

# Reconhecimento de Emoções em Imagens Utilizando Redes Neurais Convolucionais

1<sup>st</sup> Felipe Natan Zanqueta Macaúbas

*Universidade Tecnológica Federal do Paraná - (UTFPR)*

*Apucarana, Paraná, Brasil*

*felipenatan@alunos.utfpr.edu.br*

2<sup>nd</sup> Michael Pariz Pereira

*Universidade Tecnológica Federal do Paraná - (UTFPR)*

*Apucarana, Paraná, Brasil*

*michaelpariz@alunos.utfpr.edu.br*

3<sup>rd</sup> Nicolas de Paulo Romano

*Universidade Tecnológica Federal do Paraná - (UTFPR)*

*Apucarana, Paraná, Brasil*

*nicolasromano@alunos.utfpr.edu.br*

**Abstract**—Este trabalho apresenta o desenvolvimento de um sistema de reconhecimento de emoções baseado em visão computacional utilizando redes neurais convolucionais (CNN). O modelo foi treinado com o dataset público FER2013, composto por imagens faciais em escala de cinza rotuladas em sete classes emocionais básicas. A metodologia inclui tratamento e análise exploratória do dataset, normalização das imagens, aplicação de técnicas de data augmentation, definição de uma arquitetura convolucional profunda e avaliação quantitativa por meio de acurácia, matriz de confusão e análise qualitativa de erros. O melhor modelo alcançou acurácia de 60,94% no conjunto de teste do FER2013, desempenho compatível com CNNs de complexidade semelhante treinadas do zero nesse conjunto de dados. Por fim, o classificador foi integrado a uma aplicação em tempo real utilizando webcam, demonstrando a viabilidade de uso em cenários práticos como interfaces afetivas e monitoramento comportamental.

**Index Terms**—Convolutional Neural Networks, Emotion Recognition, FER2013, Deep Learning, Computer Vision.

## I. INTRODUÇÃO

O reconhecimento automático de emoções a partir de expressões faciais tem ganhado relevância significativa nas últimas décadas, impulsionado pelos avanços em visão computacional e aprendizagem profunda. A interpretação de estados emocionais é um componente essencial na interação humano-computador, possibilitando o desenvolvimento de sistemas mais responsivos, naturais e empáticos. Aplicações dessa tecnologia abrangem áreas como monitoramento comportamental, saúde mental, análise de engajamento, jogos, segurança e interfaces afetivas.

A consolidação de redes neurais convolucionais (CNNs) como principal abordagem para tarefas de classificação de imagens permitiu que sistemas de detecção emocional atingissem níveis de desempenho antes inviáveis com métodos tradicionais de extração manual de características. Com sua capacidade de aprender padrões complexos e invariantes em dados visuais, CNNs tornaram-se especialmente adequadas para lidar com detalhamentos sutis das expressões faciais humanas.

Neste trabalho, apresenta-se o desenvolvimento de um modelo baseado em CNN treinado no dataset FER2013, que

contém mais de 35 mil imagens de faces anotadas em sete classes emocionais. O objetivo principal é investigar a eficácia de uma arquitetura convolucional profunda na classificação de emoções em imagens de baixa resolução, analisando o impacto de técnicas como normalização, data augmentation e seleção criteriosa de hiperparâmetros. Além disso, é construída uma aplicação prática capaz de reconhecer emoções em tempo real utilizando a webcam, demonstrando a viabilidade do sistema em cenários reais.

## II. TRABALHOS RELACIONADOS

A literatura sobre reconhecimento de emoções faciais tem evoluído substancialmente desde os trabalhos clássicos de Ekman e Friesen, que demonstraram evidências de universalidade em um conjunto de emoções básicas, alegria, tristeza, medo, raiva, surpresa, nojo e estado neutro [1]. Inicialmente, os sistemas automáticos baseavam-se em extração manual de características, utilizando descritores como Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP) e Eigenfaces combinados com classificadores como SVM ou k-Nearest Neighbors. Embora úteis, essas abordagens apresentam limitações quando confrontadas com variações de iluminação, pose, oclusões e ruído.

