VIABILIDADE DE CNNs COMPACTAS NO DIAGNÓSTICO DA MPOX

Maicon Almeida Mian[1], Reginaldo José da Silva[2], Angela Leite Moreno[3]

[1]Universidade Federal de Alfenas, Alfenas, Brasil (maicon.mian@sou.unifal-mg.edu.br)
[2]Universidade Estadual Paulista "Júlio de Mesquita Filho", Ilha Solteira, Brasil
[3]Universidade Federal de Alfenas, Alfenas, Brasil

Resumo: A MPOX é uma zoonose viral cujo diagnóstico por PCR é oneroso e depende de estrutura laboratorial. CNNs compactas surgem como alternativa para triagem visual de lesões. Este estudo compara quatro arquiteturas (ResNet50, EfficientNetB4, B0 e MobileNetV3) no conjunto Mpox-Vision, via testes estatísticos rigorosos (Wilcoxon, TOST, Friedman). A EfficientNetB0 destacou-se como melhor compromisso: desempenho equivalente a modelos maiores com segundo menor tempo de inferência.

Palavras-chave: Diagnóstico Assistido por Computador; Mpox-Vision; CNNs Compactas; Imagens dermatológicas.

INTRODUÇÃO

O ressurgimento de doenças infecciosas, como a MPOX (historicamente conhecida como varíola dos macacos), evidencia a necessidade de métodos de diagnóstico rápidos, acessíveis e escaláveis.

A MPOX é uma doença viral zoonótica, que se tornou objeto de atenção internacional, após surtos registrados fora de seu continente de origem (África), chegando a motivar alertas sanitários pela OMS em 2022 (Nuzzo, Borio e Gostin, 2022). Clinicamente, manifesta-se em duas fases: a inicial, na qual os sintomas são semelhantes ao quadro gripal, ao passo que a secundária é marcada por erupções cutâneas (Van Nispen et al., 2023).

O diagnóstico de referência é baseado principalmente em testes moleculares por PCR realizados em amostras coletadas das lesões do paciente. Embora apresente elevada precisão diagnóstica, seu desempenho pode ser influenciado por fatores pré-analíticos, como a qualidade da coleta, o transporte e o processamento das amostras, além de riscos de contaminação e limitações do ensaio, que podem gerar resultados falso-negativos (Huggett et al., 2022).

Nesse cenário, a análise visual das lesões apresenta-se como alternativa complementar de triagem. Contudo, a realização da análise de forma manual é difícil, pois cada lesão pode variar de localização, morfologia, estágio de evolução, dentre outras características (Almeida et al., 2023). Nesse contexto, ferramentas de visão computacional, como as CNNs (*Convolutional Neural Networks*), surgem como alternativa para automatizar essa análise, sendo capazes de aprender e distinguir os padrões visuais característicos de cada condição. Embora os *Vision Transformers* (ViT) também venham obtendo bons resultados em tarefas semelhantes, optou-se pelas CNNs neste trabalho por seu menor custo computacional: característica alinhada ao objetivo de execução em dispositivos restritos (Aruk, Pacal e Toprak, 2026).

Alguns autores empregaram o uso das CNNs para identificação da MPOX, obtendo resultados promissores (Zhang e Zheng, 2026; Bala et al., 2023). Em geral, esses modelos são treinados em conjuntos de dados que incluem doenças com manifestações cutâneas semelhantes, permitindo que a rede aprenda a diferenciá-las da MPOX.

Observa-se que grande parte da literatura se concentra na avaliação do desempenho preditivo de arquiteturas profundas, como ResNet, DenseNet, *Inception*, EfficientNet e *Vision Transformers*, utilizando principalmente métricas de classificação para comparação entre modelos (Akter et al., 2024). Em contrapartida, aspectos relacionados à eficiência computacional recebem menor atenção. Em particular, poucos trabalhos sobre MPOX (i) quantificam, com rigor estatístico, o impacto da redução da complexidade dos modelos sobre o desempenho preditivo e (ii) reportam medidas objetivas de eficiência computacional, como o tempo de inferência.

Diante dessa lacuna, este trabalho analisa a influência de quatro arquiteturas de CNNs, correlacionando a forma que são organizadas (blocos utilizados, profundidade) e número de parâmetros à eficácia dos resultados, comparando redes maiores com as menores. Todas elas foram treinadas sob as mesmas condições, utilizando o conjunto de dados Mpox-Vision, que conta com quatro doenças virais: Acne, Catapora, Sarampo e MPOX (Hossain, Ahmed e Rahman, 2025). Tem-se, como objetivo, analisar a

viabilidade de modelos compactos, voltados para eficiência computacional em dispositivos móveis, mensurando se a redução do número de parâmetros e tamanho da rede acarreta um comprometimento do desempenho.

