Deepfake Detection: A Comparative Approach

Julia Campanelli Granja [1], **Marcela Xavier Ribeiro** [1]

[1]Federal University of São Carlos (UFSCar)
Rodovia Washington Luís (SP-310), Km 235 – São Carlos – SP – Brazil

julia.campanelli@estudante.ufscar.br, marcelaxr@ufscar.br

Abstract. *The rapid advancement of synthetic image generation techniques, such as Generative Adversarial Networks (GANs), has enabled the creation of realistic deepfakes, raising concerns regarding misinformation, privacy, and digital security. This work investigates automatic methods for detecting fake facial images, focusing on convolutional neural networks (CNNs) and classical techniques such as SVM and Random Forest. Experiments were conducted using a merged dataset composed of the Real and Fake Face Detection and Human Faces datasets, totaling 10,025 images. The results indicate that the tested methods achieve satisfactory performance, demonstrating the feasibility of using machine learning approaches for automatic deepfake detection. Performance and interpretability aspects are discussed, showing the superiority of CNNs in classifying fake images, achieving an average accuracy of 94.2% in the conducted tests.*

1. Introduction

Deepfakes are synthetic media created with AI, in which the faces or voices of real people are manipulated to appear as though they are saying or doing things that never happened. Such content poses growing risks to privacy, digital security, and public trust, being used in disinformation campaigns, non-consensual pornography, financial fraud, and political manipulation. The ease of access to tools that enable the creation of such content, coupled with their increasingly convincing appearance, has made *deepfake* detection a global priority [Malik et al. 2022].

In computer science, researchers have been developing automated methods to identify subtle traces left by GANs in manipulated images and videos. Among the most effective approaches are those based on Computer Vision and Machine Learning, particularly using Convolutional Neural Networks (CNNs). These models can extract texture patterns, inconsistencies, and visual artifacts imperceptible to the human eye, enabling the distinction between real and synthetic images [Arora and Soni 2021, Ankolekar et al. 2022].

However, detecting *deepfakes* is not trivial. Generative models are constantly improving, reducing artificial traces and making automated differentiation increasingly challenging [Heidari et al. 2024]. Moreover, model interpretability becomes a concern: in real scenarios, understanding why an image is classified as fake is essential, especially in legal or media applications [Khoo et al. 2022, Guarnera et al. 2022].

Beyond its technical relevance, this study also aims to promote reflection on the ethical, regulatory, and social challenges associated with the use of automated *deepfake* detection systems. Public trust in visual content is being continuously eroded, and digital

authentication mechanisms are becoming increasingly essential to preserve information integrity in contemporary society [TAVARES 2024].

In this context, this work aims to investigate machine learning and computer vision techniques applied to *deepfake* detection, focusing on developing effective and interpretable models. Public datasets, AI tools, and visualization techniques are employed not only to train accurate classifiers but also to understand their limitations and potential in addressing the growing issue of digital misinformation. The urgent need to develop secure and explainable technological solutions lies at the core of this study, aiming to keep pace with the rapid evolution of generative networks and mitigate their negative impacts on the global information ecosystem.

2. Related Work

Deepfake is a technology that uses *deep learning* algorithms to manipulate original digital content in order to produce fake but realistic-looking media, which can include audio, video, and photos. Several methods are available in the literature for creating *deepfakes*, such as GAN (*Generative Adversarial Network*) and *Autoencoders* [Chauhan et al. 2023]. One of the open-source tools available for generating *deepfakes* is Deep Face Lab [Perov et al. 2020].

Among GANs, there are: StyleGAN and StyleGAN2, both developed by NVIDIA, which generate highly realistic human faces; ProGAN, which progressively generates high-resolution images, including faces; and CycleGAN, which transforms images from one domain to another, converting sketches into realistic images [Karras et al. 2020].

Another way to create fake face images is through facial transformation (*face morphing*), which involves blending two or more facial images to create a hybrid face that retains characteristics from all input faces. These techniques can be implemented using the `dlib`[1] and `OpenCV`[2] libraries for marking, warping, and blending images to transform faces. And filters from AI-based apps and manual editing in software like Photoshop are widely used to create and manipulate facial images.

