompanhamento de distúrbios emocionais e comportamentais, como é o caso de pacientes com o Transtorno do Déficit de Atenção com Hiperatividade (TDAH).

Através da análise automatizada de expressões faciais e comportamentos não-verbais, o reconhecimento facial pode fornecer dados sobre padrões de distração, impulsividade e falta de foco, características centrais do TDAH. Isso pode facilitar não apenas o diagnóstico precoce, mas também o monitoramento contínuo do paciente durante o tratamento, oferecendo um suporte complementar para avaliações clínicas tradicionais (Flynn *et al*., 2020).

Atualmente, a detecção de emoções por meio do reconhecimento facial tem atraído o interesse de pesquisadores no contexto das emoções básicas, que são: alegria, tristeza, raiva, medo, nojo, desprezo e surpresa, conforme descritas por Paul Ekman (Schiller, 2021). A identificação dessas expressões faciais pode ser essencial para uma análise mais profunda do estado emocional dos indivíduos.

Neste cenário, surgem desafios relacionados à qualidade das imagens utilizadas para o reconhecimento facial. Imagens não tratadas ou com imperfeições, como baixa nitidez, posicionamento inadequado do rosto e variações de iluminação,

são comuns em situações práticas. Assim, é de suma importância que os modelos de reconhecimento facial mantenham uma alta taxa de acurácia, mesmo em condições adversas, para que possam ser aplicados de forma eficaz em ambientes clínicos. Modelos robustos devem ser capazes de detectar emoções em imagens que refletem a realidade, onde as condições não são controladas e as expressões faciais podem ser sutis ou difíceis de identificar (Khan, 2022).

No presente estudo, foram comparados os desempenhos dos modelos de reconhecimento facial *DeepFace* e *FaceNet* em detecção de emoções e identificação de comportamentos afetivos. Ambos os modelos foram testados quanto à sua eficácia no reconhecimento de emoções em condições que simulam um ambiente clínico, onde pacientes podem demonstrar emoções de forma sutil ou atenuada (Zhao, 2024; Cowen, 2021).

Ambos os modelos têm mostrado resultados promissores em vários contextos, no entanto, sua eficácia na detecção de emoções ainda necessita de investigação detalhada. Estudos recentes de ZHAO indicam que esses modelos podem ser adaptados para identificar emoções, mas há lacunas em pesquisas específicas sobre seu desempenho em condições que simulam o ambiente clínico, onde pacientes podem não demonstrar emoções de maneira clara.

Neste estudo, foi analisada a precisão de ambos os modelos, especialmente ao lidar com imagens tratadas e não tratadas. Os resultados indicaram que o *FaceNet* apresentou uma leve vantagem em termos de acurácia e robustez em condições adversas, mostrando-se mais eficiente na identificação de expressões faciais sutis e estados afetivos em imagens de baixa qualidade. Na análise de vídeos, o *FaceNet* também superou o *DeepFace* na detecção de comportamentos dinâmicos como distração e engajamento, comportamentos esses associados ao TDAH. Esses resultados reforçam o potencial do *FaceNet* para ser aplicado em contextos clínicos, onde a precisão na leitura de emoções e comportamentos afetivos, mesmo em condições não ideais, pode apoiar diagnósticos e monitoramentos mais eficazes.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os conceitos essenciais para a compreensão da proposta e da problemática investigada, com foco em reconhecimento facial, reconhecimento emocional e suas aplicações em contextos clínicos. A seguir, são detalhados os principais fundamentos que sustentam o desenvolvimento desta investigação.

2.1 Reconhecimento Facial com Convolutional Neural Network (CNN)

O reconhecimento facial tem como objetivo identificar e verificar a identidade de indivíduos por meio da análise de suas características faciais. A partir de uma imagem ou vídeo, algoritmos de reconhecimento facial capturam pontos de referência da face, como a distância entre os olhos, o comprimento do nariz e a largura da boca, para criar uma "impressão facial" única, permitindo assim a distinção entre diferentes pessoas. Esse processo envolve a extração de padrões biométricos específicos que podem ser armazenados e posteriormente comparados com novas imagens faciais (Begaj, 2020).

Entre as técnicas mais utilizadas nos sistemas modernos de reconhecimento facial, destaca-se o uso de *Convolutional Neural Network* (*CNN*), que são um tipo de rede neural profunda projetada para a análise de dados visuais, como imagens e vídeos. As *CNNs* consistem em várias camadas de neurônios interconectados que processam e extraem características visuais em diferentes níveis de abstração, desde detalhes simples, até formas mais complexas, como os contornos de um rosto (Onyema, 2021).

As camadas convolucionais são a base das CNNs e funcionam aplicando filtros (ou *kernels*) à imagem de entrada para extrair características específicas, como bordas, texturas e contornos faciais. Esses filtros deslizam sobre a imagem, processando pequenas seções de cada vez e gerando o chamado *feature map*, que representa aspectos únicos da imagem. Esse processo permite à rede neural detectar desde detalhes simples, como linhas e formas, até padrões mais complexos, como partes específicas de um rosto.

Após a convolução, as camadas de *pooling* entram em ação para reduzir a dimensionalidade dos mapas de características gerados. Essa redução é feita resumindo informações importantes e eliminando detalhes redundantes. Uma técnica comum, chamada *max pooling*, seleciona o valor máximo em uma região específica da imagem, preservando as características essenciais ao mesmo tempo em que reduz a quantidade de dados.

Outro componente essencial das *CNNs* é a função de ativação *Rectified Linear Unit0* (*ReLU*), que é aplicada às saídas das camadas convolucionais para introduzir não-linearidade no modelo. Isso é importante, pois permite à rede aprender relações complexas entre as características da imagem, essenciais para identificar padrões complexos como expressões faciais. A função *ReLU* é particularmente útil para modelar essas relações e tornar a rede capaz de distinguir detalhes sutis em um rosto.

Finalmente, as camadas *Fully Connected Layer* estão posicionadas próximo ao final da *CNN*. Elas integram todas as características extraídas nas camadas anteriores em um vetor de características único, conhecido como *embedding*, que representa a "impressão facial" da imagem. Essa consolidação de informações permite que a rede produza uma predição final, seja para identificar uma emoção específica ou para classificar a identidade de um rosto. As camadas totalmente conectadas, portanto, têm um papel decisivo na saída do modelo, unindo os padrões detectados para uma análise completa.

2.2 As Emoções Básicas e a Teoria de Paul Ekman

A base para o reconhecimento emocional está na teoria das emoções básicas de Paul Ekman, que identifica alegria, tristeza, raiva, medo, nojo, surpresa e desprezo como emoções universais, expressas de maneira semelhante em diferentes culturas. Essas emoções podem ser detectadas por meio de padrões faciais específicos, conforme Ekman demonstrou em seus estudos pioneiros.

