Reconhecimento de expressões faciais com MediaPipe

Daniel Squinalli Casanova[1],Pedro Luiz de Paula Filho[1], Kelyn Schenatto[1], Ricardo Sobjak[2],
Universidade Tecnológica Federal do Paraná - UTFPR, Medianeira, Brasil[1],
Universidade Federal do Paraná - UFPR - Curitiba, Brasil[2],
Email: danielcasanova@alunos.utfpr.edu.br, pedrol@utfpr.edu.br, kelynschenatto@gmail.com, ricardosobjak@utfpr.edu.br,

***Abstract*—Facial Expression Recognition (FER) is an important subfield of computer vision and artificial intelligence, with applications ranging from human-computer interaction to emotional monitoring in clinical contexts. Despite advances, most studies focus on analyses based on complete images, underestimating the viability of facial landmarks as an alternative that balances privacy and computational efficiency. This paper proposes and compares three different approaches: (1) Complete Images; (2) Rendered Landmarks; and (3) Vectorized Landmarks. The results indicate that, although the Complete Images approach achieved the best performance with an F1-Score of 0.6723, Precision of 0.672, and Recall of 0.676, demonstrating the robustness of this technique, the Rendered Landmarks, especially with the Connected Points Map (MPC), emerge as a promising alternative, balancing accuracy, efficiency, and privacy.**

***Keywords*—Emotion Detection; Facemesh; Facial Landmarks.**

***Resumo*—O reconhecimento de expressões faciais (Facial Expression Recognition - FER) é uma importante subárea da visão computacional e inteligência artificial, com aplicações que variam desde a interação humano-computador até o monitoramento emocional em contextos clínicos. Apesar dos avanços, a maioria dos estudos se concentra em análises baseadas em imagens completas, subestimando a viabilidade dos landmarks faciais como uma alternativa que equilibra privacidade e eficiência computacional. Este artigo propõe e compara três abordagens diferentes: (1) Imagens Completas; (2) Landmarks Renderizados; e (3) Landmarks Vetoriais. Os resultados indicam que, embora a abordagem de Imagens Completas tenha obtido o melhor desempenho em F1-Score: 0,6723, Precision: 0,672 e Recall: 0,676, demonstrando a robustez desta técnica. Contudo, os Landmarks Renderizados, especialmente com o Mapa de Pontos Conectados (MPC), emergem como uma alternativa promissora, equilibrando precisão, eficiência e privacidade.**

***Palavras-chave*—Detecção de Emoções; Facemesh; Landmarks Faciais.**

I. Introdução

O reconhecimento de expressões faciais (*Facial Expression Recognition - FER*) é uma subárea da visão computacional e da inteligência artificial.Seu foco está na identificação automática das expressões humanas [1]. As expressões faciais são elementos fundamentais da comunicação não verbal, desempenhando um papel importante na interação social e na transmissão de emoções. A possibilidade de reconhecer essas expressões de forma automática abre portas para uma vasta gama de aplicações, incluindo desde a interação humano-computador até o monitoramento de estados emocionais em contextos clínicos e educacionais.

O estudo de técnicas de aprendizado profundo tem proporcionado melhorias significativas no reconhecimento de expressões faciais. Ferramentas como redes neurais convolucionais (CNNs) têm sido amplamente utilizadas para extrair e classificar características faciais, permitindo que sistemas sejam capazes de identificar emoções com uma precisão que, em condições controladas, muitas vezes ultrapassa o desempenho humano [2], [3]. Esses avanços possibilitaram o desenvolvimento de aplicações em áreas como interação humano-computador e monitoramento de emoções.. De acordo com [4], a utilização de técnicas de reconhecimento de expressões faciais tem sido cada vez mais explorada como uma solução eficiente para problemas reais em diversas áreas, como por exemplo, a tele-reabilitação. Estudos recentes demonstram que a integração da Inteligência Artificial Emocional com sistemas de reabilitação pode melhorar significativamente o acompanhamento remoto do estado emocional e cognitivo dos pacientes, contribuindo para tratamentos mais personalizados e eficazes.

A construção de sistemas robustos de reconhecimento de expressões faciais enfrenta desafios significativos devido à variabilidade das características faciais, como idade, gênero e etnia. Esses fatores introduzem um nível adicional de complexidade, pois as expressões faciais podem se manifestar de maneiras distintas dependendo dessas variáveis. Conforme mencionado por [5], essa variabilidade afeta diretamente o desenvolvimento de sistemas eficazes, exigindo abordagens que considerem as diferentes manifestações das expressões faciais em diversos grupos demográficos.