MATERIAL E MÉTODOS

Esse tópico apresenta os métodos utilizados na pesquisa para treinamento e comparação dos modelos.

Banco de dados

Para o treinamento, o banco de dados utilizado foi o Mpox-Vision, com imagens provenientes de conjuntos de dados públicos e de buscas na web, posteriormente validadas quanto à sua autenticidade (Hossain, Ahmed e Rahman, 2025). Todas as imagens foram revisadas cuidadosamente por dois especialistas e rotuladas para quatro classes: Acne, Catapora, Sarampo e MPOX. Um exemplo de cada classe pode ser visto na Figura 1. O conjunto de dados é balanceado, com 200 imagens para cada uma das condições.

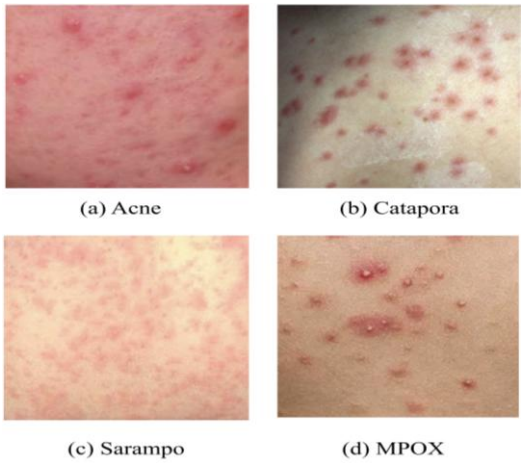


Figura 1. Exemplos de imagens para cada classe.

Arquiteturas

A fim de analisar se o número de parâmetros influência no resultado, foram selecionadas quatro arquiteturas para cobrir diferentes faixas de complexidades. A maior delas foi a ResNet50 (He et al., 2016; R50), com aproximadamente 23 milhões de parâmetros, sendo uma das mais consolidadas para classificação de imagens. A EfficientNetB4 (B4, com aproximadamente 17 milhões de parâmetros) e EfficientNetB0 (B0, com aproximadamente 4 milhões de parâmetros), ambas da mesma família EfficientNets (Tan e Le, 2019), foram selecionadas para representar as redes de médio-grande porte e as de médio-pequeno porte, respectivamente. Por fim, a menor rede selecionada foi a MobileNet V3 (MV3) em sua versão pequena (*small*), pensada para dispositivos móveis, com aproximadamente 900 mil parâmetros (Howard et al., 2019).

Condições dos Experimentos

Cada uma das redes foi submetida às mesmas condições de treinamento. As imagens foram redimensionadas para 224x224 pixels antes da entrada na rede. Durante o treinamento, utilizou-se tamanho de lote igual a 16. Os modelos foram inicializados com pesos pré-treinados no ImageNet e submetidos ao *fine-tuning*, mantendo congeladas apenas as camadas de *Batch Normalization*, estratégia adotada para preservar as estatísticas de normalização aprendidas durante o pré-treinamento e evitar instabilidades durante o *fine-tuning*.

O otimizador Adam foi utilizado com taxa de aprendizado de 0,00005 para maior estabilidade entre épocas. O treinamento foi limitado a 50 épocas devido à convergência do modelo e ao tempo computacional disponível. Para reduzir o risco de sobreajuste, foi empregada a parada antecipada pelo valor da perda da validação e paciência de seis épocas, número definido empiricamente por proporcionar tempo suficiente para a recuperação de pequenas oscilações. Ademais, a rede foi conectada a uma camada de saída com 4 neurônios para previsão, com os mapas de características sendo comprimidos pelo *Global Average Pooling*.

Cada arquitetura foi submetida a 10 sementes, que foram aplicadas exclusivamente à divisão dos dados em conjuntos de treino, validação e teste (70%, 15% e 15%, respectivamente). Todos os demais elementos aleatórios foram fixados com uma única semente, de modo que as variações nos resultados refletem apenas a sensibilidade dos modelos à partição dos dados.