Neste estudo, utilizamos como referência o trabalho de (Schiller, 2021), que oferece uma análise atualizada sobre o reconhecimento dessas sete emoções básicas e suas implicações nas áreas clínica e tecnológica. A escolha do estudo de Schiller

deve à sua abordagem, que, embora traga novas perspectivas, mantém os conceitos fundamentais propostos por Ekman. Schiller reforça a ideia da universalidade das emoções e a importância dos padrões faciais para a identificação emocional, conceitos centrais na teoria de Ekman.

A contribuição de Paul Ekman é amplamente reconhecida por estabelecer uma base científica para o estudo das expressões faciais, demonstrando que as emoções básicas são inatas e não aprendidas, sendo manifestadas de maneira consistente em diversas culturas. Essa teoria tem sido fundamental para o desenvolvimento de tecnologias de reconhecimento emocional.

2.3 Reconhecimento Emocional e Transtornos Mentais

O TDAH é uma condição neuropsiquiátrica marcada por desatenção, hiperatividade e impulsividade persistentes. No artigo de (Gupta, 2024), intitulado "*A Critical Review of Applied Behavior Analysis (ABA)*", os autores discutem abordagens de análise do comportamento aplicada (ABA) para tratar os comportamentos típicos do TDAH. Entre eles, destacam-se a dificuldade em manter a atenção, a inquietação excessiva e a incapacidade de seguir instruções. O estudo reforça a importância do tratamento comportamental para modificar esses padrões e promover o desenvolvimento de habilidades mais adequadas.

Os autores também mencionam que o reconhecimento desses comportamentos pode ser feito por meio da observação direta em diferentes contextos e do uso de instrumentos clínicos, como questionários e escalas, que avaliam a frequência e a intensidade dos sintomas. O desafio atual é aplicar modelos de reconhecimento facial para identificar esses comportamentos, o que pode oferecer uma nova dimensão para a avaliação do TDAH, complementando os métodos tradicionais de diagnóstico.

2.4 DeepFace

O *DeepFace* utiliza uma arquitetura avançada baseada em *Convolutional Neural Network* (*Deep CNN*), que são especialmente eficazes em tarefas de visão

computacional. O modelo processa imagens faciais ao passar por várias camadas convolucionais, onde são extraídas características hierárquicas de diferentes níveis de abstração, desde bordas e texturas até representações faciais mais complexas. Uma das principais técnicas utilizadas no *DeepFace* é a normalização tridimensional dos rostos, na qual as imagens são alinhadas para reduzir variações de perspectiva e de posição da cabeça. Isso ajuda o modelo a criar um mapeamento mais robusto de rostos em diferentes ângulos e expressões, projetando-os em um espaço vetorial latente (DU, 2022).

2.5 FaceNet

O *FaceNet* é um modelo de reconhecimento facial que adota uma abordagem diferente das técnicas tradicionais de classificação facial. Em vez de simplesmente classificar imagens, o *FaceNet* transforma cada rosto em um *embedding*, que por sua vez, é projetado para que rostos semelhantes estejam próximos entre si em um espaço vetorial, enquanto rostos diferentes fiquem distantes. Esse método permite não apenas o reconhecimento de identidade, mas também a comparação, verificação e agrupamento de rostos com alta precisão.

A arquitetura do *FaceNet* utiliza uma técnica chamada *triplet loss*. Essa técnica funciona ao otimizar a rede para minimizar a distância entre *embeddings* de imagens da mesma pessoa e maximizar a distância entre *embeddings* de pessoas diferentes. No processo de treinamento, o modelo processa grupos de três imagens (um âncora, uma positiva e uma negativa), onde a âncora e a positiva são do mesmo indivíduo, enquanto a negativa pertence a outra pessoa. A rede é ajustada para que a distância entre a âncora e a positiva seja menor que a distância entre a âncora e a negativa, aprimorando a discriminação entre faces (Liu, 2021).

2.6 Métricas de avaliação dos modelos

Para esta seção de métricas, fundamentaremos a escolha das métricas de avaliação para os modelos *DeepFace* e *FaceNet* com base na análise discutida por Chicco *et al*. (2021) no artigo "*The coefficient of determination R-squared is more*

informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation". Embora o estudo destaque o coeficiente de determinação R^2 como uma métrica informativa em análises de regressão, a comparação entre as métricas auxilia na compreensão da aplicabilidade de cada uma em cenários de *machine learning* e visão computacional.

As métricas escolhidas aqui — Acurácia, *F1-Score, Mean Squared Error (MSE), Mean Absolute Error (MAE) e Mean Absolute Percentage Error (MAPE)* — refletem diferentes perspectivas do desempenho dos modelos e foram aplicadas com base nas características específicas de cada tarefa e conjunto de dados. Embora o problema central do estudo seja de classificação (detecção de emoções e comportamentos afetivos), as métricas de regressão *MSE, MAE,* e *MAPE* foram utilizadas para oferecer uma avaliação adicional, especialmente em condições adversas.

2.6.1 Acurácia

A Acurácia é uma das métricas fundamentais para avaliar a eficácia de um modelo de classificação. Ela indica a proporção de previsões corretas, considerando tanto os Verdadeiros Positivos (TP) quanto os Verdadeiros Negativos (TN), em relação ao total de previsões realizadas. Essa métrica é amplamente utilizada para determinar o percentual de acertos do modelo em um conjunto de dados, proporcionando uma visão geral de seu desempenho.

Na representação abaixo, é apresentada a equação utilizada para o cálculo da acurácia, onde:

- **TP**: Representa o número de predições corretas de casos positivos;
- **TN**: Representa o número de predições corretas de casos negativos;
- **FP**: Refere-se às predições incorretas em que o modelo classificou um caso negativo como positivo;
- **FN**: Refere-se às predições incorretas em que o modelo classificou um caso positivo como negativo.

A acurácia pode ser expressa pela seguinte fórmula:

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN}$$

2.6.2 F1-Score

O *F1-Score* é uma métrica que combina Precisão e Revocação, proporcionando uma avaliação equilibrada entre a capacidade do modelo de identificar corretamente os casos positivos e a habilidade de minimizar falsos negativos. A Precisão mede a proporção de predições positivas corretas em relação ao total de predições positivas feitas pelo modelo, enquanto a Revocação avalia a proporção de predições positivas corretas em relação ao total de casos realmente positivos. O *F1-Score* é especialmente útil em cenários onde há um desbalanceamento entre as classes, pois considera tanto a quantidade de erros quanto a exatidão das predições positivas.