Além disso, a proteção da privacidade tornou-se uma preocupação central na coleta e armazenamento de dados faciais, especialmente com a implementação da Lei Geral de Proteção de Dados (LGPD) no Brasil. A LGPD, estabelecida

pela Lei nº 13.709 de 2018 [6], impõe diretrizes rigorosas para o tratamento de dados pessoais, proibindo, por exemplo, o armazenamento de imagens faciais sem o consentimento explícito dos indivíduos, a fim de proteger a privacidade e os direitos dos titulares dos dados.

Uma das ferramentas que têm se destacado no contexto do reconhecimento de expressões faciais é o MediaPipe[1]. Desenvolvido pelo Google, o MediaPipe é uma biblioteca de código aberto que facilita a implementação de pipelines para a detecção e rastreamento de *landmarks* faciais em tempo real. *Landmarks* são pontos de referência específicos no rosto, como os contornos dos olhos, boca e nariz, usados para mapear a estrutura facial com precisão [7]. A partir daqui, utilizaremos o termo *landmarks* para nos referirmos a esses pontos de referência faciais.. Sua capacidade de extrair pontos faciais detalhados, como os contornos dos olhos, boca e nariz, o torna uma escolha eficiente para aplicações que requerem precisão na análise de expressões faciais. Essa ferramenta permite que apenas os *landmarks* sejam armazenados, em vez de imagens completas, o que oferece uma vantagem significativa em termos de privacidade.

Um exemplo de aplicação prática dessa tecnologia é a detecção de dor em crianças com autismo, um processo notoriamente complexo devido às dificuldades de comunicação inerentes ao transtorno. Nessas situações, o uso de modelos de inteligência artificial surge como uma abordagem promissora para identificar expressões faciais de dor de forma eficaz [7]. Nesse contexto, este artigo tem como objetivo comparar a eficácia dos diferentes tipos de *landmarks* extraídos pelo MediaPipe com as imagens completas no reconhecimento de expressões faciais. Essa comparação permitirá uma análise detalhada sobre a viabilidade de substituir imagens faciais completas por *landmarks* no desenvolvimento de sistemas de reconhecimento de expressões faciais, visando soluções mais seguras e eficientes, com respeito à privacidade dos indivíduos.

II. Metodologia

Neste estudo, foi utilizada a técnica de *Transfer Learning* aplicada à rede neural convolucional ResNet50, uma escolha estratégica para o reconhecimento de expressões faciais. A ResNet50, conforme discutido por [8], é uma rede profunda composta por 50 camadas que utiliza uma arquitetura de aprendizado residual, o que permite um treinamento mais eficiente em redes profundas.

O *Transfer Learning*, conforme descrito por [9], aproveita modelos previamente treinados em grandes bases de dados, como o ImageNet[2], que contém milhões de imagens categorizadas, facilitando a adaptação a novas tarefas com menos dados de treinamento. Tal afirmação está representada na Figura 4, que ilustra como o *Transfer Learning* utiliza os conhecimentos adquiridos de uma base de dados extensa para melhorar o desempenho em uma tarefa específica.

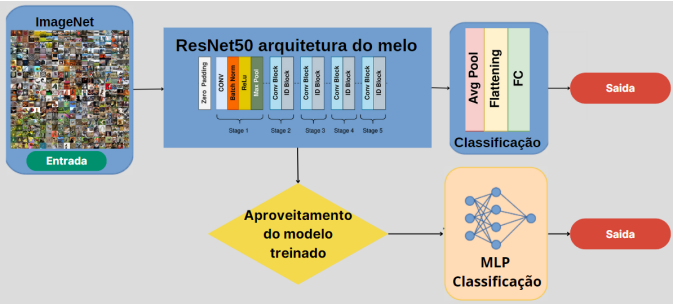

Fig. 1. *Transfer Learning* aplicado ao reconhecimento de expressões faciais.

A ResNet50 foi ajustada para o processamento de imagens completas e *landmarks* extraídos pelo MediaPipe. Para otimizar a precisão na detecção de emoções faciais, o modelo foi pré-treinado com o ImageNet e, em seguida, adaptado ao problema específico de reconhecimento facial. Foram utilizados os seguintes parâmetros durante o treinamento: as últimas quatro camadas da ResNet50 foram descongeladas, permitindo sua atualização durante o treinamento, enquanto as demais foram mantidas congeladas para preservar os conhecimentos pré-adquiridos. A última camada *fully connected* da ResNet50 foi substituída por uma sequência de camadas personalizadas: uma camada linear de 512 neurônios, seguida por uma ativação ReLU e uma camada de *dropout* com taxa de 0,5, antes de finalizar com uma camada linear ajustada para o número de classes do problema e uma função *LogSoftmax*. A função de erro utilizada foi a *NLLLoss* (*Negative Log Likelihood Loss*), combinada com o otimizador Adam, com uma taxa de aprendizado inicial definida em 0,0001. O modelo foi treinado por 20 épocas, e a melhor performance foi monitorada durante o treinamento para salvar o modelo com a maior acurácia de validação. Essa estratégia permitiu um ajuste fino do modelo, maximizando a precisão e minimizando o risco de *overfitting*.