Métricas e Estatísticas

Para realizar a comparação, algumas métricas foram selecionadas para avaliar cada um dos modelos: Acurácia (Acur.), que mede a taxa geral de acertos; Sensibilidade (Sens.), que identifica corretamente as amostras de cada classe em relação ao total de amostras reais da respectiva classe; F-1 Score (F1), que balanceia a precisão e a sensibilidade; Área sob a curva *Receiver Operating Characteristic* (AUC-ROC), que avalia a capacidade do modelo em distinguir as classes.

O conjunto de dados inicial foi criado para a identificação da classe MPOX. Por essa razão, as métricas foram contabilizadas tanto pela média macro entre as quatro classes, quanto pelo desempenho obtido em cada classe individualmente

A fim de avaliar a eficiência temporal das arquiteturas, analisou-se o tempo de inferência de cada modelo durante a etapa de testes. Todos os

experimentos foram executados em um ambiente composto por um processador AMD Ryzen 5 4600G, 16 GB de RAM, sem GPU dedicada, utilizando o sistema operacional Windows 10.

Visando uma análise estatística para corroborar com a significância da diferença entre os modelos, alguns métodos foram aplicados.

Primeiramente, para cada métrica, foram estimados intervalos de confiança de 95% com base na distribuição t-*Student*, utilizando a média, o desvio padrão e o número de execuções (sementes). Ademais, alguns testes de hipóteses sobre os resultados foram realizados. O primeiro deles foi o teste de Shapiro-Wilk (Shapiro e Wilk, 1965), para analisar a normalidade dos dados, em que p-valores maiores que 0,05 indicam que a distribuição da métrica em cada semente segue uma distribuição normal.

Com o resultado do teste de Shapiro-Wilk, para comparar os modelos, foi aplicado um teste de paridade com cada uma das combinações de modelos disponíveis, utilizando a correção de Holm (Holm, 1979). Adotou-se o seguinte critério: se qualquer uma das métricas apresentasse não-normalidade, aplicava-se o teste de Wilcoxon (Wilcoxon, 1945). Caso todos os dados fossem normalmente distribuídos, utilizava-se o teste t-*Student* (Student, 1908). Em ambos os casos, p-valores menores que 0,05 são considerados indicativos de diferença significativa entre os modelos.

A partir dos resultados do teste de paridade, foi realizado o teste de TOST (Schuirmann, 1987) para analisar a equivalência entre os modelos, com a aplicação de versões paramétricas ou não paramétricas. Inicialmente, foram excluídas dessa etapa as comparações que já haviam apresentado diferença estatisticamente significativa no teste de paridade (considerados diferentes). Em seguida, aplicou-se o teste TOST às comparações restantes. As margens de equivalência foram estabelecidas com base no contexto de uso da ferramenta como auxiliar de triagem, não como substituto do diagnóstico molecular.

Para métricas de classificação, adotou-se 1%, limiar conservador considerando que variações naturais de desempenho entre ambientes clínicos em sistemas de IA dermatológica superam 5% (Tjiu e Lu, 2025). Para o tempo de inferência, 500 ms corresponde ao limiar de responsividade perceptível em sistemas interativos (Nielsen, 1993), acima do qual o atraso interfere no fluxo clínico. Reconhece-se que a definição objetiva de margens de equivalência em benchmarks de aprendizado de máquina permanece um problema em aberto (Koobs e Koning, 2026), sendo a escolha aqui guiada por critérios práticos da aplicação. Considerou-se evidência de equivalência estatística quando os testes unilaterais inferior e superior apresentaram p-valores menores que 0,05. Caso uma das duas hipóteses não fosse rejeitada, a comparação era classificada como inconclusiva.

Como último passo, foi realizado o teste de Friedman (Friedman, 1937) para verificar se havia diferença significativa entre os modelos em cada uma das métricas. Diante da presença dessa diferença (p-valores menores que 0,05), realizou-se a construção de gráficos Post-Hoc (Demšar, 2006) para comparar os modelos em cada métrica.

RESULTADOS E DISCUSSÃO

O resultado dos intervalos de confiança em cada uma das métricas, para todos os modelos testados nas 10 sementes distintas, pode ser observado na Tabela 1 (desempenho global) e na Tabela 2 (desempenho por classe). Já as matrizes de confusão normalizadas, obtidas a partir da média dos resultados das sementes para cada arquitetura, são apresentadas na Figura 2. Enquanto os gráficos AUC-ROC para cada classe em cada arquitetura podem ser vistos na Figura 3.