Na representação abaixo, observa-se que o *F1-Score* é a média harmônica entre a Precisão e a Revocação, onde:

- **P**: representa a proporção de verdadeiros positivos em relação ao total de positivos preditos *(TP/(TP + FP))*;
- **R**: representa a proporção de verdadeiros positivos sobre o total de positivos reais *(TP/(TP + FN))*;

A fórmula para o cálculo do *F1-Score* é dada por:

$$\text{F1-Score} = 2 \cdot \frac{P \cdot R}{P + R}$$

2.6.3 Mean Squared Error (MSE)

O *Mean Squared Error* (MSE), mede a diferença média ao quadrado entre os valores preditos pelo modelo e os valores reais. Essa métrica penaliza predições com grandes erros, elevando as diferenças ao quadrado, o que faz com que o MSE seja sensível a *outliers*.

Na Figura 3, a fórmula do MSE é apresentada como:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Onde:

- y_i: Valor real da observação;
- $\hat{y}_i$: Valor predito pelo modelo;
- η: Número total de observações.

2.6.4 Mean Absolute Error (MAE)

O *Mean Absolute Error* (MAE), mede a média das diferenças absolutas entre os valores reais e os valores preditos. Ao contrário do MSE, o MAE não eleva os erros ao quadrado, o que o torna menos sensível a outliers e, portanto, mais interpretável em termos de erro médio.

A fórmula do MAE, apresentada na Figura 4, é:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Onde:

- y_i: Valor real da observação;
- $\hat{y}_i$: Valor predito pelo modelo;
- η: Número total de observações.

2.6.5 Mean Absolute Percentage Error (MAPE)

O *Mean Absolute Percentage Error* (MAPE), mede o erro médio em termos percentuais entre os valores preditos e os valores reais. Ele é útil para avaliar a precisão de previsões em relação aos valores reais, especialmente quando se deseja expressar o erro em termos percentuais.

A fórmula do *MAPE*, apresentada na Figura 5, é:

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

Onde:

- y_i: Valor real da observação;
- $\hat{y}_i$: Valor predito pelo modelo;
- η: Número total de observações.

2.7 Base de Dados Utilizadas no Estudo

Nesta seção, são apresentados os conjuntos de dados utilizados para o treinamento e a avaliação dos modelos *FaceNet* e *DeepFace*. Foram selecionadas três bases de dados com características distintas: *Facial Emotion Recognition 2013 Dataset* (FER-2013), *Labeled Faces in the Wild* (LFW) e *Dataset for Affective States* (DAiSEE). Cada base foi escolhida por sua relevância no reconhecimento facial, emocional e de estados afetivos, possibilitando uma análise do desempenho dos modelos em diversas condições. Na tabela a seguir, são descritas as principais informações de cada conjunto de dados, como o número de imagens ou vídeos, o tipo de dado (imagem estática ou sequência de vídeo), as emoções ou estados rotulados e as características específicas que tornam esses *datasets* adequados para o estudo do reconhecimento facial e emocional. Assim, o leitor encontrará uma visão geral de cada base e sua aplicação no treinamento dos modelos, facilitando a compreensão dos desafios enfrentados pelo *FaceNet* e *DeepFace* em diferentes cenários (Noboa, 2022).

Quadro 1 - Descrição das bases de dados utilizadas

Dataset	Tipo de Dados	Número de Exemplares	Emoções/Estados	Características
FER-2013	Imagens em tons de cinza	35.887 imagens	Alegria, tristeza, raiva, medo, nojo, surpresa, neutro	Conjunto de dados amplamente aceito para reconhecimento emocional; sete emoções básicas rotuladas.
LFW	Imagens de rostos em ambientes não controlados	13.233 imagens	N/A	Foco em reconhecimento facial com variações de iluminação e ângulo; ambiente desafiador.
DAiSEE	Vídeos com estados afetivos	9.000 vídeos	Distração, confusão, engajamento, frustração	Focado em estados afetivos, utilizado para a análise de comportamentos associados a transtornos mentais como o TDAH.

3 METODOLOGIA

Este estudo utilizou uma abordagem quantitativa, visto que envolveu a análise empírica de dados para comparar a precisão dos dois modelos, *FaceNet* e *DeepFace*.

A pesquisa também se enquadra como uma comparação experimental, onde diferentes bases de dados e condições são utilizadas para avaliar o desempenho dos modelos em cenários variados, envolvendo o reconhecimento de emoções e comportamentos afetivos.

Para garantir uma comparação objetiva dos resultados entre *DeepFace* e *FaceNet*, foram utilizadas cinco métricas principais: Acurácia, *F1-Score*, *MSE*, *MAE* e *MAPE*. Cada uma dessas métricas forneceu uma perspectiva diferente sobre o desempenho dos modelos, o que permitiu uma análise detalhada em diversos aspectos. Na tabela abaixo, detalha-se o papel de cada métrica na análise, explicando como elas foram aplicadas nos diferentes contextos de avaliação dos modelos.

Acurácia	Avaliada principalmente nas bases *FER-2013* e *DAISEE* para verificar a taxa geral de acertos no reconhecimento de emoções básicas e comportamentos afetivos. Usada para entender como os modelos se comportam em um cenário geral de classificações corretas versus incorretas.
F1-Score	Utilizada na base *FER-2013* para avaliar a capacidade dos modelos em detectar emoções mais raras ou sutis, como o medo ou nojo, equilibrando o impacto de falsos positivos e falsos negativos.
MSE	Aplicada especialmente na base *LFW*, onde há condições adversas. O *MSE* foi usado para identificar grandes desvios entre as predições dos modelos e os valores reais em imagens não tratadas.
MAE	Utilizada em todas as bases, mas com destaque na *DAISEE*, onde o *MAE* ajudou a medir o erro médio linear entre as predições de estados afetivos e os valores reais ao longo dos vídeos.
MAPE	Aplicada na base *DAISEE* para calcular o erro percentual nas predições dos estados emocionais como distração e engajamento, facilitando a análise da precisão relativa dos modelos em identificar comportamentos ao longo dos vídeos.

Para realizar a análise comparativa dos modelos , foram utilizadas três bases de dados distintas: *FER-2013*, *LFW* e *DAiSEE*. A base *FER-2013* foi utilizada para avaliar a precisão no reconhecimento de emoções básicas. Para isso, as imagens

foram organizadas em dois subconjuntos: treinamento e validação, sendo que aproximadamente 80% das imagens foram usadas para treinar os modelos e 20% para validar os resultados. Ambas as divisões seguiram um critério aleatório, garantindo que os modelos fossem testados em diferentes amostras emocionais. Os modelos foram treinados diretamente nas imagens brutas da *FER-2013*, sem tratamento adicional, para que pudessem captar as expressões faciais rotuladas com emoções.