Para uma das abordagens metodológicas, utilizou-se uma rede MLP (*Multi-Layer Perceptron*) para classificar os *landmarks* extraídos pelo MediaPipe. As MLPs são redes neurais amplamente utilizadas em tarefas de classificação devido à sua capacidade de aprender padrões complexos em dados. Para uma descrição mais detalhada da arquitetura e funcionamento das MLPs, consulte o artigo de [10]. A MLP foi configurada para receber os vetores de pontos dos *landmarks* como entrada, permitindo que realizasse a classificação das expressões faciais com base nesses dados.

[1]https://ai.google.dev/edge/mediapipe/solutions/guide?hl=pt-br
[2]https://paperswithcode.com/dataset/imagenet

O desenvolvimento foi realizado com a linguagem Python e as bibliotecas PyTorch e MediaPipe, ambas gratuitas e de código aberto, assegurando a acessibilidade e a reprodução do estudo. Os experimentos foram conduzidos em um ambiente com um processador Intel Core i5, 16 GB de RAM e uma GPU NVIDIA GeForce GTX 1660, proporcionando o desempenho necessário para a execução eficiente dos modelos.

Para facilitar a compreensão do processo metodológico, a Figura 2 apresenta um fluxograma com as etapas desde a entrada das imagens até a avaliação final dos modelos. As imagens passam por um pré-processamento e seguem por três fluxos distintos: no primeiro, são processadas diretamente pela ResNet50; no segundo, os *landmarks* são extraídos com o MediaPipe, gerando uma imagem (*Facemesh*), antes de serem processados pela ResNet50; e no terceiro, cria-se um vetor de pontos dos *landmarks* e este vetor é inserido em uma MLP.

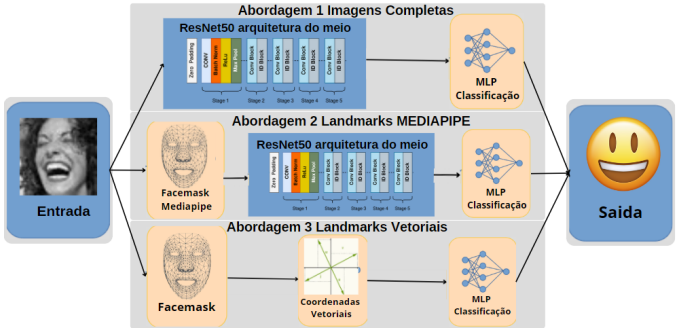

Fig. 2. Exemplificando as abordagens que usaremos com o MediaPipe.**Adaptado de [11]**

Para avaliar o desempenho dos modelos, foram utilizadas as métricas F1-Score, Precision, Recall e AUC-ROC. Essas métricas foram escolhidas por sua capacidade de fornecer uma avaliação abrangente e robusta, especialmente em cenários com classes desbalanceadas, como é o caso do dataset FER-2013. Estudos, como os de [12] e [13], discutem a importância dessas métricas na avaliação de modelos de classificação. A combinação dessas métricas permite captar nuances do desempenho do modelo, garantindo uma análise mais precisa e confiável.

Para o treinamento e teste dos modelos, foi utilizado o dataset FER-2013[3]. Este conjunto de dados é composto por 35.887 imagens de rostos em escala de cinza, com resolução de 48x48 pixels, nas quais as faces estão centralizadas, ocupando aproximadamente a mesma área em cada imagem. O objetivo do desafio proposto pelo dataset é classificar cada face em uma das sete categorias de emoções, que originalmente são: *Angry* (4953), *Disgust* (547), *Fear* (5121), *Happy* (8989), *Sad* (6077), *Surprise* (4002), e *Neutral* (6198). No entanto, devido à quantidade insuficiente de exemplos na categoria *Disgust*, esta classe foi removida para manter a consistência e a robustez do modelo. A exclusão dessa classe, embora necessária, pode impactar a capacidade do modelo de generalizar adequadamente para todas as expressões faciais, especialmente em contextos em que a expressão de desgosto é relevante. Essa limitação deve ser considerada ao aplicar o modelo em cenários reais, uma vez que a ausência de uma classe de emoção pode resultar em uma cobertura incompleta das possíveis expressões faciais humanas.