Desempenho geral de classificação

Analisando as médias entre as classes, apresentadas na Tabela 1, as arquiteturas apresentaram, de forma geral, elevado desempenho, com destaque para B4, que obteve as maiores médias na maioria das métricas avaliadas. A B0 também se destacou, apresentando desempenho ligeiramente inferior, porém consistente e próximo ao da B4. MV3, por sua vez, apresentou desempenho inferior em relação às EfficientNets, mas demonstrou maior estabilidade entre as execuções, evidenciada pelos menores intervalos de confiança.

Tabela 1. Desempenho global (média entre as quatro classes), com intervalos de confiança de 95%.

Modelo	Acur. (%)	Sens. (%)	F1
R50	91,33 [88,90 - 93,77]	91,33 [88,90 - 93,77]	0,9130 [0,8885 - 0,9374]
B4	92,58 [90,89 - 94,28]	92,58 [90,89 - 94,28]	0,9257 [0,9086 - 0,9427]
B0	92,42 [90,59 - 94,25]	92,42 [90,59 - 94,25]	0,9238 [0,9051 - 0,9425]
MV3	89,50 [88,37 - 90,63]	89,50 [88,37 - 90,63]	0,8943 [0,8825 - 0,9060]

Análise por classe

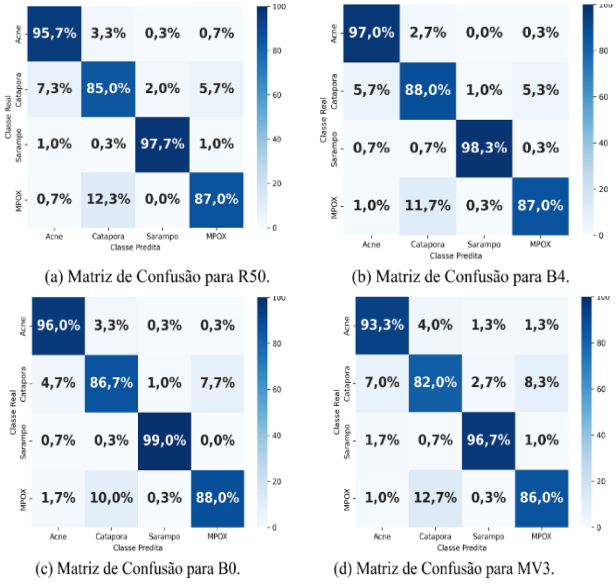
Analisando os resultados por classe apresentados na Tabela 2, observa-se que a MPOX apresentou as menores médias entre as quatro classes, com valores de sensibilidade reduzidos para a faixa de [80%, 90%].

Tabela 2. Desempenho por classe, com intervalos de confiança de 95%.

	Mod.	Sens. (%)	F1.	AUC-ROC
Acne	R50	95,67 [92,68 - 98,65]	93,57 [90,74 - 96,39]	0,9944 [0,9907 - 0,9981]
	B4	97,00 [95,24 - 98,76]	94,99 [93,13 - 96,85]	0,9974 [0,9952 - 0,9995]
	B0	96,00 [94,12 - 97,88]	94,62 [92,67 - 96,58]	0,9965 [0,9945 - 0,9986]
	MV3	93,33 [88,70 - 97,97]	91,87 [89,33 - 94,42]	0,9889 [0,9835 - 0,9944]
Catapora	R50	82,00 [75,82 - 88,18]	82,12 [79,25 - 84,98]	0,9563 [0,9488 - 0,9638]
	B4	88,00 [82,95 - 93,05]	86,69 [83,39 - 89,98]	0,9712 [0,9611 - 0,9813]
	B0	86,67 [80,51 - 92,82]	86,43 [82,58 - 90,29]	0,9753 [0,9666 - 0,9840]
	MV3	85,00 [78,52 - 91,48]	84,50 [80,26 - 88,75]	0,9686 [0,9589 - 0,9783]
Sarampo	R50	96,67 [94,72 - 98,61]	96,21 [95,01 - 97,41]	0,9981 [0,9967 - 0,9995]
	B4	98,33 [96,31 - 100]	98,49 [97,17 - 99,82]	0,9999 [0,9997 - 1,0000]
	B0	99,00 [97,85 - 100]	98,68 [97,77 - 99,60]	0,9999 [0,9998 - 1,0000]
	MV3	97,67 [95,40 - 99,93]	97,66 [96,36 - 98,95]	0,9994 [0,9989 - 1,0000]
MPOX	R50	86,00 [82,48 - 89,52]	87,52 [84,74 - 90,30]	0,9799 [0,9745 - 0,9852]
	B4	87,00 [82,30 - 91,69]	0,9010 [86,76 - 93,44]	0,9834 [0,9769 - 0,9900]
	B0	88,00 [83,48 - 92,52]	0,8979 [86,87 - 92,71]	0,9843 [0,9793 - 0,9893]
	MV3	87,00 [81,79 - 92,21]	89,46 [86,13 - 92,80]	0,9852 [0,9787 - 0,9916]

