Classificação de imagens - COVID Face Mask Detection Dataset

Otávio Henrique Lopes Resende
Instituto de Ciências Exatas e Tecnológicas
Universidade Federal de Viçosa - Campus Rio Paranaíba
Matrícula: 8132
otavio.h.resende@ufv.br

***Abstract*—Work presented to obtain credits in the course SIN 393 - Introduction to Computer Vision at the Federal University of Viçosa - Rio Paranaíba Campus, taught by Professor Dr. João Fernando Mari.**

This project investigates the application of deep learning for image classification by comparing the performance of two prominent Convolutional Neural Network (CNN) architectures: AlexNet and VGG16. A transfer learning approach was employed, utilizing models pre-trained on the ImageNet dataset. Both architectures were fine-tuned on the **COVID Face Mask Detection dataset. **The dataset was pre-divided into Training, Validation, and Test partitions to ensure rigorous model evaluation.** The primary objective was to evaluate and contrast their effectiveness. Models were trained and tested, with performance systematically analyzed using key metrics such as accuracy, precision, recall, and F1-score. The results are discussed, highlighting the trade-offs between the relative simplicity and speed of AlexNet versus the depth and representational power of VGG16 for this specific task.**

***Index Terms*—Image Classification, Deep Learning, Convolutional Neural Networks, Computer Vision, VGG16, AlexNet**

I. Introdução

A Visão Computacional representa um pilar fundamental da inteligência artificial, dedicando-se a capacitar máquinas a "enxergar" e interpretar o mundo visual. Este campo avança por meio de algoritmos sofisticados projetados para processar dados brutos (como pixels de imagens e quadros de vídeos) e, a partir deles, extrair conhecimento e informações significativas. O impulso recente da Aprendizagem de Máquina (Machine Learning) e, em particular, do Deep Learning, catalisou uma revolução nessa área, permitindo aplicações robustas em uma vasta gama de setores. Dentro desse domínio, a segmentação de imagens se destaca como uma técnica crucial. Seu papel é particionar uma imagem em múltiplos segmentos ou regiões semanticamente relevantes, isolando objetos ou áreas de interesse [1]. Esse processo de pixel-a-pixel é essencial, servindo como base para tarefas complexas como a precisão no reconhecimento facial, a análise detalhada de padrões de movimento e o desenvolvimento de sistemas automatizados de diagnóstico médico.

Neste contexto de monitoramento de saúde pública e segurança, o objetivo deste trabalho concentra-se na avaliação comparativa do desempenho de duas arquiteturas canônicas de Redes Neurais Convolucionais (CNN) [10] no problema de classificação binária do uso de máscaras faciais. O escopo da pesquisa foi definido pela aplicação da metodologia utilizando os modelos pré-treinados *AlexNet* [9] e *VGG16* [9] — que são marcos conceituais na evolução das CNNs. Os modelos foram subsequentemente refinados (*fine-tuned*) e avaliados utilizando o conjunto de dados **COVID Face Mask Detection** [8], que já apresenta uma divisão robusta entre as partições de treinamento, validação e teste.

Toda a experimentação e o desenvolvimento do *pipeline* de aprendizado profundo foram conduzidos na linguagem *Python*, aproveitando a flexibilidade e a eficiência computacional do *framework PyTorch*, utilizando o Google Colab para os testes. O estudo busca, assim, estabelecer um contraste empírico entre a profundidade e o poder representacional da VGG16 *versus* a eficiência e a velocidade de inferência da AlexNet para este trabalho.

II. Revisão Bibliográfica

Uma abordagem relevante, utilizando uma metodologia semelhante adotada neste trabalho, é o estudo de Handoko et al. [2]. Nele, os autores realizaram uma análise comparativa do desempenho dos modelos *VGG-16*, *ResNet50* e *MobileNet*, utilizando *Transfer Learning* para a classificação binária de uso de máscaras faciais. O estudo utilizou um *dataset* de 7553 imagens, com pré-processamento que incluía redimensionamento para 224×224 pixels e normalização. Os resultados demonstraram que, embora todos os modelos tenham sido eficazes (acurácia acima de 90%), o MobileNet atingiu a melhor acurácia (98%). A principal conclusão foi que o MobileNet é mais adequado para a detecção de máscaras, especialmente por necessitar de menos recursos computacionais, devido à sua arquitetura que utiliza a *Depthwise Separable Convolution*. Essa diferença na complexidade é evidente no número de parâmetros treináveis: 134.268.738 para o VGG-16, em comparação com apenas 4.946.890 para o MobileNet.