Com isso, o dataset ficou com o total de 35.340 imagens, sendo elas divididas em três subconjuntos: 28.272 imagens para treinamento (80%), 3.534 para validação (10%) e 3.534 para teste (10%).

A Figura 3 apresenta amostras em que cada imagem representa uma das categorias emocionais incluídas na análise, garantindo a diversidade e a representatividade no treinamento e teste dos modelos.

Fig. 3. Exemplos de imagens representativas de cada uma das classes emocionais do conjunto de dados FER-2013. As imagens estão em escala de cinza e organizadas de forma a ilustrar a diversidade de expressões analisadas no estudo. **Adaptado de [11]**.

O processo de classificação e filtragem das imagens iniciou-se com a aplicação da biblioteca MediaPipe, mais especificamente o módulo *Facemesh*, responsável pela detecção e rastreamento de *landmarks* faciais. Durante o pré-processamento, algumas imagens que não permitiam uma detecção adequada dos *landmarks* foram descartadas, assegurando que apenas imagens com precisão razoável na detecção dos pontos faciais fossem mantidas no *dataset* final. Este passo é essencial para garantir que o modelo trabalhe com dados de alta qualidade, minimizando a influência de ruídos e variações indesejadas nos resultados finais.

Na segunda abordagem, que utiliza o MediaPipe, foram

[3]https://ai.google.dev/edge/mediapipe/solutions/guide?hl=pt-br. Acesso em: 22 out. 2024.

empregados três tipos de mapeamentos de pontos (*landmarks*) durante a fase de análise: Mapa de Pontos Simples (MPS), Mapa de Pontos Detalhado (MPD) e Mapa de Pontos Conectados (MPC).A escolha desses mapeamentos foi baseada na necessidade de balancear a precisão e a complexidade do modelo. O MPS foi utilizado para avaliar o desempenho com um número reduzido de pontos focados em áreas-chave da face, como olhos, sobrancelhas e boca, o que reduz o custo computacional. O MPD, por sua vez, inclui um conjunto completo de pontos, capturando um maior nível de detalhe facial, ideal para cenários em que a precisão é importante. O MPC foi escolhido por adicionar conexões entre os pontos do rosto, formando uma malha estrutural que aprimora a representação espacial das expressões, o que potencialmente aumenta a robustez do modelo diante de variações faciais.

Na terceira abordagem, que também utiliza *landmarks* extraídos pelo MediaPipe, foram testados apenas os mapeamentos MPS e MPD. Essa escolha foi baseada em resultados preliminares que indicaram que o MPC, embora mais robusto, aumentava significativamente a complexidade computacional sem oferecer melhorias proporcionais em relação ao MPD. Por essa razão, a terceira abordagem focou em testar a eficácia de uma rede MLP (*Multi-Layer Perceptron*) com os mapeamentos MPS e MPD, priorizando simplicidade e eficiência computacional, enquanto mantinha uma precisão aceitável na classificação das expressões faciais.

Para ilustrar o processo de extração de *landmarks* faciais utilizando o MediaPipe, a Figura 4 ilustra o *pipeline* do *FaceMesh* no MediaPipe, mostrando o processo desde a entrada da imagem facial até a geração de três tipos de mapeamentos de pontos faciais: Mapa de Pontos Simples (MPS), Mapa de Pontos Detalhado (MPD) e Mapa de Pontos Conectados (MPC).

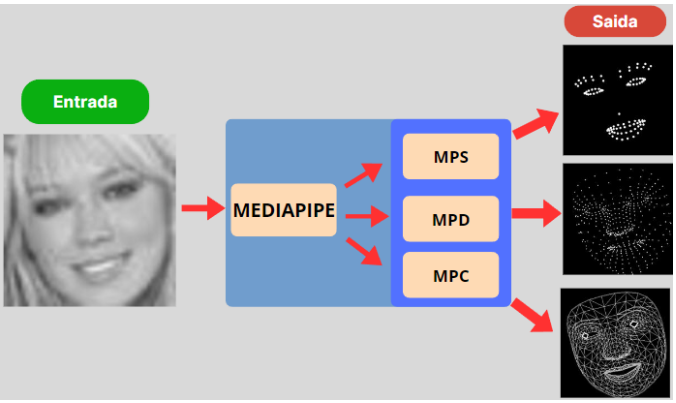

Fig. 4. Exemplo do processamento das imagens com o MediaPipe.

A escolha desses mapeamentos permitiu uma análise abrangente, considerando diferentes níveis de detalhamento na captura dos pontos faciais e sua influência no desempenho dos modelos de reconhecimento de expressões faciais.