A base *LFW*, composta por imagens não tratadas e capturadas em condições não controladas, foi utilizada para testar a robustez dos modelos em situações adversas. Os modelos foram previamente treinados na *FER-2013* e, em seguida, testados na *LFW* sem adaptação adicional. O foco aqui foi verificar a estabilidade e precisão dos modelos em imagens de baixa qualidade, simulando cenários do mundo real. As métricas de erro absoluto e percentual, como *MAE* e *MAPE*, foram especialmente importantes nesta fase, permitindo uma comparação clara entre as predições corretas e as falhas.

A base *DAiSEE* foi usada para avaliar a detecção de comportamentos afetivos em vídeos. Esta base foi aplicada para medir a capacidade dos modelos em reconhecer expressões mais sutis e dinâmicas em tempo real. Como o foco é em vídeos, a abordagem para este conjunto de dados foi diferente das imagens estáticas. Os vídeos foram processados quadro a quadro, com os modelos sendo aplicados a cada imagem individualmente. Ao longo do vídeo, as expressões foram avaliadas cumulativamente, permitindo a análise do comportamento afetivo ao longo do tempo, o que é essencial para identificar padrões relacionados ao TDAH. A precisão na detecção de mudanças emocionais ao longo dos vídeos foi uma das métricas chave na avaliação da eficácia dos modelos nesse contexto clínico.

Todas as bases foram pré-processadas de maneira consistente, sem grandes alterações nos dados originais, para garantir que os resultados refletissem a performance dos modelos em condições próximas da realidade. As métricas de avaliação, foram aplicadas igualmente a todos os conjuntos de dados, permitindo uma comparação justa entre os dois modelos.

3.1 Ferramentas Utilizadas

Para a implementação e análise comparativa dos modelos, foram utilizadas diversas ferramentas e bibliotecas que desempenharam papéis essenciais no pré-processamento dos dados, treinamento dos modelos e análise dos resultados. Os modelos foram implementados utilizando as bibliotecas *TensorFlow* e *Keras*, que são usadas em tarefas de aprendizado profundo. A biblioteca *OpenCV* foi utilizada para o pré-processamento das imagens e vídeos, ajudando na detecção de faces, redimensionamento e normalização das imagens antes de serem alimentadas nos modelos.

Para as bases de dados de imagens, como *FER-2013* e *LFW*, o *OpenCV* foi utilizado para garantir que todas as imagens tivessem o mesmo tamanho e formato, ajustando-as conforme necessário para que fossem compatíveis com os modelos treinados. No caso da base *DAiSEE*, composta por vídeos, o *OpenCV* foi utilizado para processar os vídeos quadro a quadro, extraindo as imagens faciais e alimentando os modelos de forma sequencial. Esse processamento foi importante para garantir que os modelos pudessem detectar comportamentos afetivos dinâmicos, em vez de imagens estáticas.

A biblioteca *NumPy* foi muito utilizada para manipulação de dados e cálculos estatísticos necessários para os cálculos das métricas de desempenho. Após cada rodada de testes, os resultados das predições foram armazenados em *arrays NumPy*, facilitando o cálculo das métricas e a geração de gráficos comparativos entre os modelos.

Para visualizar os resultados e facilitar a interpretação dos dados, foi utilizada a biblioteca *Matplotlib*. Essa ferramenta foi usada para gerar gráficos que compararam o desempenho dos modelos nas diferentes bases de dados. Gráficos de barras e linhas foram gerados para ilustrar a acurácia, *F1-Score* e os erros absolutos e percentuais dos modelos, permitindo uma visualização clara de como cada modelo se comportou nos diferentes cenários testados. Com essas ferramentas, foi possível conduzir a análise comparativa entre *DeepFace* e *FaceNet* de maneira eficiente, permitindo replicabilidade e controle sobre cada etapa do experimento.

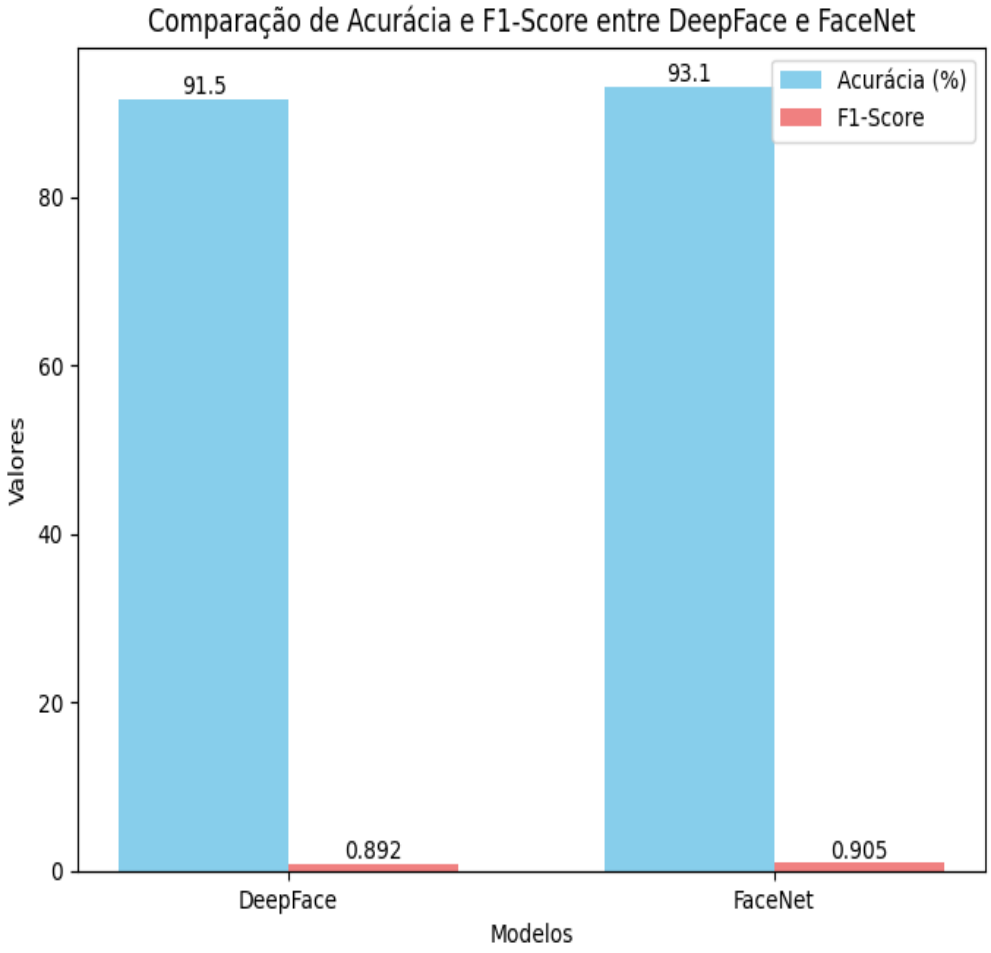
4 DESENVOLVIMENTO E RESULTADOS

A seguir, são apresentados os resultados obtidos na análise comparativa dos modelos com base nas métricas de avaliação descritas na metodologia. Cada experimento foi conduzido conforme as especificações das bases de dados *FER-2013*, *LFW* e *DAiSEE*, utilizando as ferramentas de visualização e processamento descritas. Para melhor entendimento dos resultados, gráficos e descrições detalham o desempenho dos modelos em cada conjunto de dados e os aspectos específicos que foram avaliados em cada caso.