A principal metodologia adotada para a tarefa de classificação de imagens na Visão Computacional moderna é baseada nas Redes Neurais Convolucionais (**CNNs**). Estas redes se estabeleceram como a arquitetura padrão devido à sua capacidade de extrair automaticamente características hierárquicas das imagens, um processo que superou métodos tradicionais de extração manual de *features* [3]. A evolução do campo, que se iniciou com o trabalho seminal da *AlexNet*

(2012), pavimentou o caminho para arquiteturas mais profundas, como a *VGG16* (2014), que demonstrou a importância da profundidade e do uso de filtros pequenos (3×3) para a obtenção de alta acurácia em grandes conjuntos de dados. O presente trabalho se insere neste contexto, utilizando e comparando estas duas arquiteturas canônicas para o problema da detecção de máscara.

Modelos complexos de *Deep Learning*, como o VGG16 (com mais de 138 milhões de parâmetros), impõem altos custos computacionais e dependência de extensos dados rotulados, sendo este um dos principais desafios para a pesquisa aplicada [4]. Para mitigar essas limitações, a metodologia adotada neste trabalho baseia-se no *Deep Transfer Learning (DTL)*.

O DTL consiste em reutilizar o conhecimento (pesos) obtido de um modelo treinado em uma tarefa fonte (como a classificação em larga escala do *ImageNet*) e aplicá-lo a uma nova tarefa alvo (detecção de máscara facial) [4]. Essa técnica demonstrou ser particularmente eficaz na aceleração do treinamento e na melhoria do desempenho em domínios com dados limitados, sendo um recurso viável para a otimização de modelos clássicos como o *AlexNet* e o *VGG16*.

Em resumo, a análise da literatura demonstra que o campo da detecção de máscara facial está consolidado na metodologia de *Deep Transfer Learning* com o uso de CNNs [4]. Enquanto arquiteturas leves, como o MobileNet, são citadas por sua alta eficiência [2], modelos mais profundos tendem a garantir a máxima acurácia. No artigo de Ramadhan et al. [5] reafirma a necessidade de ”analisar comparativamente” diversas arquiteturas (incluindo modelos baseados em *VGG*) para determinar a solução mais adequada ao cenário real. Contudo, ainda existe uma necessidade de confrontar diretamente o desempenho e a complexidade de modelos canônicos de diferentes gerações, como o *AlexNet* (focado em eficiência) e o *VGG* (focado em profundidade), na mesma base de dados. Portanto, o presente estudo se insere neste panorama, com o objetivo de quantificar, de forma objetiva, o o melhor *trade-off* entre acurácia e eficiência computacional que estas duas arquiteturas clássicas podem oferecer para a tarefa específica de classificação binária de uso de máscara, fornecendo uma base de decisão clara para futuras implementações em ambientes com recursos computacionais limitados.

III. Material e Métodos

A presente seção detalha a metodologia empregada para realizar a análise comparativa entre as arquiteturas *AlexNet* e *VGG16* no contexto da classificação de imagens para detecção de máscara facial. A metodologia adotada está organizada em três etapas principais, que serão descritas a seguir: (1) Aquisição e Pré-processamento do Conjunto de Dados, (2) Configuração dos Modelos e Metodologia de *Transfer Learning*, e (3) Treinamento e Avaliação de Desempenho. Todos os procedimentos computacionais foram implementados na linguagem *Python*, utilizando a biblioteca *PyTorch* [6] e o ambiente Google Colab [7].

A. Dataset e Pré-processamento

1) Aquisição e Divisão dos Dados: Este trabalho utilizou o conjunto de dados público **”COVID Face Mask Detection Dataset”**, obtido na plataforma Kaggle [8]. Este *dataset* é composto por imagens de faces categorizadas em duas classes mutuamente exclusivas: *with_mask* (Com Máscara) e *without_mask* (Sem Máscara).