III. Resultados

Nas três abordagens testadas, diferentes métodos foram aplicados para processar e classificar as expressões faciais, com o objetivo de avaliar como os níveis de detalhamento dos *landmarks* influenciam a precisão do reconhecimento.

Na primeira abordagem, as imagens completas do *dataset* FER-2013, pós-processadas, foram usadas para o treinamento e teste do modelo ResNet50, após pré-processamento. Esta abordagem serviu como linha de base para comparação com as outras abordagens que utilizam *landmarks*. O modelo treinado com imagens completas obteve um F1-Score de 0,6723, acurácia de 0,676, *Precision* de 0,672, *Recall* de 0,676 e uma *AUC-ROC* de 0,9057, indicando sua eficácia em reconhecer expressões faciais sem o uso de *landmarks* específicos. Estes resultados sugerem que o uso de imagens completas proporciona uma representação mais rica e detalhada das características faciais, resultando em um reconhecimento mais preciso das expressões.

Na segunda abordagem, foram realizados três experimentos distintos, cada um utilizando configurações específicas de *landmarks* extraídos pelo MediaPipe: MPS, MPD e MPC. Esses experimentos visavam avaliar a eficiência de diferentes níveis de detalhamento na detecção de expressões faciais. A Tabela I resume os F1-Scores, *Precision*, *Recall* e *AUC-ROC* obtidos em cada configuração, permitindo uma comparação direta entre as técnicas aplicadas.

TABELA I
Desempenho das Configurações de *Landmarks* na Abordagem 2

Experimentos	F1-Score	Precision	Recall	AUC-ROC
1 - MPS	0,5589	0,5568	0,5693	0,8596
2 - MPD	0,5380	0,5507	0,5471	0,8459
3 - MPC	0,5777	0,5812	0,5781	0,8612

Na terceira abordagem, utilizou-se uma rede MLP (*Multi-Layer Perceptron*) para a classificação dos *landmarks* extraídos pelo MediaPipe. Essa abordagem seguiu os níveis de detalhamento previamente estabelecidos nos experimentos 1 e 2 da Abordagem 2, visando explorar a eficácia da MLP em processar e classificar os diferentes mapeamentos de *landmarks*. A Tabela II abaixo apresenta os resultados obtidos, incluindo F1-Scores, *Precision*, *Recall* e *AUC-ROC*, permitindo uma comparação direta com os resultados das abordagens anteriores.

Para complementar a análise quantitativa apresentada nas Tabelas I e II, a Figura 5 apresenta um gráfico comparativo das três abordagens testadas. Este gráfico facilita a visualização das

TABELA II
DESEMPENHO DA REDE MLP COM *Landmarks* VETORIAIS NA ABORDAGEM 3

Experimentos	F1-Score	Precision	Recall	AUC-ROC
1 - MPS	0,5657	0,5704	0,5756	0,8603
2 - MPD	0,5636	0,5744	0,5761	0,8683

diferenças de desempenho entre as abordagens em termos das métricas F1-Score, Precision, Recall e AUC-ROC, permitindo uma análise mais clara de qual técnica oferece o melhor equilíbrio entre precisão e eficiência.

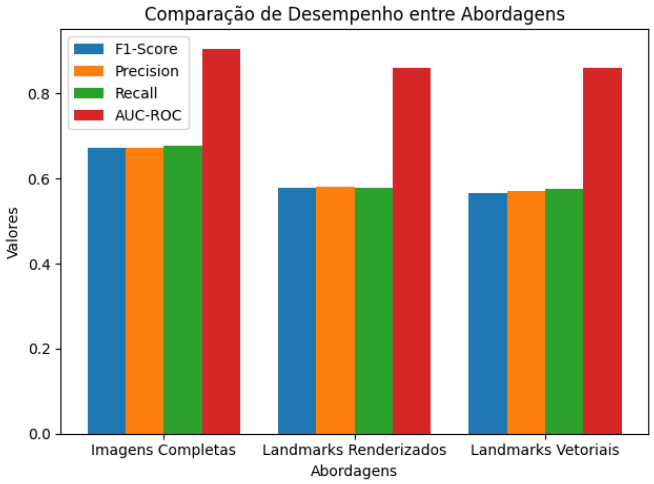

Fig. 5. Comparação de Desempenho das Três Abordagens Testadas

Além das métricas agregadas apresentadas, foram geradas matrizes de confusão para cada abordagem, a fim de fornecer uma análise detalhada das predições feitas pelos modelos em relação às classes reais. As Figuras 6, 7 e 8 mostram as matrizes de confusão dos melhores resultados de cada abordagem.