4.1 Treinamento e Avaliação nos Dados de Emoções Básicas (*FER-2013*)

O primeiro experimento envolveu o treinamento e validação dos modelos na base *FER-2013*. Esse conjunto de dados foi selecionado para avaliar a precisão dos modelos no reconhecimento de emoções básicas. As imagens foram divididas em conjuntos de treinamento (80%) e validação (20%), sem tratamento adicional. Os modelos foram testados com as métricas de Acurácia e *F1-Score*, que são particularmente relevantes para a avaliação de classificação correta de emoções. Os gráficos representados na Figura 1, ilustram visualmente as métricas de Acurácia e *F1-Score*.

Figura 2 – Comparação de acurácia e F1-Score entre os modelos na base *FER-2013*



Fonte: Os Autores (2024).

Os resultados representados na Figura 1, mostram que o *FaceNet* apresentou uma ligeira vantagem sobre o *DeepFace* tanto em termos de acurácia quanto em *F1-Score*. O gráfico de barras sugere que o *FaceNet* possui uma leve vantagem, devido ao uso de *embeddings*, que auxiliam na representação das características emocionais das expressões.

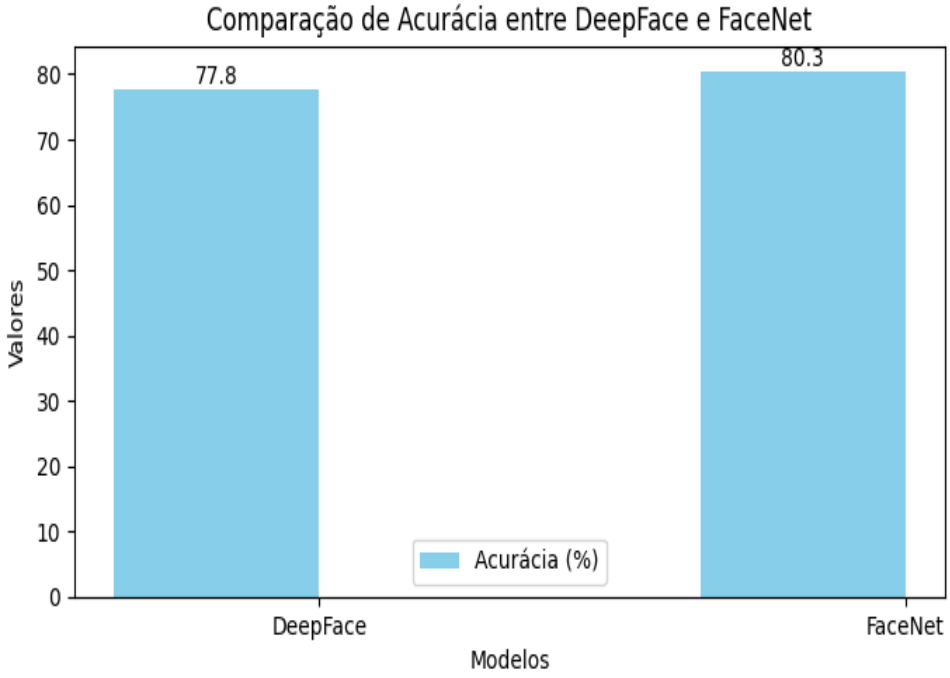
4.2 Desempenho dos modelos em imagens não tratadas (LFW)

Na segunda fase, os modelos foram testados na base *LFW*, composta por imagens de rostos capturados em condições adversas e não controladas. Esse experimento teve como objetivo avaliar o desempenho dos modelos em cenários mais

próximos do mundo real. As métricas de *MSE* e *MAE* foram utilizadas para analisar o desempenho em imagens com qualidade inferior.

Na Figura 2, é possível visualizar a comparação da acurácia de ambos os modelos ao lidarem com imagens de menor qualidade. O gráfico mostra uma queda de desempenho nos dois modelos devido às condições adversas, mas também evidencia que o *FaceNet* manteve uma taxa de acurácia ligeiramente superior à do *DeepFace*.

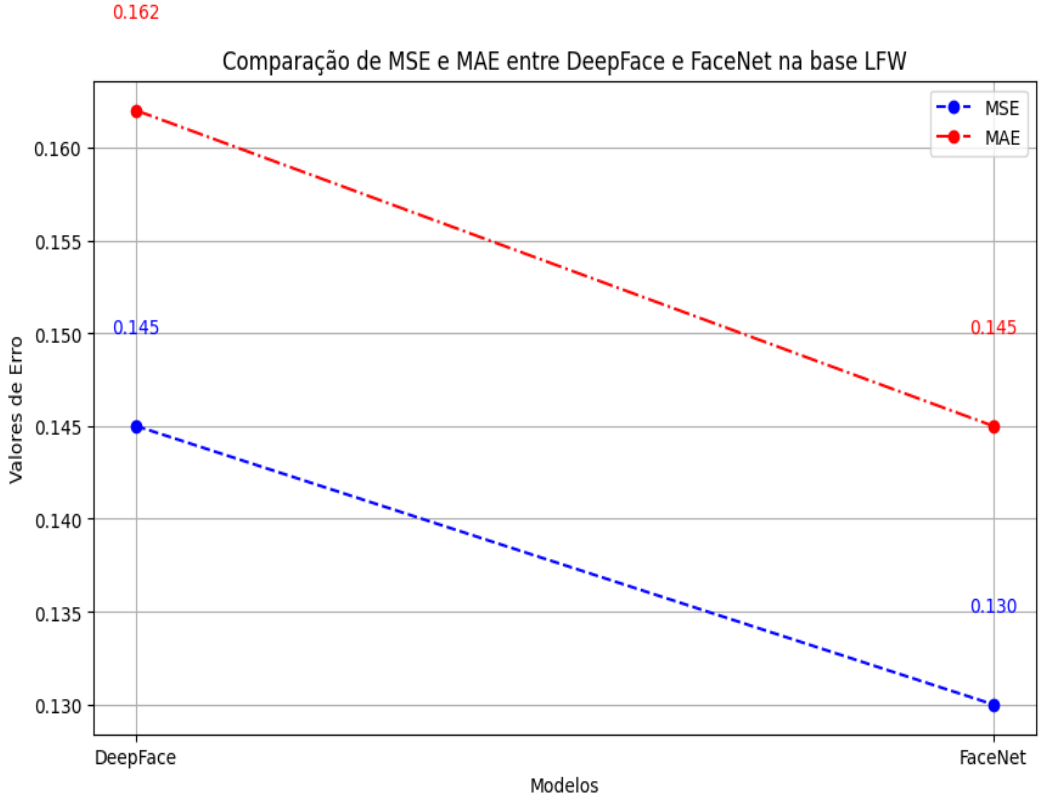
Figura 2 – Comparação de acurácia entre os modelos na base *LFW*



Fonte: Os Autores (2024).