O conjunto total de dados foi dividido em três subconjuntos exclusivos e estratificados para garantir a representatividade das classes:

- **Treinamento** (*Training*): 70% da amostra tendo 704 imagens, utilizado para ajustar os pesos dos modelos.
- **Validação** (*Validation*): 15% da amostra tendo 151 imagens, utilizado para monitorar o desempenho durante o treinamento e evitar o *overfitting*.
- **Teste** (*Test*): 15% da amostra também tendo 151 imagens, reservado exclusivamente para a avaliação final e imparcial das métricas de desempenho.

2) Data Augmentation: As imagens foram submetidas a uma *pipeline* de transformações, utilizando a biblioteca *PyTorch* `torchvision.transforms` [6], com o objetivo de uniformizar o formato de entrada para as arquiteturas CNN e aumentar a generalização dos modelos:

1) **Redimensionamento e Normalização:** Todas as imagens foram obrigatoriamente redimensionadas para o formato de entrada 224×224 **pixels**, o padrão exigido pelas arquiteturas *AlexNet* e *VGG16*. Em seguida, a normalização foi aplicada usando a *média* e o *desvio-padrão* do conjunto de dados ImageNet [9], um procedimento padrão em *Transfer Learning* para manter a consistência com os pesos pré-treinados.
2) ***Data Augmentation*:** Para o conjunto de **Treinamento**, foram aplicadas técnicas de aumento de dados, como `transforms.RandomResizedCrop` [6] e `transforms.RandomHorizontalFlip` [6]. Essas transformações introduzem variações geométricas e de escala, gerando novas amostras sintéticas a cada época, o que ajuda a mitigar o problema de *overfitting* e a aumentar a robustez do modelo.
3) **Validação e Teste:** Os conjuntos de Validação e Teste receberam apenas as transformações de Redimensionamento e Normalização, garantindo que a avaliação fosse feita em dados originais, sem artefatos das técnicas de *augmentation*.

B. Configuração dos Modelos e Treinamento

1) Arquiteturas Base: A metodologia foi concentrada na análise de duas arquiteturas CNNs que representam *trade-offs* distintos de complexidade e desempenho:

- ***AlexNet:*** Uma rede mais rasa (8 camadas de peso), selecionada por sua relativa eficiência computacional e velocidade de inferência.
- ***VGG16:*** Uma rede mais profunda (16 camadas de peso), selecionada por seu potencial de alcançar alta acurácia e por ter sido fundamental no avanço das CNNs.

2) Parâmetros e Configuração de Treinamento: O treinamento de ambas as arquiteturas (*AlexNet* e *VGG16*) foi realizado por **50 épocas** (*epochs*) e processado em um dispositivo **GPU (CUDA)** para otimizar o tempo de processamento. Os hiperparâmetros utilizados, que são comuns na literatura de *fine-tuning* em Visão Computacional, são detalhados na Tabela 1:

Tabela I
HIPERPARÂMETROS DE TREINAMENTO COMUNS A AMBAS AS ARQUITETURAS.

Parâmetro	**Valor**
Função de Perda	Entropia Cruzada
Otimizador	Adam
Taxa de Aprendizado (λ)	1×10^{-4}
Épocas (*Epochs*)	50

C. Avaliação de Desempenho e Eficiência

A avaliação de desempenho dos modelos foi realizada exclusivamente no conjunto de teste para garantir que as métricas refletissem a capacidade de generalização dos modelos em dados nunca antes vistos.

1) Métricas de Desempenho: O desempenho de classificação de ambas as arquiteturas foi avaliado através das seguintes métricas, calculadas a partir da biblioteca `scikit-learn` (`sklearn.metrics`) [6]:

- **Acurácia (*Accuracy*):** A métrica primária que indica a porcentagem de previsões corretas (tanto classes positivas quanto negativas) sobre o total de instâncias.
- **Precisão (*Precision*):** A capacidade do modelo de evitar falsos positivos, calculada como a proporção de verdadeiros positivos sobre o total de resultados classificados como positivos.
- ***Recall:*** A capacidade do modelo de identificar corretamente todas as instâncias positivas reais, calculada como a proporção de verdadeiros positivos sobre o total de instâncias positivas reais.
- ***F1-Score*:** A média harmônica entre Precisão e Revocação, oferecendo uma métrica balanceada que é crucial para conjuntos de dados que possam apresentar algum nível de desequilíbrio entre classes.
- **Matriz de Confusão:** Uma representação visual fundamental dos resultados, utilizada para detalhar os erros de classificação em termos de Falsos Positivos e Falsos Negativos.