Com o avanço do aprendizado profundo, redes neurais convolucionais passaram a dominar esta área, pois aprendem representações discriminativas diretamente dos pixels da imagem. O dataset FER2013, introduzido no desafio “Challenges in Representation Learning” da ICML 2013, tornou-se um dos principais benchmarks para avaliação desses modelos [2]. Trabalhos baseados em CNNs profundas e ensembles, como os de Pramerdorfer e Kampel [3] e Khaireddin e Chen [4], reportam acurácias entre 70% e 75% no FER2013 ao empregar arquiteturas modernas (por exemplo, VGGNet) e estratégias de fine-tuning.

Revisões recentes de literatura evidenciam um crescimento contínuo de aplicações de deep learning em reconhecimento de emoções faciais, bem como desafios ainda em aberto, tais como desbalanceamento de classes, robustez a variações demográficas e implementação em tempo real [5],

[6]. Neste contexto, o presente estudo se posiciona como uma implementação educacional de uma CNN de média complexidade, treinada do zero no FER2013, enfatizando a clareza da metodologia, a análise detalhada de erros e a demonstração de uma aplicação prática em webcam.

### III. METODOLOGIA DE DESENVOLVIMENTO

O desenvolvimento do sistema proposto segue um pipeline estruturado que inclui carregamento e preparação do dataset, processamento das imagens, definição da arquitetura da CNN, estratégias de treinamento e avaliação final.

Como ponto de partida para a implementação, foi utilizada a estrutura de código apresentada no notebook público “Emotion Detection using CNN” [7], disponibilizado no Kaggle. O código original foi adaptado, reorganizado e amplamente comentado, de forma a alinhar a arquitetura da rede, o fluxo de pré-processamento e os procedimentos de treinamento aos objetivos específicos da disciplina e às definições metodológicas descritas nesta seção.

#### A. Conjunto de Dados

Utilizou-se o dataset FER2013, disponibilizado no Kaggle, composto por 35.887 imagens faciais em escala de cinza, com resolução de 48x48 pixels [2]. As amostras estão divididas em sete categorias emocionais (raiva, nojo, medo, feliz, triste, surpreso e neutro) e organizadas em três subconjuntos: 28.709 imagens para treinamento, 3.589 para validação (PublicTest) e 3.589 para teste (PrivateTest). O arquivo original em formato CSV já fornece essa divisão, bem como os rótulos numéricos das emoções para cada exemplo.

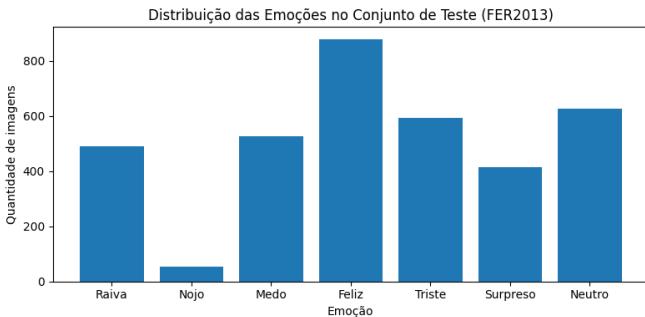


Fig. 1. Distribuição das emoções no conjunto de teste do dataset FER2013.

A Figura 1 apresenta a distribuição das emoções no conjunto de teste do FER2013. Nota-se que as classes estão desbalanceadas, com predominância das categorias “Feliz”, “Triste” e “Medo”, enquanto emoções como “Nojo” apresentam quantidade significativamente inferior de amostras. Esse desbalanceamento impacta diretamente o desempenho do modelo, especialmente nas classes minoritárias, que tendem a possuir menor representatividade durante o processo de treinamento. A análise dessa distribuição é fundamental para interpretar corretamente os resultados obtidos nas avaliações posteriores.