Na Figura 3, observa-se que, embora ambos os modelos tenham tido uma queda na acurácia devido à natureza das imagens da *LFW*, o *FaceNet* ainda apresentou desempenho superior. O menor valor de *MSE* e *MAE* no *FaceNet* sugere que ele é mais resistente a condições adversas e imperfeições visuais, mantendo uma menor média de erro em relação ao valor real das emoções. A análise visual desses resultados indica que o *FaceNet* possui maior tolerância a ruídos e variações nas imagens, uma característica essencial para aplicações clínicas.

Figura 3 – Comparação de MSE MAE entre os modelos na base *LFW*

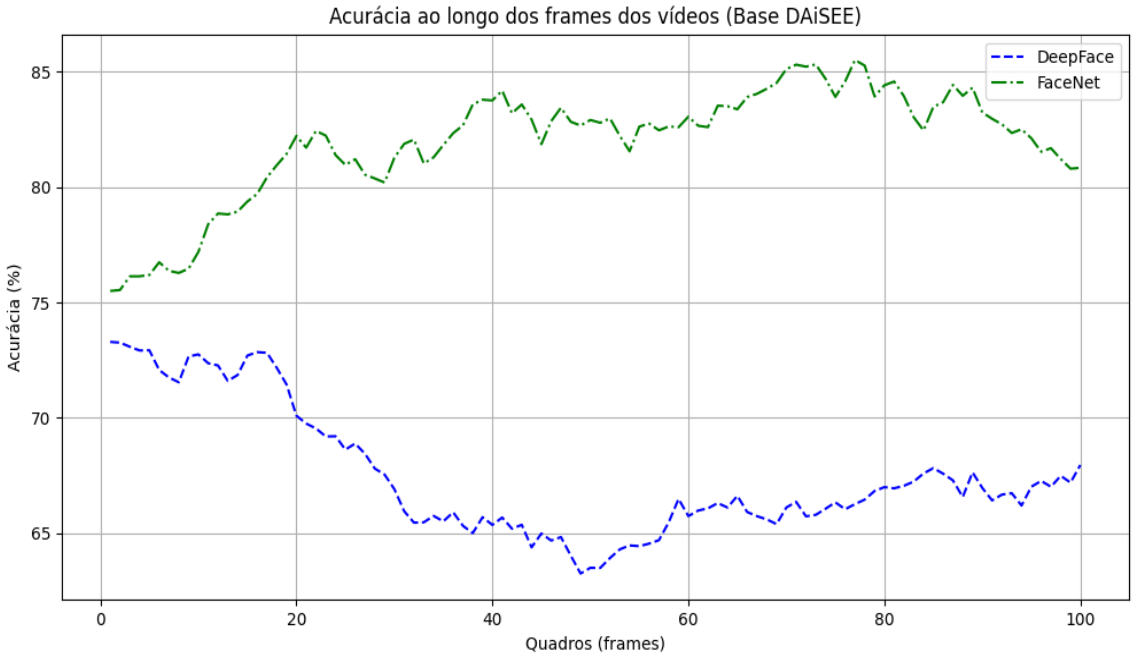


Fonte: Os Autores (2024).

4.3 Detecção de Comportamentos Afetivos em Vídeos (DAiSEE)

O terceiro experimento utilizou a base DAiSEE, composta por vídeos rotulados com estados afetivos como distração e confusão. Esse experimento foi conduzido para avaliar a capacidade dos modelos em identificar padrões emocionais dinâmicos ao longo do tempo, simulando um ambiente clínico onde mudanças de comportamento são relevantes. Os vídeos foram processados quadro a quadro, com cada quadro sendo analisado individualmente para garantir que as expressões fossem capturadas de forma cumulativa. Para essa avaliação, foi dada especial atenção às métricas de MAPE e Acurácia. Na Figura 4, nota-se uma oscilação nos valores de acurácia ao longo dos quadros, especialmente em momentos onde o modelo precisou captar mudanças emocionais dinâmicas.

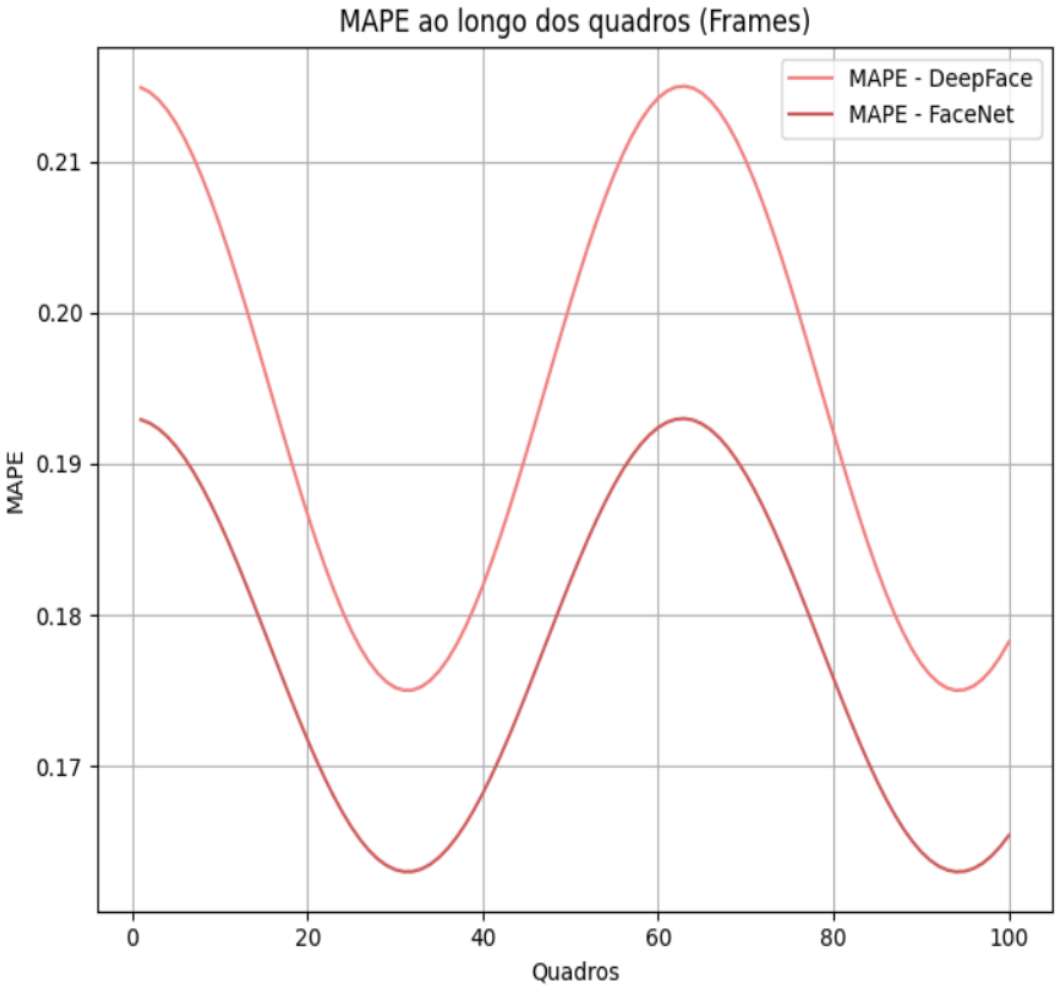
Figura 4 – Acurácia ao longo dos frames dos videos (Base DAiSEE)