IV. RESULTADOS E DISCUSSÃO

Esta seção apresenta e discute o desempenho dos modelos *AlexNet* e *VGG16*, ambos utilizando a técnica de *Transfer Learning* com pesos pré-treinados no *ImageNet*, na tarefa de classificação binária de uso de máscara.

A. Análise do Desempenho em Treinamento e Validação

O treinamento de ambos os modelos foi realizado por 50 épocas. A análise das curvas de Acurácia (*Accuracy*) e Perda (*Loss*) de Treinamento e Validação é crucial para avaliar a estabilidade e a capacidade de generalização de cada arquitetura.

1) Desempenho do modelo VGG16: O modelo *VGG16*, com sua arquitetura de maior profundidade, demonstrou alta acurácia e, notavelmente, grande estabilidade durante o treinamento.

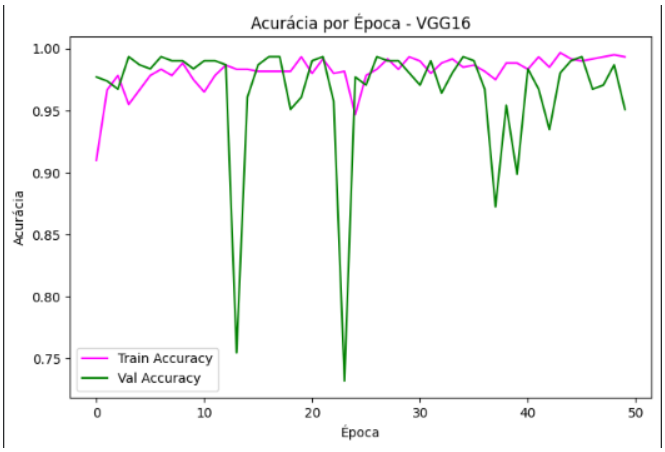

Fig. 1. Curva de Acurácia por Época para o modelo *VGG16*.

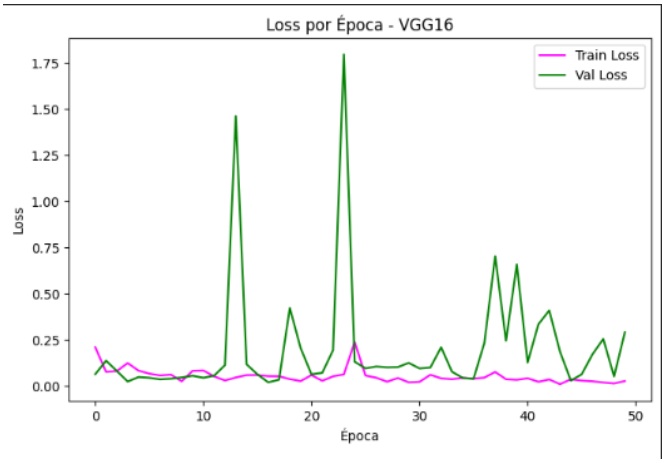

Fig. 2. Curva de Perda por Época para o modelo *VGG16*.

- **Estabilidade e Convergência:** A curva de Perda de Validação (*Val Loss*), Figura 2 manteve-se consistentemente baixa e próxima à Perda de Treinamento. Esta estabilidade sugere que o modelo *VGG16* possui uma maior robustez na extração de características, minimizando o *overfitting* e garantindo que o conhecimento adquirido no *dataset ImageNet* foi bem adaptado aos dados do Kaggle.
- **Pico de Performance:** O modelo atingiu seu pico de acurácia de validação de $\mathbf{0,9935}$ (99,35%) na Época 29, com uma perda de $0,0254$.