Analisando as matrizes de confusão apresentadas na Figura 2, é possível perceber que isso ocorre por uma dificuldade de discriminação entre as classes, com cerca de 10% das imagens de MPOX sendo confundidas com Catapora em todas as arquiteturas, denotando que esse é um ponto difícil para todos os modelos. Isso pode ser explicado pelo fato de que ambas as doenças compartilham características visuais semelhantes, como lesões com alto contraste em relação à pele de fundo (isto é, áreas com grande diferença de intensidade de cor e textura), distribuídas em diferentes regiões do corpo e, em alguns casos, apresentando conteúdo purulento.

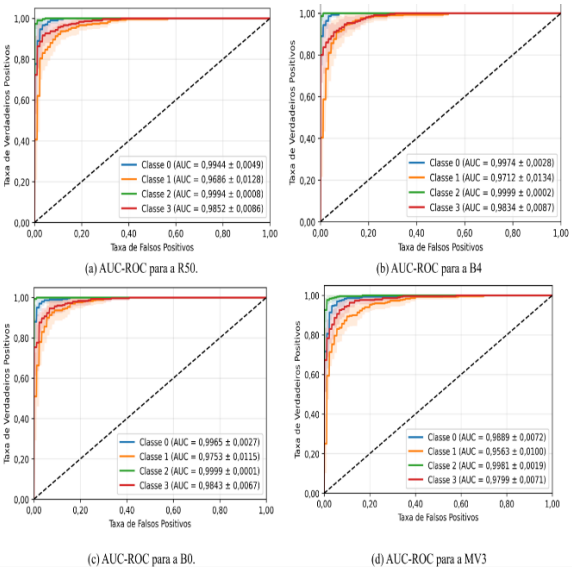
Por fim, a análise das curvas AUC-ROC apresentadas na Figura 3, confirma o elevado desempenho de todos os modelos avaliados, com valores próximos de um para todas as classes. Isso demonstra a alta capacidade de discriminação dos modelos em distinguir a MPOX de outras doenças com lesões de pele semelhantes em diferentes limiares de decisão. Além disso, os baixos desvios padrão evidenciam que esse desempenho foi consistente entre as diferentes execuções, reforçando a estabilidade e a capacidade de generalização dos modelos.



(a) Matriz de Confusão para R50. (b) Matriz de Confusão para B4.

(c) Matriz de Confusão para B0. (d) Matriz de Confusão para MV3.

Figura 2: Matrizes de Confusão normalizadas para cada modelo.



(a) AUC-ROC para a R50. (b) AUC-ROC para a B4

(c) AUC-ROC para a B0. (d) AUC-ROC para a MV3

Figura 3: AUC-ROC médias para cada modelo e seus respectivos desvios padrão.

Tempo de inferência

Como esperado, as arquiteturas com menor número de parâmetros apresentaram os menores tempos médios de inferência, com intervalos de confiança reduzidos e sem sobreposição entre os modelos avaliados. A MV3 destacou-se por apresentar o menor tempo médio, inferior a um segundo.

Observa-se que a R50, apesar de possuir o maior número de parâmetros, não foi a que demandou maior tempo médio de execução. Este resultado pode ser explicado pelo fato de a R50 ser composta majoritariamente por convoluções padrão, para as quais existem implementações altamente otimizadas em CPUs modernas (Kelefouras e Keramidas, 2023).

Enquanto as arquiteturas EfficientNets dependem de convoluções *depthwise-separable* e blocos *squeeze-excite* que, embora possuam menos parâmetros, tendem a ser limitados pelo acesso à memória nesse tipo de processador.

Tabela 3: Tempo de inferência médio (s) por modelo, com intervalos de confiança de 95%.

Modelo	Tempo Inferência (s)
R50	4,1527 [4,0697 - 4,2356]
B4	5,4477 [5,3318 - 5,5637]
B0	2,3763 [2,2701 - 2,4824]
MV3	0,8270 [0,7930 - 0,8610]

Análise Estatística

No que se refere aos testes estatísticos, a avaliação de normalidade por meio do teste de Shapiro-Wilk revelou ausência de normalidade em 27 métricas, especialmente envolvendo B0, com destaque para a classe de Sarampo em múltiplas métricas, o que justificou a adoção de testes não paramétricos subsequentes (Wilcoxon).