Fonte: Os Autores (2024).

Na Figura 5, nota-se uma oscilação nos valores de MAPE ao longo dos quadros, especialmente em momentos onde o modelo precisou captar mudanças emocionais dinâmicas. Ainda assim, o FaceNet manteve uma performance ligeiramente melhor em termos de MAPE, sugerindo que ele consegue captar com mais precisão pequenas variações emocionais, o que é essencial na análise de vídeos clínicos. O maior valor de MAPE no DeepFace indica que o modelo apresentou maior desvio percentual em relação aos valores reais ao longo do tempo, o que pode impactar na sua aplicação para detecção precisa de mudanças afetivas em contextos clínicos.

Figura 5 – MAPE ao longo dos frames dos videos (Base DAiSEE)

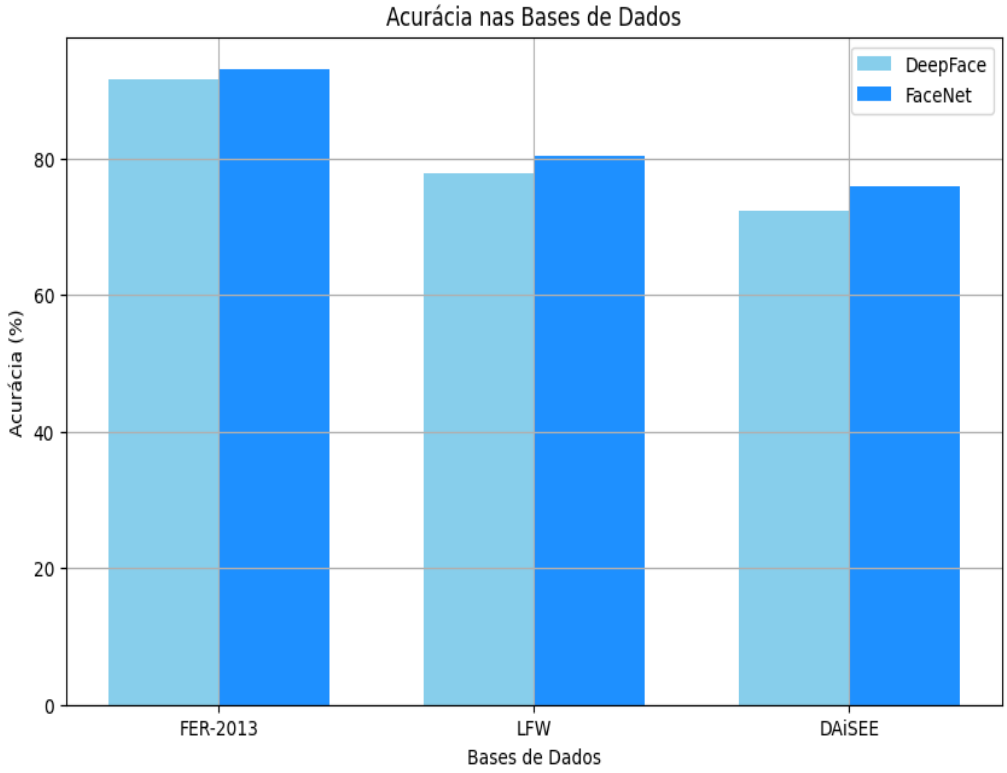


Fonte: Os Autores (2024).

4.4 Análise Geral dos Resultados

A seguir, são apresentados os resultados gerais obtidos nos experimentos realizados. Na Figura 6 abaixo ilustra a acurácia dos modelos nas três bases de dados avaliadas, proporcionando uma comparação visual do desempenho dos modelos em diferentes cenários de teste. O gráfico permite observar como os modelos se comportam em cada conjunto de dados, destacando as variações de precisão associadas às características específicas de cada base utilizada.

Figura 6 – Comparação de Acurácia entre as 3 bases de dados



Fonte: Gabriel Gonçalves Pereira (2024)

O diferencial do *FaceNet* reside em seu sistema de *embeddings*, que cria representações vetoriais das características faciais, permitindo que o modelo mantenha a precisão mesmo em condições adversas e cenários dinâmicos. Esse mapeamento em espaço vetorial torna o *FaceNet* mais resiliente ao ruído visual e permite que ele capture variações sutis nas expressões ao longo do tempo, características essenciais para análises clínicas de detecção de estados afetivos em ambientes onde as condições de captura podem variar.

A escolha das métricas foi pensada para oferecer uma análise adequada a cada base, utilizando *MSE* e *MAE* em situações com maior ruído e *MAPE* em contextos temporais, enquanto acurácia e *F1-Score* complementam a avaliação onde as emoções eram bem rotuladas. Essa abordagem permitiu uma visão completa das capacidades e limitações de cada modelo para diferentes cenários de aplicação.

5 CONCLUSÃO

Este estudo comparou o desempenho dos modelos de reconhecimento facial *DeepFace* e *FaceNet* em diferentes contextos, focando no reconhecimento de emoções básicas e na identificação de comportamentos afetivos em ambientes adversos e dinâmicos. Os experimentos realizados nas bases *FER-2013, LFW* e *DAiSEE* mostraram que o *FaceNet* é o modelo mais adequado para aplicações clínicas, devido à sua precisão e resiliência em condições variadas. Na detecção de emoções básicas, o *FaceNet* obteve uma leve vantagem em acurácia e *F1-Score*. Em imagens não tratadas da base *LFW*, o modelo se mostrou mais robusto a ruídos, com menor *MSE* e *MAE* em relação ao *DeepFace*. Além disso, na análise de comportamentos afetivos com a base *DAiSEE*, o *FaceNet* demonstrou maior capacidade de identificar mudanças emocionais dinâmicas, essencial para monitoramento clínico de estados como distração e engajamento.

