VeritasAI: UMA PROPOSTA PARA DETECÇÃO DE IMAGENS ARTIFICIAIS

LEONARDO SOLOVIJOVAS SANTOS[1], JOÃO PEDRO MACHADO SILVA[1]
GABRIEL MARCELINO ALVES[2]

[1] Graduando em Bacharelado em Ciência da Computação Instituto Federal de São Paulo, São João da Boa Vista – SP.

[2] Docente do Instituto Federal de São Paulo, São João da Boa Vista – SP.

Área de conhecimento (Tabela CNPq): 1.03.00.00-7 Ciência da Computação

Resumo: Este trabalho apresenta o VeritasAI, um sistema de classificação de imagens baseado em redes neurais convolucionais, capaz de distinguir imagens reais de imagens que foram geradas por Inteligência Artificial. A arquitetura EfficientNetB0 foi empregada, combinada com técnicas de *transfer-learning*, *fine-tuning* e *data augmentation* para aumentar a robustez e generalização do modelo. O treinamento foi realizado com um conjunto balanceado de 70 mil imagens e a performance foi avaliada pelas métricas acurácia, F1-Score e matriz de confusão. Além disso, uma zona de incerteza foi implementada para identificar previsões menos confiáveis, aumentando a transparência e a confiabilidade. O sistema alcançou acurácia de 95,6% indicando a viabilidade de uma abordagem automatizada para verificação de autenticidade visual que pode apoiar a detecção de conteúdos artificiais em ambientes digitais.

Palavras-chave: Classificação de Imagens; Redes Neurais Convolucionais; Inteligência Artificial; Deepfakes.

VeritasAI: A PROPOSAL FOR ARTIFICIAL IMAGE DETECTION

Abstract: This work presents VeritasAI, an image classification system based on convolutional neural networks capable of distinguishing real images from those generated by Artificial Intelligence. The EfficientNetB0 architecture was employed, combined with transfer learning, fine-tuning, and data augmentation techniques to enhance the model's robustness and generalization. The training was conducted using a balanced dataset of 70,000 images, and performance was evaluated through accuracy, F1-score, and confusion matrix metrics. Additionally, an uncertainty zone was implemented to identify less reliable predictions, increasing transparency and trustworthiness. The system achieved an accuracy of 95.6%, indicating the feasibility of an automated approach for visual authenticity verification, which can support the detection of artificial content in digital environments.

Keywords: Image Classification; Convolutional Neural Networks; Artificial Intelligence; Deepfakes.

INTRODUÇÃO

O avanço da inteligência artificial generativa possibilitou a criação de imagens com elevado realismo, trazendo benefícios em áreas como design, publicidade e entretenimento, mas também desafios relacionados à falsificação digital, desinformação e ética. A UNESCO (2021) destaca a importância de mecanismos que permitam distinguir conteúdo real de sintético, alertando para os riscos de desconfiança em instituições e na percepção da realidade. Visando soluções para esse problema, o uso de sistemas que auxiliam a distinção entre imagens reais e artificiais é fundamental. Diante desse cenário, este trabalho apresenta o VeritasAI, um sistema cujo modelo baseado em redes neurais convolucionais (CNNs) foi treinado com o conjunto de dados *"AI vs. Human-Generated Images"* do Kaggle (Sala, 2025) e propõe verificar a autenticidade de imagens digitais, contribuindo para a mitigação da desinformação em ambientes digitais.

OBJETIVOS

O objetivo geral deste trabalho é desenvolver um sistema de classificação de imagens, baseado em redes neurais convolucionais capaz de identificar com precisão se uma imagem é real ou gerada por inteligência artificial. Como objetivos específicos, este trabalho propõe: 1) reunir e organizar um conjunto balanceado de imagens reais e artificiais; 2) implementar e validar o modelo CNN com EfficientNetB0 e *fine-tuning*, por meio da Acurácia, F1-Score e Matriz de Confusão.

REVISÃO DA LITERATURA

A família EfficientNet, proposta por Tan e Le (2019), inovou o design de redes neurais convolucionais (CNNs) ao introduzir o método de *compound scaling*. Paralelamente, o avanço da inteligência artificial generativa transformou a criação de imagens sintéticas (Goodfellow et. al., 2014). O *Latent Diffusion Model* popularizou a geração de imagens de alta resolução, tornando-as praticamente indistinguíveis das reais (Rombach et al. 2022). No entanto, Hopf e Timofte (2025) propuseram uma solução para aumentar a robustez em cenários reais. Nesse contexto, Liu et al. (2021) reforçam que as CNNs continuam eficazes na identificação de padrões artificiais mesmo em bases complexas de *deepfakes*. Alam, Kartowisastro e Wicaksono (2022) mostraram que a EfficientNet com *fine-tuning* melhora o desempenho de classificadores binários, sendo ideal para sistemas com restrições computacionais. Complementarmente, Naskar et al. (2024) destacaram que estratégias de *meta-learning*, combinadas com zonas de incerteza baseadas em probabilidades, aumentam a transparência dos modelos e possibilitam revisão humana em casos ambíguos.

METODOLOGIA

O estudo utilizou o conjunto de dados “AI vs Human-Generated Images” (Sala, 2025), composto por 70 mil imagens com dimensões superiores a 320x112 e balanceadas entre reais (Shutterstock) e artificiais (DeepMedia). As imagens foram divididas em treino (80%) e validação (20%), normalizadas para [0,1], submetidas a *data augmentation* e padronizadas em 224x224 pixels. O modelo adotado foi EfficientNetB0 com *transfer-learning* e *fine-tuning*, devido à sua eficiência em termos de acurácia e baixo custo computacional, aproveitando a transferência de aprendizado para melhorar a performance com menos parâmetros. O treinamento ocorreu em duas etapas: congelamento inicial do backbone (*warm-up*) e *progressive unfreezing*, com otimizador Adam e função de perda Entropia Cruzada Binária.