Tabela 4. P-valores significativos obtidos pelo teste de Wilcoxon.

	Comparação	B0×MV3	R50×MV3	B4×MV3
Geral	**Acur.**	0,0411	-	0,0411
	Sens.	0,0411	-	0,0411
	F1	0,0414	-	0,0108
	AUC-ROC	0,0093	-	0,0144
	Inferência	0,0059	0,0059	0,0059
Acne	**AUC-ROC**	0,0108	-	0,0090
Catapora	**AUC-ROC**	0,0080	-	0,0415
Sarampo	**Sens.**	0,0371	-	-
	F1.	0,0107	0,0494	0,0165
	AUC-ROC	0,0141	-	0,0141

Os resultados do teste de Wilcoxon podem ser vistos na Tabela 4, em que, por concisão, apenas as comparações com diferença estatisticamente significativa ($p < 0,05$) foram reportadas; as demais combinações de modelo, métrica e classe não apresentaram diferença significativa e foram omitidas da tabela. Os resultados evidenciam que todas as arquiteturas diferem significativamente da MV3 em alguma métrica de desempenho. Em particular, as EfficientNets apresentam diferenças em todas as métricas da média macro entre as classes. Adicionalmente, todas as arquiteturas apresentam diferenças significativas quando comparadas no tempo de inferência.

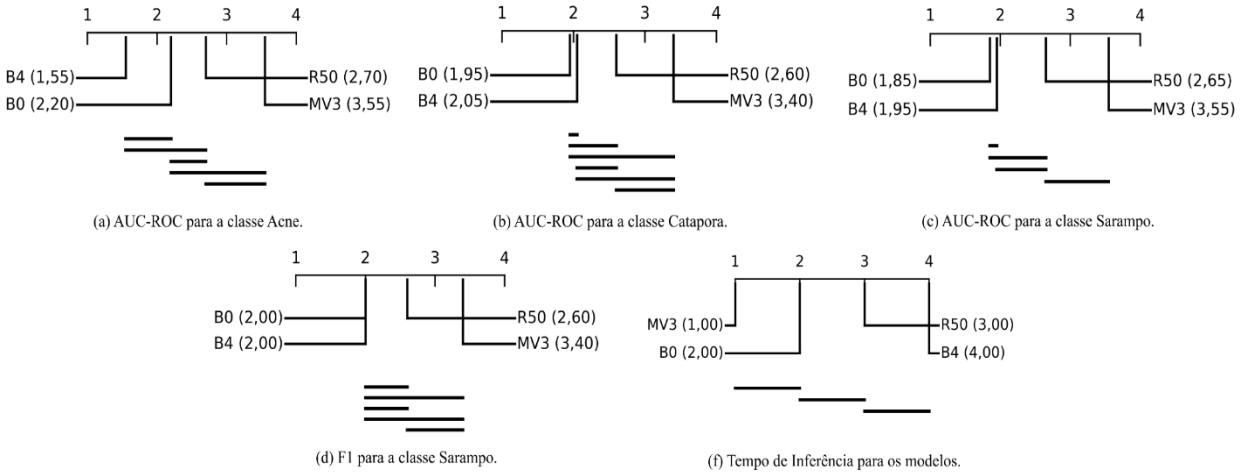
Analisando a tabela de equivalências obtida pelo teste de TOST (Tabela 5), em que I: e S: indicam, respectivamente, os p-valores dos testes unilaterais inferior e superior que compõem o TOST (ambos devem ser menores que 0,05 para se considerar equivalência estatística), observa-se que a métrica de AUC-ROC apresentou evidências de equivalência estatística.

Em particular, a B0, segunda menor arquitetura em termos de complexidade, demonstrou equivalência em AUC-ROC com os modelos de maior porte em todas as classes, exceto Catapora. Esses resultados sugerem que, apesar de sua menor complexidade, a B0 mantém desempenho estatisticamente equivalente ao de arquiteturas maiores em métricas centrais de discriminação e equilíbrio entre precisão e sensibilidade, o que reforça seu potencial como uma alternativa mais leve sem perda significativa de desempenho.

Tabela 5. P-valores significativos obtidos pelo teste TOST (bilateral).