A contribuição deste estudo está em destacar a importância da escolha do modelo para o reconhecimento facial em contextos clínicos, oferecendo uma avaliação comparativa detalhada que auxilia na decisão entre *FaceNet* e *DeepFace* para diferentes condições de uso. Futuras pesquisas podem explorar ajustes nos *embeddings* do *FaceNet* e melhorias no pré-processamento do *DeepFace*, além do desenvolvimento de bases de dados mais realistas que simulam ambientes clínicos. Tais avanços podem ampliar a eficácia desses modelos em saúde mental e educação, promovendo diagnósticos mais precisos e monitoramento comportamental em tempo real.

REFERÊNCIAS

ANDRÉS, Noboa; GONZALEZ, Omar; FREDDY, Tapia. Use of the Student Engagement as a Strategy to Optimize Online Education, Applying a Supervised Machine Learning Model Using Facial Recognition. *In*: APPLIED TECHNOLOGIES. International Conference On Applied Technologies, 4., 2022, Quito, Equador. **Proceedings** [...]. Quito, Equador: Springer. 2022, p. 283–295.

BEGAJ, Sabrina; TOPAL, Ali Osman; ALI, Maaruf. Emotion recognition based on facial expressions using convolutional neural network (CNN). *In*: INTERNATIONAL CONFERENCE ON COMPUTING, NETWORKING, TELECOMMUNICATIONS &

ENGINEERING SCIENCES APPLICATIONS (CoNTESA), 2020, Tirana, Albania. **Proceedings** [...]. Tirana, Albania: IEEE, 2020, p. 58–63.

CHICCO, Davide; WARRENS, Matthijs J.; JURMAN, Giuseppe. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in evaluating the accuracy of regression analysis. **PeerJ Computer Science,** v. 7, e623, 2021. DOI: 10.7717/peerj-cs.623.

COWEN, Alan S. *et al*. Sixteen facial expressions occur in similar contexts worldwide **Nature,** v. 589, p. 251–257, 2021.

COWEN, Alan S.; KELTNER, Dacher; SCHROFF, Florian; JOU, Brendan; ADAM, Hartwig; PRASAD, Gautam. Sixteen facial expressions occur in similar contexts worldwide. **Nature**, London, v. 589, n. 7841, p. 251–257, 2021.

DU, Hang *et al*. The elements of end-to-end deep face recognition: A survey of recent advances. **Computing Surveys**, New York, v. 54, n. 10, p. 142, 2022.

DU, S.; MARTINEZ, A. M. Expressões faciais compostas de emoção: da investigação básica às aplicações clínicas. **Diálogos em Neurociência Clínica**, v. 17, n. 4, p. 443–455, 2015.

FLYNN, Maria *et al*. Assessing the Effectiveness of Automated Emotion Recognition in Adults and Children for Clinical Investigation. **Frontiers in Human Neuroscience,** v. 14, 2020. ISSN: 1662-5161. DOI: 10.3389/fnhum.2020.00070. URL: https://www.frontiers.org/journals/human-neuroscience/articles/10.3389/fnhum.2020.0007.

GUO, Rufang *et al*. Desenvolvimento e aplicação da tecnologia de reconhecimento de emoções - uma revisão sistemática da literatura. **BMC Psychology,** v. 12, n. 95, 2024. DOI: 10.1186/s40359-024-01581-4.

GUPTA, Mayank; KHURANA, Priyal; GUPTA, Nihit. A Critical Review of Applied Behavior Analysis (ABA): Trends & Gaps. **Prerpints.org**, 2024.

KHAN, Amjad Rehman. Facial emotion recognition using conventional machine learning and deep learning methods: current achievements, analysis and remaining challenges. **Information,** v. 13, n. 6, p. 268, 2022.

LEONARDI, Jan Luiz; RUBANO, Denize Rosana. Fundamentos empíricos da análise do comportamento aplicada para o tratamento do transtorno do déficit de atenção e hiperatividade (TDAH). **Perspectivas em Análise do Comportamento**, v. 3, n. 1, p. 1-19, 2012.

LIU, Shuhua *et al*. Speech emotion recognition based on transfer learning from the FaceNet framework. **The Journal of the Acoustical Society of America,** v. 149, n. 2, p. 1338–1345, 2021.

MELLOUK, Wafa, HANDOUZI, Wahida. CNN-LSTM for automatic emotion recognition using contactless photoplythesmographic signals. **Biomedical Signal Processing and Control**, v.85, p. 104907, 2023.

ONYEMA, Edeh Michael *et al*. Enhancement of patient facial recognition through deep learning algorithm: ConvNet. **Journal of Healthcare Engineering,** v.1, 5196000, 2021.

ONYEMA, Edeh Michael; SHUKLA, Piyush Kumar; DALAL, Surjeet; MATHUR, Mayuri Neeraj; ZAKARIAH, Mohammed; TIWARI, Basant. Enhancement of patient facial recognition through deep learning algorithm: ConvNet. **Journal of Healthcare Engineering**, v. 2021, n. 1, p. 5196000, 2021.

SCHILLER, Devon. The face and the faceness: Iconicity in the early faciasemiotics of Paul Ekman, 1957–1978. **Σημειωτ κ´η-Sign Systems Studies,** v. 49, n. 3-4, p. 361–382, 2021.

SCHWAM JUNIOR, J. G. **Capacidade de reconhecimento facial de emoções em pacientes com espasmo hemifacial**. 2018. Dissertação (Mestrado em Neurociências) - Faculdade de Medicina de Ribeirão Preto, Ribeirão Preto, 2018. Disponível em: https://doi.org/10.11606/D.17.2019.tde-30052019-163537.

TAIGMAN, Yaniv *et al*. Deepface: Closing the gap to human-level performance in face verification. *In*: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2014, Columbus, OH, EUA. **Proceedings** [...]. Columbus, OH, EUA: IEEE, 2014. p. 1701-1708.

YOLCU, G. *et al*. Reconhecimento da expressão facial para monitoramento de distúrbios neurológicos com base na rede neural convolucional. **Ferramentas Multimédias**, v. 78, p. 31581–31603, 2019.

ZHAO, Yue *et al*. Distilling vision-language models on millions of videos. *In*: CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2024, Seattle, WA, EUA. **Proceedings** [...]. Seattle, WA, EUA: IEEE, 2024, p. 13106–13116.