RESULTADOS

O modelo VeritasAI atingiu 95,6% de acurácia no conjunto de validação, mostrando capacidade de diferenciar imagens reais de artificiais. A Figura 1 apresenta a matriz de confusão que indicou equilíbrio entre acertos e poucos erros. O F1-Score médio foi 0,96, e uma zona de incerteza para previsões entre 0,4 e 0,6 permite análise humana em casos ambíguos.

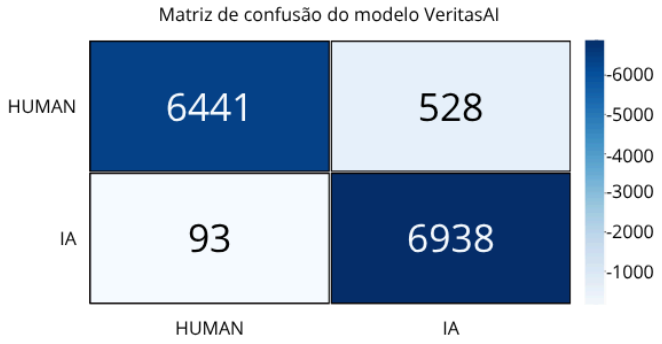

FIGURA 1. Matriz de confusão do sistema VeritasAI aplicada ao conjunto de validação.
FONTE: Autores, 2025.

DISCUSSÃO

O VeritasAI confirmou a eficácia das CNNs para detectar imagens artificiais, conforme Liu et al. (2021). A EfficientNetB0 com *fine-tuning* mostrou desempenho elevado, alinhado a Alam, Kartowisastro e Wicaksono (2022). A acurácia acima de 95% e a zona de incerteza reforçam a interpretabilidade e transparência do modelo, corroborando Naskar et al. (2024). Assim, os resultados validam e expandem as abordagens existentes, demonstrando aplicabilidade prática na mitigação da desinformação visual.

CONCLUSÕES

O VeritasAI atingiu os objetivos propostos, demonstrando capacidade de diferenciar imagens reais de artificiais. Limitações observadas em imagens de baixa qualidade ou muito estilizadas indicam a necessidade de futuras melhorias, sem comprometer o desempenho do sistema. O trabalho confirma que é possível oferecer uma ferramenta técnica eficiente para apoiar a detecção de conteúdos artificiais e a mitigação da desinformação digital.

AGRADECIMENTOS

Agradecemos ao Instituto Federal de São Paulo (IFSP) pelo apoio técnico e acadêmico durante a realização deste projeto.

REFERÊNCIAS

ALAM, I.N., Kartowisastro, I.H., Wicaksono, P. (2022). **Transfer learning technique with EfficientNet for facial expression recognition system**. Revue d'Intelligence Artificielle, Vol. 36, No. 4, pp. 543-552. https://doi.org/10.18280/ria.360405

GOODFELLOW, Ian J.; POUGET-ABADIE, Jean; MIRZA, Mehdi; XU, Bing; WARDE-FARLEY, David; OZAIR, Sherjil; COURVILLE, Aaron; BENGIO, Yoshua. **Generative Adversarial Networks**. arXiv, 2014. Disponível em: https://arxiv.org/abs/1406.2661. Acesso em: 20 out. 2025.

HOPF, Benedikt; TIMOFTE, Radu. **Practical Manipulation Model for Robust Deepfake Detection.** arXiv, 2025. Disponível em: https://arxiv.org/abs/2506.05119. Acesso em: 20 out. 2025.

LIU, Jiarui; ZHU, Kaiman; LU, Wei; LUO, Xiangyang; ZHAO, Xianfeng. **A lightweight 3D [2506.05119] Practical Manipulation Model for Robust Deepfake Detectionconvolutional neural network for deepfake detection**. International Journal of Imaging Systems and Technology, v. 31, n. 1, p. 200-210, 2021. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1002/int.22499. Acesso em: 17 ago. 2025.

NASKAR, Gourab; MOHIUDDIN, Sk; MALAKAR, Samir; CUEVAS, Erik; SARKAR, Ram. **Deepfake detection using deep feature stacking and meta-learning**. *Heliyon*, v. 10, n. 4, p. e25933, 2024. https://doi.org/10.1016/j.heliyon.2024.e25933

ROMBACH, Robin; BLATTMANN, Andreas; LORENZ, Dominik; ESSER, Patrick; OMMER, Björn. **High-Resolution Image Synthesis with Latent Diffusion Models**. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. p. 10684-10695. Disponível em: https://arxiv.org/abs/2112.10752. Acesso em: 17 ago. 2025.

SALA, Alessandra. **AI vs Human-Generated Images**. 2025. Disponível em: kaggle.com/datasets/alessandrasala79/ai-vs-human-generated-dataset. Acesso em: 17 mai. 2025.

TAN, Mingxing; LE, Quoc V. ***EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks***. arXiv, 2019. Disponível em: https://arxiv.org/abs/1905.11946. Acesso em: 22 out. 2025.

UNESCO. **Synthetic content and its implications for AI policy:** a primer. 2021. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000392181. Acesso em: 17 ago. 2025.