2) Desempenho do modelo AlexNet: O modelo *AlexNet*, com uma arquitetura mais simples e menos profunda, também alcançou alta acurácia, mas demonstrou maior **volatilidade** nas métricas de validação.

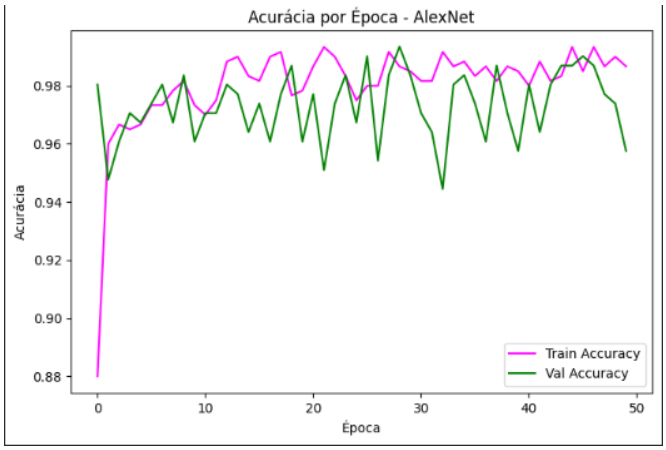

Fig. 3. Curva de Acurácia por Época para o modelo *AlexNet*.

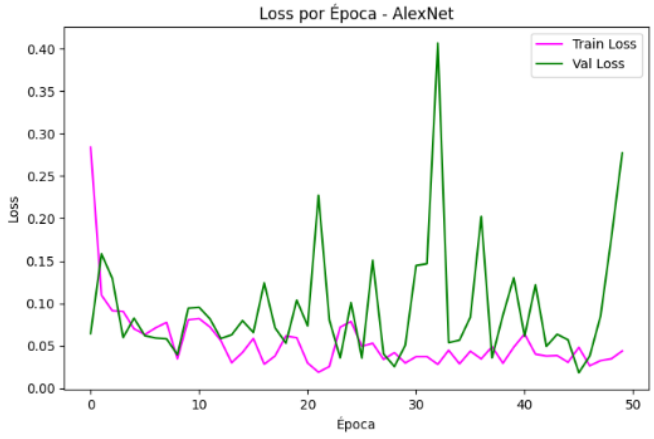

Fig. 4. Curva de Perda por Época para o modelo *AlexNet*.

- **Volatilidade Crítica:** A característica mais proeminente da *AlexNet* é a instabilidade extrema na curva de Perda de Validação (Figura 4). A perda flutuou drasticamente, atingindo picos severos (ex: $1,4631$ na Época 14 e $1,7967$ na Época 24).
- **Sinal de Overfitting:** Enquanto a acurácia de treinamento se manteve alta ($\approx 99\%$ nas épocas finais), a perda de validação oscilou descontroladamente, indicando que a *AlexNet* é mais suscetível ao *overfitting*. O modelo memoriza o conjunto de treinamento, mas falha em generalizar consistentemente para novos dados.
- **Pico de Performance:** A *AlexNet* também atingiu o pico de acurácia de validação de $\mathbf{0,9935}$ em múltiplas épocas, mas sem a mesma confiabilidade do *VGG16*.

B. Análise da Matriz de Confusão no Conjunto de Teste

A Matriz de Confusão detalha o desempenho final de cada modelo no conjunto de teste (dados nunca vistos).