	AUC-ROC			
	Geral	**Acne**	**Sarampo**	**Mpox**
B0×R50	I: 0,0030 S: 0,0054	I: 0,0030 S: 0,0054	I: 0,0028 S: 0,0028	I: 0,0028 S: 0,0054
B0×MV3	-	-	-	I: 0,0040 S: 0,0333
B0×B4	I: 0,0030 S: 0,0095	I: 0,0029; S: 0,0029	I: 0,0021 S: 0,0021	I: 0,0125 S: 0,1621
R50×MV3	-	-	I: 0,0029 S: 0,0029	-
R50×B4	I: 0,0125 S: 0,0054	I: 0,0095 S: 0,0029	I: 0,0028 S: 0,0028	I: 0,0125 S: 0,0333

Os gráficos de *Post-Hoc* com diferença crítica (CD = 1,483), para as métricas em que o teste de Friedman identificou diferença significativa entre os modelos podem ser vistos na Figura 4.



(a) AUC-ROC para a classe Acne. (b) AUC-ROC para a classe Catapora. (c) AUC-ROC para a classe Sarampo.

(d) F1 para a classe Sarampo. (f) Tempo de Inferência para os modelos.

Figura 4: Gráficos do teste Post Hoc realizados em cada métrica com CD = 1,483.

Para as métricas AUC-ROC e F1, observa-se que as arquiteturas B4 e B0 obtiveram os melhores rankings médios, porém sem diferença estatisticamente significativa entre si, indicando desempenho equivalente nessas métricas. Em contrapartida, ambas apresentaram diferença significativa em relação à MV3, que ocupou consistentemente a última posição, enquanto a R50 apresentou desempenho intermediário, sem se destacar de forma consistente em relação às demais arquiteturas.

Para as métricas AUC-ROC e F1, observa-se que as arquiteturas B4 e B0 obtiveram os melhores rankings médios, porém sem diferença estatisticamente significativa entre si, indicando desempenho equivalente nessas métricas. Em contrapartida, ambas apresentaram diferença significativa em relação à MV3, que ocupou consistentemente a última posição, enquanto a R50 apresentou desempenho intermediário, sem se destacar de forma consistente em relação às demais arquiteturas.

Em relação ao tempo de inferência, a MV3 obteve o menor ranking médio (1,00), seguida da B0 (2,00), evidenciando sua maior eficiência computacional. Entretanto, esse ganho de velocidade pela MV3 ocorreu à custa de uma redução significativa no desempenho preditivo.

Assim, os resultados dos testes estatísticos evidenciam um equilíbrio entre custo computacional e desempenho, indicando que a B0 apresenta o melhor equilíbrio entre elevada capacidade de classificação, com equivalência estatística em relação à B4 em algumas métricas e diferenças significativas em relação às arquiteturas de menor desempenho, além de baixo tempo de inferência.

CONCLUSÃO

Após a realização das avaliações nas quatro arquiteturas e dos testes estatísticos, algumas conclusões podem ser destacadas. Primeiramente, no contexto deste trabalho, a MV3 apresentou diferenças estatisticamente significativas em relação às demais arquiteturas em algumas métricas de desempenho, embora tenha obtido menores intervalos de confiança na média macro entre as classes. Esses resultados sugerem que, neste estudo, a MV3 pode não representar a melhor relação entre eficiência e acurácia, indicando que a redução significativa do número de parâmetros pode afetar o desempenho em determinados cenários de aplicação.

Porém, o comportamento da B0 constitui um resultado relevante: mesmo sendo a segunda menor arquitetura, apresentou equivalência estatística com os modelos de maior porte nas métricas de AUC, sem diferença significativa nas métricas de classificação. Além disso, apresentou o segundo menor tempo médio de inferência, reforçando seu equilíbrio favorável entre desempenho preditivo e eficiência computacional.

A métrica de tempo de inferência no conjunto de testes revelou distinções claras. Por isso, embora não se possa apontar um modelo superior em capacidade classificativa, torna-se possível identificar quais arquiteturas são mais eficientes em termos de latência e, portanto, mais adequadas a contextos específicos de aplicação

Sob essa ótica, a decisão entre as arquiteturas torna-se um *trade-off* entre acurácia e latência. A MV3, a menor delas, mostrou perda de desempenho, enquanto a B0 obteve boas médias de classificação e o segundo menor tempo de inferência, configurando uma opção de equilíbrio intermediário. Não há, portanto, um único modelo superior, mas escolhas adequadas a diferentes restrições de aplicação.