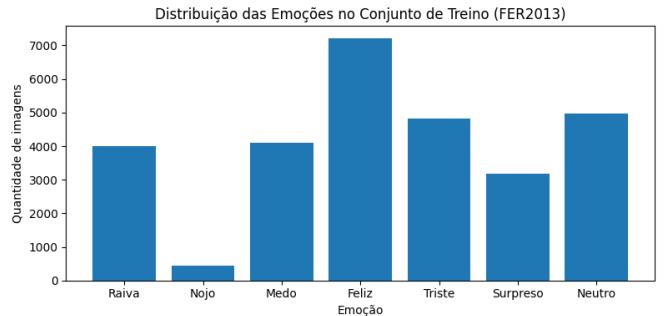


Fig. 2. Distribuição das emoções no conjunto de treino do dataset FER2013.

A Figura 2 mostra a distribuição das emoções no conjunto de treino, onde se observa o mesmo padrão de desbalanceamento identificado no conjunto de teste. A classe “Feliz” possui o maior número de amostras, enquanto “Nojo” continua sendo a categoria menos representada. Esse aspecto justifica a utilização de técnicas de data augmentation, que visam minimizar os efeitos do desbalanceamento ao fornecer maior variabilidade às classes com menor quantidade de exemplos. Essa desigualdade nas amostras por classe também contribui para explicar diferenças de desempenho entre as emoções avaliadas.



Fig. 3. Exemplos de cada uma das sete classes de emoção do dataset FER2013 após o pré-processamento.

A Figura 3 apresenta um exemplo representativo de cada classe de emoção presente no FER2013: raiva, nojo, medo, felicidade, tristeza, surpresa e neutro. Mesmo com resolução reduzida (48x48 pixels) e em escala de cinza, observa-se que traços faciais como abertura da boca, contração das sobrancelhas e formato dos olhos permanecem visíveis, o que justifica a aplicação de redes convolucionais para extração automática dessas características.

#### B. Pré-processamento

As imagens são inicialmente convertidas de strings contendo 2304 valores para matrizes de 48x48. Em seguida, são normalizadas para o intervalo [0, 1] e reorganizadas para o formato (48, 48, 1), compatível com camadas convolucionais. Os rótulos são convertidos para codificação one-hot.

#### C. Data Augmentation

A fim de reduzir overfitting e aumentar a robustez do modelo, aplicou-se data augmentation incluindo:

- rotações suaves
- deslocamento horizontais e verticais
- zoom parcial
- espelhamento horizontal

Essas transformações ajudam o modelo a generalizar para variações naturais da face humana, melhorando a acurácia em dados não vistos.

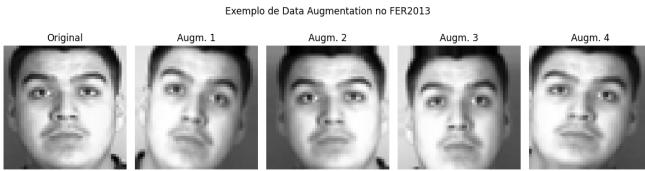


Fig. 4. Exemplo de data augmentation aplicado a uma imagem do FER2013: versão original e quatro variações geradas (rotações, translações e zoom).

A Figura 4 apresenta um exemplo de data augmentation aplicado às imagens do FER2013. A partir de uma face original, são geradas variações com pequenas rotações, translações e alterações de escala. Essas transformações simulam variações naturais de pose e enquadramento, aumentando a diversidade do conjunto de treino sem alterar a emoção subjacente. Na prática, essa estratégia contribuiu para reduzir overfitting e melhorar a capacidade de generalização do modelo, como observado no comportamento das curvas de treino e validação.

#### D. Arquitetura da CNN

A rede neural desenvolvida é composta por três blocos convolucionais, cada um contendo camadas Conv2D seguidas de Batch Normalization, MaxPooling e Dropout. Após a extração de características, utiliza-se Flatten, seguido por uma camada totalmente conectada com 256 neurônios e, por fim, uma camada Softmax com sete saídas correspondentes às emoções.

A estrutura foi projetada para equilibrar profundidade, capacidade de generalização e custo computacional.