3. Methodology

The study was conducted using Python 3.10 on Google Colab with a Tesla T4 GPU. The experimental dataset was constructed by aggregating the *Real and Fake Face Detection* [CIPLAb 2020] and *Human Faces* [Gupta 2021] repositories, resulting in a total of **10,025 images** after the removal of duplicates and quality verification. The preprocessing pipeline involved resizing images to 128×128 pixels, grayscale conversion, and pixel normalization (0–1), followed by the application of *data augmentation* techniques (rotation, zoom, and horizontal flipping) to enhance model robustness and mitigate overfitting.

The primary architecture consists of a Convolutional Neural Network (CNN) incorporating *Conv2D*, *MaxPooling2D*, and *Dense* layers, utilizing ReLU activations and a Sigmoid output for binary classification. The Adam optimizer and binary cross-entropy loss were employed. For comparative analysis, traditional models (SVM, Logistic Regression, and Random Forest) were implemented using HOG-based feature extraction.

[1] `http://dlib.net/`
[2] `https://opencv.org/`

The evaluation was performed using 20% of the data held as a test set. The CNN-based model achieved an average accuracy of 94.2% in the test set, consistently outperforming classical methods, which achieved between 70% and 85% accuracy.

The metrics considered included accuracy, precision, recall, and F1-score. Error analysis revealed that most misclassifications occurred in low-sharpness images or in images with neutral facial expressions, which may hinder the detection of subtle artifacts typically present in fake images.

To investigate the network's decision-making process, the Grad-CAM (Gradient-weighted Class Activation Mapping) technique was employed to visualize the regions that contributed most significantly to the model's classification.

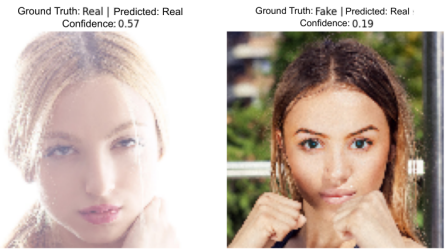

Figure 1. Visual samples utilized for model analysis. On the left, a real image correctly classified; on the right, a GAN-generated face. Source: the author.

In Figure 1, two distinct examples are presented. In real faces, the focus typically resides on well-defined facial structures such as the eyes and contours. In contrast, for images generated by GANs, the network's attention often concentrates on the eyes and mouth, where subtle synthesis inconsistencies frequently occur.

Beyond quantitative performance, Grad-CAM allows for an investigation of whether the model relies on semantically relevant features or spurious artifacts. In real images, activations concentrated on human facial morphology indicate that the network captured characteristic patterns of organic faces. In fake images, activations appeared more dispersed or focused on secondary regions, such as the transition between skin and hair, likely because artifacts introduced by adversarial processes emerge in these locations.

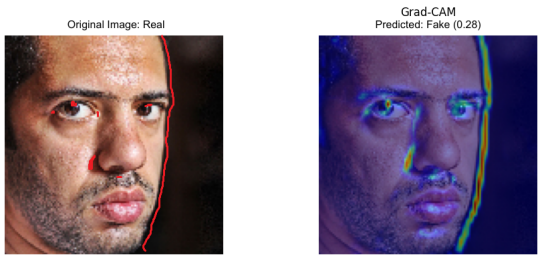

Figure 2. Grad-CAM visualization: Activation in non-informative regions during misclassification.

Figure 3. Grad-CAM visualization: Correct classification with semantic focus on facial features.

In Figure 2, the activation map confirms that when the model misclassifies an image with low confidence (0.28), it tends to focus on minimally informative regions.

Conversely, Figure 3 reinforces a more stable decision process, where the network classified the image as real with high confidence (0.95) based on significant facial cues.

4. Results and Discussion

This section presents a comprehensive quantitative and qualitative analysis of the evaluated models. The results are discussed in light of the experimental findings, focusing on predictive performance, error patterns, and the practical implications for deepfake detection.

The CNN-based model consistently outperformed classical machine learning approaches across all evaluated metrics. As summarized in Table 1, the proposed architecture achieved an accuracy of 94.2%, significantly higher than the 86.5% obtained by the SVM with HOG descriptors.

Table 1. Comparative performance of the models on the test set.

Model	Acc	Prec	Rec	F1
CNN (proposed)	0.942	0.945	0.939	0.942
HOG + SVM (RBF)	0.865	0.872	0.853	0.862
HOG + Random Forest	0.821	0.835	0.804	0.819
HOG + Log. Reg.	0.798	0.802	0.791	0.796

The high F1-score (0.942) indicates that the model maintains a robust balance between Precision and Recall, minimizing both false positives and false negatives. Additionally, the ROC-AUC value of 0.981 demonstrates the model's high discriminative capacity, effectively separating synthetic faces from authentic ones even under varying lighting conditions.