Como limitações deste trabalho, destaca-se o número reduzido de sementes e apenas um conjunto de dados analisado. Também se ressalta a ausência de

avaliações de desempenho computacional mais robustos de eficiência (como uso de memória e FLOPs). Como próximos trabalhos, espera-se incorporar mecanismos de atenção (como CBAM) às arquiteturas avaliadas, com vistas a reduzir a confusão entre MPOX e Catapora e o uso de conjuntos de dados externos para validação.

Em síntese, os resultados mostram que é possível alcançar elevado desempenho preditivo sem recorrer às arquiteturas de maior complexidade. Em especial, a B0 demonstrou um equilíbrio entre acurácia e eficiência computacional, tornando-se uma alternativa promissora para aplicações de apoio ao diagnóstico da MPOX em cenários com recursos computacionais limitados, favorecendo o desenvolvimento de sistemas mais acessíveis e passíveis de implantação em diferentes ambientes.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG, processo nº BIS-00722-25, Chamada 009/2025 – BIC STEM.

REFERÊNCIAS

Almeida, S. et al. A dermatologic assessment of 101 mpox (monkeypox) cases from 13 countries during the 2022 outbreak: Skin lesion morphology, clinical course, and scarring. Journal of the American Academy of Dermatology, v. 88, p. 1066-1073, 2023.

Akter, S. et al. Deep learning-enabled monkeypox diagnosis: A comprehensive survey, taxonomy, current challenges, limitations and future directions. Information Fusion, v. 111, p. 102631, 2024.

Aruk, I.; Pacal, I.; Toprak, A. N. A comprehensive comparison of convolutional neural network and visual transformer models on skin cancer classification. Computational Biology and Chemistry, v. 120, pt. 2, p. 108713, 2026.

Bala, D. et al. MonkeyNet: A robust deep convolutional neural network for monkeypox disease detection and classification. Neural Networks, v. 161, p. 757-775, 2023.

Demešar, J. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, v. 7, p. 1-30, 2006.

Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, v. 32, n. 199, p. 675-701, 1937.

He, K. et al. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition, v. 2016, p. 770-778, 2016.

Holm, S. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, v. 6, n. 2, p. 65-70, 1979.

Hossain, M. S.; Ahmed, M.; Rahman, M. S. From survey to solution: A deep learning framework for reliable monkeypox diagnosis using skin images. Array, v. 28, p. 100554, 2025.

Howard, A. et al. Searching for mobilenetv3. IEEE/CVF International Conference on Computer Vision, v. 2019, p. 1314-1324, 2019.

Huggett, J. F. et al. Monkeypox: another test for PCR. Euro Surveillance, v. 27, p. 2200497, 2022.

Kelefouras, V.; Keramidas, G. Design and Implementation of Deep Learning 2D Convolutions on modern CPUs. IEEE Transactions on Parallel and Distributed Systems, v. 34, n. 12, p. 3104-3116, 2023.

Koobs, S.; Koning, N. W. Equivalence testing with data-dependent and post-hoc equivalence margins. arXiv preprint arXiv:2603.16213, 2026.

Nielsen, J. Usability Engineering. San Francisco: Morgan Kaufmann, 1993. ISBN 0-12-518406-9.

Nuzzo, J. B.; Borio, L. L.; Gostin, L. O. The WHO declaration of monkeypox as a global public health emergency. JAMA, v. 328, n. 7, p. 615–617, 2022.

Schuirmann, D. J. A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, v. 15, n. 6, p. 657-680, 1987.

Shapiro, S. S.; Wilk, M. B. An analysis of variance test for normality (complete samples). Biometrika, v. 52, n. 3/4, p. 591-611, 1965.

Student. The probable error of a mean. Biometrika, v. 6, n. 1, p. 1-25, 1908.

Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. International Conference on Machine Learning, v. 97, p. 6105-6114, 2019.

Tjiu, J. W.; Lu, C. F. Equity and Generalizability of Artificial Intelligence for Skin-Lesion Diagnosis Using Clinical, Dermoscopic, and Smartphone Images: A Systematic Review and Meta-Analysis. Medicina, v. 61, n. 12, p. 2186, 2025.

Van Nispen, C. et al. Diagnosis and Management of Monkeypox: A Review for the Emergency Clinician. Annals of Emergency Medicine, v. 81, p. 20-30, 2023.

Wilcoxon, F. Individual comparisons by ranking methods. Biometrics Bulletin, v. 1, n. 6, p. 80-83, 1945.

Zhang, Z.; Zheng, S. Attention-enhanced hybrid deep learning framework for Monkeypox skin lesion classification. Electronic Research Archive, v. 34, p. 738-776, 2026.