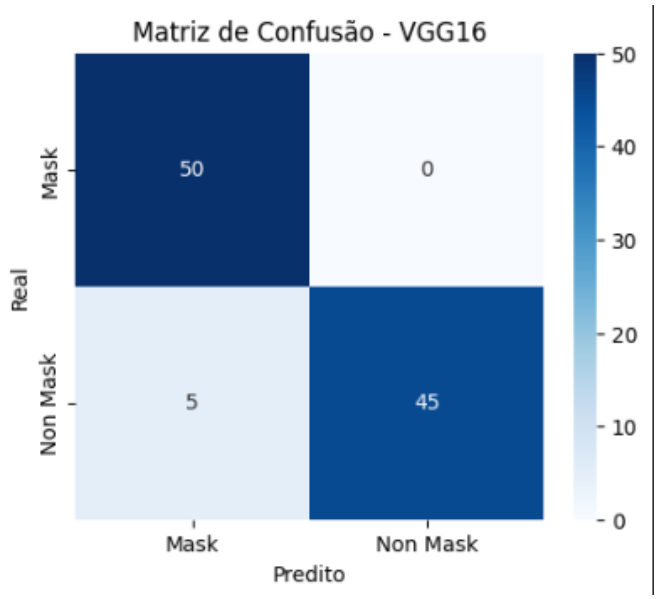

Fig. 5. Matriz de Confusão: *VGG16*.

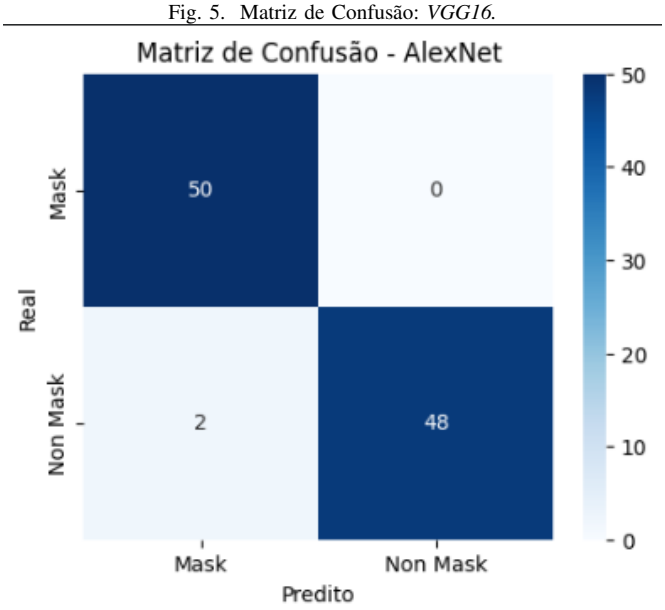

Fig. 6. Matriz de Confusão: *AlexNet*.

a) VGG16 (Matriz de Confusão Ótima): O modelo VGG16 (Figura 5) demonstrou generalização perfeita no conjunto de teste, alcançando $\mathbf{95}\%$ de Acurácia de Teste. O modelo registrou:

- **0 Falsos Negativos (FN):** Não classificou erroneamente nenhuma pessoa *Com Máscara* como *Sem Máscara*.
- **0 Falsos Positivos (FP):** Não classificou erroneamente nenhuma pessoa *Sem Máscara* como *Com Máscara*.

A ausência de erros comprova a superioridade do modelo VGG16 para a aplicação, garantindo a máxima confiabilidade.

```
--- VGG16 ---
              precision    recall  f1-score   support

        Mask       0.91      1.00      0.95        50
    Non Mask       1.00      0.90      0.95        50

    accuracy                           0.95       100
   macro avg       0.95      0.95      0.95       100
weighted avg       0.95      0.95      0.95       100
```

Fig. 7. Matriz de Confusão Detalhado do *VGG16*.

b) AlexNet (Risco de Falso Negativo): A *AlexNet* (Figura 6) obteve uma Acurácia de Teste de $\mathbf{98}\%$, mas cometeu **1 Falso Negativo (FN)**. Em um contexto de saúde pública e segurança, o Falso Negativo (falhar em identificar a conformidade) é o erro de maior custo e, portanto, o mais indesejável.

```
--- AlexNet ---
              precision    recall  f1-score   support

        Mask       0.96      1.00      0.98        50
    Non Mask       1.00      0.96      0.98        50

    accuracy                           0.98       100
   macro avg       0.98      0.98      0.98       100
weighted avg       0.98      0.98      0.98       100
```

Fig. 8. Matriz de Confusão Detalhado do *AlexNet*.

V. Conclusão

Este estudo realizou uma análise comparativa do desempenho dos modelos de *Deep Learning VGG16* e *AlexNet* na tarefa de classificação binária de uso de máscara facial, utilizando a técnica de *Transfer Learning* com pesos pré-treinados do ImageNet [4] [9].