#### E. Treinamento

O treinamento do modelo foi realizado por até 40 épocas, com batch size igual a 64 e otimizador Adam, utilizando a taxa de aprendizado padrão da implementação do Keras. Para evitar overfitting, foram empregados dois callbacks: EarlyStopping, monitorando a acurácia de validação com patience de 5 épocas e restaurando automaticamente os melhores pesos, e ModelCheckpoint, responsável por salvar em disco o modelo com maior *val\_accuracy*. A função de perda adotada foi a categorical cross-entropy e as métricas monitoradas foram acurácia em treino e validação. A validação foi conduzida utilizando o subconjunto PublicTest, enquanto o desempenho final do modelo foi medido apenas no subconjunto PrivateTest, reservado para teste.

## IV. RESULTADOS E DISCUSSÕES

As curvas de acurácia e perda apresentadas na Figura 5 demonstram um comportamento consistente de aprendizado ao longo das épocas. A acurácia de validação evoluiu de aproximadamente 2% na primeira época para valores em torno de 60% nas últimas épocas, enquanto a função de perda de validação foi reduzida de cerca de 1,8 para pouco acima

de 1,1. Esse comportamento indica que o modelo conseguiu aprender padrões relevantes do dataset sem apresentar divergência acentuada entre treino e validação, o que sugere que o overfitting foi mitigado pelo uso combinado de Dropout, Batch Normalization e data augmentation.

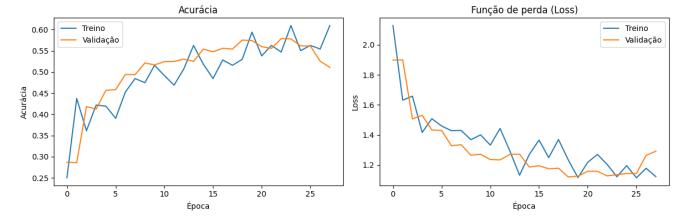


Fig. 5. Evolução da acurácia e da função de perda durante o treinamento e validação do modelo CNN ao longo das épocas.

O melhor conjunto de pesos, selecionado pelo critério de maior acurácia de validação, foi posteriormente avaliado no conjunto de teste (PrivateTest). Nesse conjunto, a CNN alcançou acurácia global de 60,94%. Considerando a simplicidade relativa da arquitetura e o treinamento realizado inteiramente a partir do zero, esse resultado é compatível com o intervalo de desempenho reportado em trabalhos que treinam CNNs de porte semelhante no FER2013, tipicamente entre 55% e 65% de acurácia [3], [5].

A matriz de confusão da Figura 6 evidencia que as classes “Feliz” e “Surpreso” apresentam as maiores taxas de acerto, o que é esperado, pois essas expressões tendem a ser mais intensas e visualmente distintas. Em contraste, emoções como “Medo”, “Raiva” e, principalmente, “Nojo” apresentam maior número de confusões. Parte dessa dificuldade está associada ao desbalanceamento do conjunto de dados, em que a classe “Nojo” contém significativamente menos exemplos em relação às demais, além da semelhança visual entre algumas expressões em imagens de baixa resolução.

A Figura 7 ilustra exemplos de classificações corretas e incorretas do modelo. Observa-se que erros são especialmente frequentes em imagens com iluminação desfavorável, expressões de baixa intensidade ou presença de oclusões parciais do rosto (mãos, objetos, cabelos). Há também casos em que a própria anotação do dataset parece ambígua, o que limita o desempenho máximo possível de qualquer modelo treinado apenas com essas amostras.

Além das avaliações em dados estáticos, o modelo foi integrado a uma aplicação de captura de vídeo em tempo real utilizando a biblioteca OpenCV. Cada frame é convertido para tons de cinza, redimensionado para 48x48 pixels, normalizado e, em seguida, classificado pela CNN. A emoção com maior probabilidade é sobreposta ao vídeo em tempo real. Testes informais com diferentes pessoas e condições de iluminação indicaram que o sistema responde em tempo hábil e é capaz de identificar corretamente emoções mais expressivas, como “Feliz” e “Surpreso”, embora mantenha as mesmas dificuldades observadas no conjunto de teste para emoções mais sutis.