Ao analisar essas matrizes, observou-se que emoções como *Happy* (feliz) foram identificadas com alta precisão em todas as abordagens, refletindo taxas de acerto superiores a 90%. Essa alta taxa de acerto pode ser atribuída às características faciais distintivas e geralmente mais evidentes associadas a essa emoção, como o sorriso, que são facilmente capturadas pelos modelos.

Embora o reconhecimento de emoções como felicidade tenha apresentado uma alta taxa de acurácia, emoções mais sutis e com características visuais semelhantes, como tristeza e medo, continuam a representar um desafio significativo. A confusão entre essas classes pode ser atribuída à similaridade das expressões faciais envolvidas, o que resulta em uma maior taxa de falsos positivos, principalmente nas abordagens que utilizam *landmarks* renderizados. Para superar esses obstáculos, propomos algumas direções futuras.

Primeiramente, a adoção de modelos híbridos pode ser uma solução eficaz. Esses modelos poderiam combinar tanto informações de *landmarks* faciais quanto imagens completas para capturar melhor as nuances de expressões faciais que compartilham características semelhantes. Ao integrar diferentes tipos de representações de dados, como o uso de convoluções em imagens completas junto com uma análise vetorial detalhada dos *landmarks*, pode-se melhorar a capacidade de discriminar emoções sutis.

Outra possibilidade seria ajustar a arquitetura dos modelos atuais, utilizando camadas adicionais que favoreçam a atenção em regiões faciais específicas, como a área dos olhos e da boca, que são particularmente relevantes para diferenciar expressões de medo e tristeza. Além disso, a aplicação de técnicas de aumento de dados (*data augmentation*), como pequenas variações na iluminação e ângulo de captura, pode ajudar o modelo a generalizar melhor para diferentes variações dessas emoções.

Finalmente, a criação de *datasets* mais balanceados, com um número maior de exemplos para emoções como medo e tristeza, contribuiria para treinar modelos que respondam melhor às sutilezas dessas expressões. Essas abordagens poderiam não apenas melhorar a precisão, mas também aumentar a robustez dos modelos ao enfrentar expressões faciais mais complexas e difíceis de identificar.

Adicionalmente, uma consideração importante seria a otimização da complexidade computacional dos modelos utilizados. O uso de arquiteturas mais leves, como as variantes do *MobileNet* ou *EfficientNet*, pode ser uma alternativa para reduzir o custo computacional sem sacrificar significativamente a precisão. Além disso, técnicas de compressão de rede, como a quantização de pesos ou *pruning*, podem ser aplicadas para otimizar o desempenho de redes maiores, reduzindo o consumo de memória e tempo de inferência. Essas abordagens são particularmente relevantes para aplicações em tempo real ou em dispositivos com capacidade computacional limitada, onde a eficiência é uma prioridade.

Ao comparar os resultados das três abordagens, observou-se que o uso de imagens completas na Abordagem 1 produziu os melhores resultados em termos de F1-Score (0,6723), *Precision* (0,6720) e *Recall* (0,6760). Esses resultados indicam que a utilização de imagens completas no *dataset* FER-2013 oferece uma representação mais rica e detalhada das características faciais, facilitando um reconhecimento mais preciso e robusto das expressões. A precisão relativamente alta sugere que o modelo é eficaz em evitar falsos positivos, enquanto o alto *Recall* indica uma capacidade sólida de identificar corretamente as expressões faciais presentes. Além disso, a *AUC-ROC* de 0,9057 reflete a

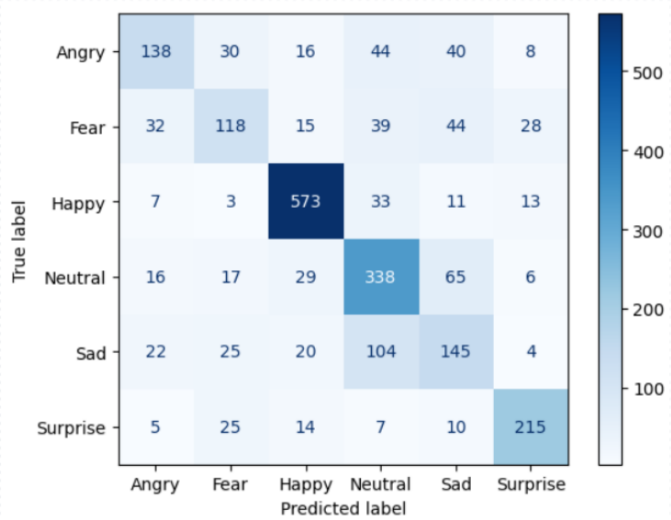

Fig. 6. Matriz de confusão resultado dos da abordagem com as imagens completas.