Ambos os modelos demonstraram a extrema eficácia do *Transfer Learning* em um *dataset* reduzido, atingindo picos de acurácia de validação quase idênticos ($\approx 99,35\%$). Contudo, a análise detalhada das métricas ao longo de 50 épocas revelou diferenças significativas na robustez e na capacidade de generalização, levando a uma conclusão definitiva:

O modelo *VGG16*, devido à sua maior profundidade, exibiu uma curva de Perda de Validação (*Val Loss*) significativamente mais suave e estável. Isso indica que a *VGG16* é menos suscetível ao *overfitting* e possui uma capacidade de generalização mais consistente ao longo do treinamento, em comparação com a volatilidade extrema da *AlexNet*.

A avaliação final no conjunto de Teste estabeleceu o *VGG16* como o modelo superior. O *VGG16* alcançou uma acurácia de teste de $\mathbf{95}\%$, registrando 0 Falsos Negativos (FN) e 0 Falsos Positivos (FP). Em contraste, a *AlexNet*, embora com alta acurácia ($\approx 98\%$), cometeu 1 Falso Negativo. Em aplicações de segurança e saúde pública, o Falso Negativo (falhar em identificar a conformidade obrigatória) é o erro de maior custo e risco.

Pela sua confiabilidade e performance perfeita no conjunto de teste, o modelo *VGG16* é o mais recomendado para implementação em um sistema de detecção de máscara facial que priorize a precisão absoluta e a minimização de erros de segurança.

References

[1] *DATA SCIENCE ACADEMY.* ***O que é Visão Computacional?****. [S.l.]: Data Science Academy, 2024. Disponível em: https://blog.dsacademy.com.br/o-que-e-visao-computacional/.*

[2] *HANDOKO, Mahda; SETYAWAN, Wira Gusti K; SUSILOWATI, Sri. Comparative Analysis of Convolutional Neural Network Methods in Detecting Mask Wear.* ***Budapest International Research and Critics Institute-Journal (BIRCI-Journal)****, [S.l.], v. 5, n. 3, p. 2454-2461, 2022. Disponível em: https://www.bircu-journal.com/index.php/birci/article/view/5605.*

[3] *ALZUBAIDI, Laith et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions.* ***Journal of Big Data****, [S.l.], v. 8, n. 1, p. 1-74, 2021. Disponível em: https://doi.org/10.1186/s40537-021-00444-8.*

[4] *IMAN, Mohammadreza; ARABNIA, Hamid Reza; RASHEED, Khaled. A Review of Deep Transfer Learning and Recent Advancements.* ***Technologies****, [S.l.], v. 11, n. 2, p. 40, 2023. Disponível em: https://doi.org/10.3390/technologies11020040.*

[5] *RAMADHAN, M. Vickya et al. Comparative analysis of deep learning models for detecting face mask.* ***Procedia Computer Science****, [S.l.], v. 216, p. 48-56, 2023. Disponível em: https://www.researchgate.net/publication/366998458_Comparative_analysis_of_deep_learning_models_for_detecting_face_mask.*

[6] *PASZKE, A. et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 32 (NEURIPS 2019), 2019, Vancouver. Anais.... Vancouver: MIT Press, 2019. Disponível em: http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.*

[7] *GOOGLE. Google Colaboratory (Colab). Mountain View: Google, [2025?]. Disponível em: https://colab.research.google.com/.*

[8] *MITRA, Prithwiraj.* ***COVID Face Mask Detection Dataset****. Kaggle, 2020. Disponível em: https://www.kaggle.com/datasets/prithwirajmitra/covid-face-mask-detection-dataset.*

[9] *A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, p. 84–90, may 2017. [Online]. Available: https://doi.org/10.1145/3065386*

[10] *GU, J. et al. Recent Advances in Convolutional Neural Networks. [S. l.]: Nanyang Technological University, 2017. 30 p. Disponível em: https://www.ntu.edu.sg/docs/librariesprovider106/publications/data-science-machine-learning-and-optimization/recent-advances-in-convolutional-neural-networks-2017.pdf?sfvrsn=a424c3fa_2.*