A Figura 6 apresenta a matriz de confusão obtida no conjunto de teste. Nota-se que a maior parte das amostras

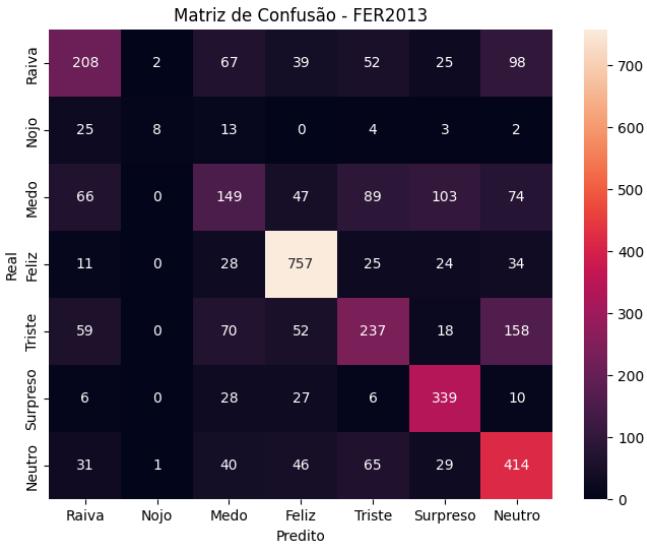


Fig. 6. Matriz de confusão do modelo CNN no conjunto de teste do FER2013. As linhas representam os rótulos reais e as colunas, as classes preditas.

concentra-se na diagonal principal, indicando boa capacidade de acerto do modelo. As emoções Feliz e Surpreso foram as mais bem classificadas, com elevado número de acertos e baixa taxa de confusão com outras classes. Por outro lado, observa-se maior confusão entre as emoções Raiva, Medo e Neutro, o que é esperado, uma vez que essas expressões podem ser mais sutis ou visualmente semelhantes em imagens de baixa resolução.

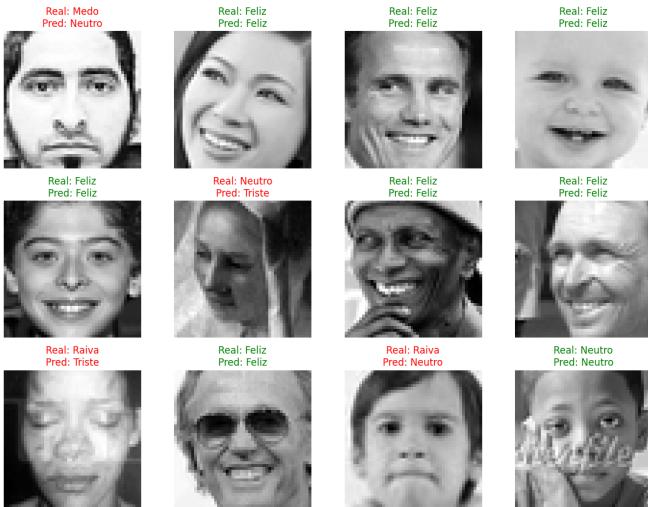


Fig. 7. Exemplos de previsões do modelo no conjunto de teste. O rótulo real é indicado na linha superior e o rótulo predito, na inferior; textos em verde representam acertos e em vermelho, erros.

A Figura 7 ilustra exemplos qualitativos de previsões do modelo. As imagens mostram casos tanto de acerto quanto de erro, destacando em verde as classificações corretas e em vermelho aquelas incorretas. Observa-se que muitos erros ocorrem em expressões de Medo e Neutro classificadas como

Feliz ou Triste, o que reforça a dificuldade do modelo em distinguir emoções de menor intensidade ou com pouca ativação muscular facial. Esse tipo de análise visual complementa a matriz de confusão e evidencia limitações relevantes do modelo em cenários de fronteira entre classes.

#### Aplicação em Tempo Real

Para validar a aplicabilidade do modelo, foi desenvolvido um script em Python utilizando a biblioteca OpenCV para captura de vídeo e a API keras para o carregamento do modelo treinado.