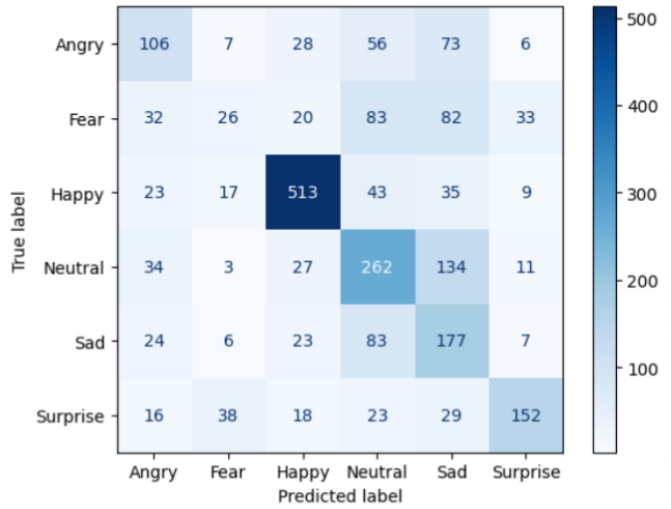

Fig. 7. Matriz de confusão com o melhor resultado da abordagem Landmarks Renderizados.

capacidade do modelo de distinguir entre diferentes classes de emoções de forma consistente, confirmando a robustez dessa abordagem, especialmente em cenários em que a precisão e a sensibilidade são cruciais.

Nas abordagens *Landmarks* Renderizados e *Landmarks* Vetoriais, que utilizaram *landmarks* faciais, notou-se que o nível de detalhamento dos *landmarks* teve um impacto significativo no desempenho do modelo. Em particular, o Mapa de Pontos

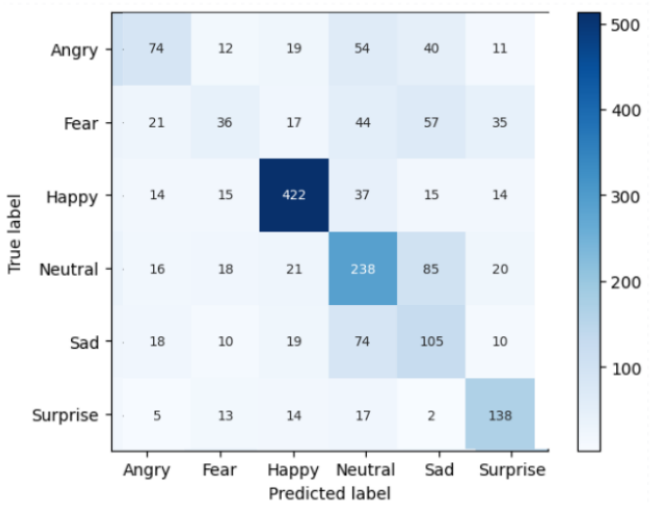

Fig. 8. Matriz de confusão com o melhor resultado da abordagem Landmarks Vetoriais.

Conectados (MPC) na Abordagem 2 apresentou um desempenho superior, com um F1-Score de 0,5777, *Precision* de 0,5812, *Recall* de 0,5781 e *AUC-ROC* de 0,8612. A análise crítica desses resultados sugere que a adição de conexões entre os pontos faciais não só melhora a representação espacial das expressões, mas também preserva relações estruturais essenciais que são fundamentais para a interpretação correta das emoções. O fato de o MPC ter superado as outras configurações de *landmarks* indica que a estrutura conectada oferece uma representação mais rica das características faciais, o que é especialmente vantajoso em cenários de emoções complexas, onde nuances sutis são importantes.

A maior precisão observada no MPC indica que essa configuração é mais eficaz em evitar falsos positivos, o que é importante para aplicações que exigem alta confiabilidade na detecção de emoções, como em contextos clínicos ou de segurança. No entanto, o aumento da complexidade computacional associado ao MPC pode limitar sua aplicabilidade em ambientes com restrições de recursos, sugerindo um *trade-off* entre precisão e eficiência.

Na abordagem de *Landmarks* Vetoriais, foi utilizada uma rede MLP, observando-se um desempenho inferior em todas as configurações de *landmarks* quando comparado à Abordagem 2. O F1-Score mais alto alcançado na Abordagem 3 foi de 0,5657 com a configuração MPS, com *Precision* de 0,5704, *Recall* de 0,5756 e *AUC-ROC* de 0,8603. Esses resultados sugerem que, embora a MLP tenha uma capacidade decente de processar *landmarks*, ela pode não ser a escolha ideal para classificar expressões faciais sem ajustes adicionais ou

modificações na arquitetura. A ligeira queda no desempenho em relação à Abordagem 2 pode ser atribuída à menor capacidade da MLP de capturar as complexidades das conexões faciais, que são melhor manejadas pela ResNet50. Isso evidencia a importância de considerar a arquitetura da rede em relação ao tipo de entrada utilizada: enquanto a ResNet50, uma rede convolucional profunda, é capaz de explorar as nuances dos *landmarks* conectados, a MLP, sendo uma rede totalmente conectada, pode não conseguir extrair a mesma profundidade de informações a partir de dados que requerem um entendimento mais sofisticado da estrutura facial.