- 1) **Aquisição:** O sistema captura o fluxo de vídeo da webcam padrão quadro a quadro
- 2) **Detecção Facial:** Utiliza-se o classificador em cascata de Haar (*Haar Cascade Classifier*) com os pesos pré-treinados para identificar as coordenadas das faces presentes na imagem.
- 3) **Pré-processamento:** Para cada face detectada, a Região de Interesse (ROI) é recortada e convertida para escala de cinza. Em seguida, aplica-se redimensionamento para  $48 \times 48$  pixels e normalização dos valores de pixel para o intervalo  $[0, 1]$ , garantindo compatibilidade com a camada de entrada da CNN.
- 4) **Inferência:** A matriz processada é submetida ao modelo .h5, que retorna um vetor de sete probabilidades. A classe com o maior valor é selecionada como a predição final.
- 5) **Interface:** A emoção predita e a probabilidade são desenhadas sobre o vídeo original em tempo real, juntamente com uma caixa delimitadora ao redor da face

Os testes demonstraram que a arquitetura leve da CNN permitiu uma execução fluida, sem latência perceptível, validando o sistema para aplicações de interação humano-computador em computadores pessoais convencionais.

#### V. CONCLUSÃO

Este estudo apresentou o desenvolvimento de um modelo baseado em redes neurais convolucionais para reconhecimento de emoções faciais utilizando o dataset FER2013. A arquitetura proposta, composta por três blocos convolucionais seguidos de uma camada densa de 256 neurônios, aliada a técnicas de normalização, Dropout e data augmentation, permitiu atingir acurácia de 60% no conjunto de teste, com melhor desempenho nas classes “Feliz” e “Surpreso”.

Além do treinamento e avaliação offline, foi desenvolvida uma aplicação que realiza a classificação em tempo real a partir de uma webcam, evidenciando o potencial prático do sistema em aplicações como interfaces afetivas, monitoramento e análise comportamental. O fluxo de processamento captura do frame, pré-processamento, inferência com a CNN e exibição da emoção prevista mostrou-se suficientemente leve para execução em tempo real em um computador de uso geral.

Apesar dos resultados satisfatórios, algumas limitações foram identificadas, sobretudo nas classes menos representadas e em expressões de baixa intensidade, como “Nojo” e “Medo”. Como trabalhos futuros, destacam-se: (i) exploração

de arquiteturas pré-treinadas (por exemplo, VGGNet e ResNet) via transferência de aprendizado; (ii) uso de técnicas adicionais para tratar o desbalanceamento, como ponderação de classes, focal loss ou oversampling; e (iii) inclusão de etapas de detecção e alinhamento facial mais robustas antes da classificação. Essas extensões têm potencial para aproximar o desempenho do sistema do estado da arte reportado na literatura [3]–[5].

#### REFERENCES

- [1] P. Ekman and W. V. Friesen, “Constants across cultures in the face and emotion,” *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, 1971.
- [2] I. J. Goodfellow, D. Erhan, P. L. Carrier *et al.*, “Challenges in representation learning: A report on three machine learning contests,” *Neural Networks*, vol. 64, pp. 59–63, 2015.
- [3] C. Pramerdorfer and M. Kampel, “Facial expression recognition using convolutional neural networks: State of the art,” *arXiv preprint arXiv:1612.02903*, 2016.
- [4] Y. Khairuddin and Z. Chen, “Facial emotion recognition: State of the art performance on FER2013,” *arXiv preprint arXiv:2105.03588*, 2021.
- [5] M. Sajjad, S. Hussain, A. Ullah *et al.*, “A comprehensive survey on deep facial expression recognition: Challenges, applications, and future guidelines,” *Alexandria Engineering Journal*, vol. 68, no. 6, pp. 817–840, 2023.
- [6] R. Rajvanshi and K. Thakar, “A survey on facial expression recognition using deep learning approach,” 2020, unpublished.
- [7] O. Eravci, “Emotion Detection using CNN,” Kaggle, 2021. [Online]. Available: <https://www.kaggle.com/code/oykuer/emotion-detection-using-cnn/notebook>.