Além disso, a análise dos resultados sugere que a escolha dos *landmarks* e da arquitetura da rede deve ser feita com base no objetivo específico da aplicação. Para tarefas em que a precisão e a sensibilidade são essenciais, a utilização de *landmarks* conectados em uma rede como a ResNet50 pode oferecer vantagens significativas. Por outro lado, para aplicações que exigem maior eficiência computacional com um compromisso menor em termos de precisão, abordagens mais simples, como MPS em uma MLP, podem ser suficientes.

IV. Conclusão

Os resultados obtidos reforçam a superioridade da abordagem com imagens completas, especialmente em contextos em que a precisão e a sensibilidade são fundamentais, como no reconhecimento detalhado de expressões faciais. No entanto, em situações em que o uso de imagens faciais completas não é viável — seja por restrições legais, preocupações com a privacidade ou limitações de recursos computacionais — a utilização de *landmarks* faciais surge como uma alternativa promissora.

A configuração conectada dos *landmarks*, como o MPC, mostrou-se eficaz em fornecer uma representação detalhada das características faciais, permitindo um equilíbrio interessante entre desempenho e eficiência computacional. Embora as imagens completas ofereçam um reconhecimento mais preciso e robusto, os *landmarks* proporcionam uma solução viável em cenários em que a proteção da privacidade e a redução da complexidade computacional são prioridades.

Portanto, apesar de as imagens faciais completas serem a escolha ideal para aplicações que exigem a máxima precisão, os *landmarks* faciais podem ser adequadamente utilizados em contextos em que essas exigências são mitigadas por outras considerações, como a privacidade ou a necessidade de menor carga computacional. Estudos futuros podem explorar ajustes adicionais na arquitetura MLP ou investigar a integração de técnicas híbridas que combinam a profundidade das imagens completas com a leveza dos *landmarks*, ampliando o escopo de aplicações práticas e seguras do reconhecimento de expressões faciais.

Referências

[1] A. V. Savchenko, "Facial expression and attributes recognition based on multi-task learning of lightweight neural networks," *arXiv preprint*, 2023. [Online]. Available: http://arxiv.org/abs/2103.17107v3

[2] B. Fang, Y. Zhao, G. Han, and J. He, "Expression-guided deep joint learning for facial expression recognition," *Sensors*, vol. 23, no. 16, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/16/7148

[3] T. Kopalidis, V. Solachidis, N. Vretos, and P. Daras, "Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets," *Information*, vol. 15, no. 3, 2024. [Online]. Available: https://www.mdpi.com/2078-2489/15/3/135

[4] D. Ciraolo, M. Fazio, R. S. Calabrò, M. Villari, and A. Celesti, "Facial expression recognition based on emotional artificial intelligence for tele-rehabilitation," *Biomedical Signal Processing and Control*, vol. 92, p. 106096, 2024.

[5] C. C. Chibelushi and F. Bourel, "Facial expression recognition: A brief tutorial overview," Staffordshire University, ORSYP, Tech. Rep., 2002. [Online]. Available: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=860287296e960dcc54508813b9bd55c89f5c23ea

[6] Brasil, "Lei geral de proteção de dados pessoais, lei nº 13.709, de 14 de agosto de 2018," http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm, 2018.

[7] P. V. K. Sandeep and N. S. Kumar, "Pain detection through facial expressions in children with autism using deep learning," *Soft Computing*, vol. 28, pp. 4621–4630, 2024. [Online]. Available: https://doi.org/10.1007/s00500-024-09696-x

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778. [Online]. Available: https://arxiv.org/abs/1512.03385

[9] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5288526

[10] Z. Zhao, S. Xu, B. H. Kang, M. M. J. Kabir, Y. Liu, and R. Wasinger, "Investigation and improvement of multi-layer perceptron neural networks for credit scoring," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3508–3516, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414007726

[11] T. D. Science, "The annotated resnet-50," *Towards Data Science*, 2021. [Online]. Available: https://towardsdatascience.com/the-annotated-resnet-50-a6c536034758

[12] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306457309000259

[13] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240. [Online]. Available: https://minds.wisconsin.edu/handle/1793/60482