Despite the high accuracy, a detailed error analysis revealed that misclassification are often linked to specific image characteristics. False negatives predominantly occurred in low-sharpness images or those featuring neutral expressions with minimal facial artifacts. These cases often involve high-quality StyleGAN2 samples, which are known for their photorealistic textures that can bypass standard convolutional filters.

As shown in Figure 5, the output score distribution provides insight into the model's confidence. While the majority of samples are correctly clustered near the extremes (0.0 for fake and 1.0 for real), there is a noticeable overlap in the 0.4 to 0.6 range.

In these intermediate regions, the model exhibits lower confidence, likely due to the absence of clear generative artifacts or the presence of noise that mimics organic textures. This quantitative data supports the qualitative findings from the Grad-CAM analysis, reinforcing that the model is sensitive to semantically relevant facial features.

Although the focus of this work is technical, it is essential to highlight the ethical dilemmas associated with the generation and detection of deepfakes. Automated detection systems must be applied responsibly, respecting individual rights, privacy, and avoiding discriminatory biases in the models.

Poorly calibrated detectors may yield false positives with serious consequences, such as inappropriate censorship, defamation, or unjust accountability. Additionally, the

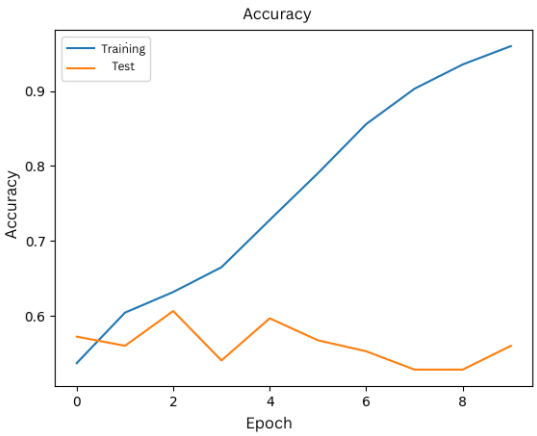

Figure 4. Training and testing accuracy evolution. The convergence after the 8th epoch suggests that the model reached its optimal learning capacity without overfitting.

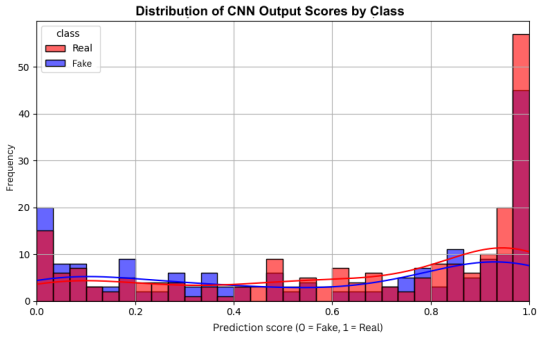

Figure 5. Distribution of CNN output scores. The clear separation between classes validates the effectiveness of the features learned during training.

concentration of such technology in the hands of a few entities raises concerns about informational monopoly and algorithmic governance.

Thus, the importance of practices such as external auditing, open-source repositories, and clear regulatory guidelines is reinforced. These measures are fundamental to ensure the safe, ethical, and transparent use of AI-based detection systems.

5. Conclusion

This study presents a comparative analysis of different approaches for automatic deepfake image detection, emphasizing computer vision and machine learning techniques. Through experimentation with classical methods based on handcrafted feature extraction—such as HOG combined with SVM, Random Forest, and Logistic Regression classifiers—and modern approaches using convolutional neural networks (CNNs), it was possible to observe the clear superiority of CNNs in predictive performance. The results confirm the state of the art reported in the literature, in which deep learning models, due to their ability to learn hierarchical representations directly from raw data, demonstrate higher accuracy and robustness in identifying fake images, even under variations in quality, lighting, and style. Nevertheless, the work also highlights challenges inherent in deepfake detection, such as the difficulty in classifying images generated by advanced models like StyleGAN2, and limitations in low-quality or neutral-expression scenarios. These findings emphasize the complexity of the problem and the need for approaches that combine accuracy with interpretability and adaptability. Future work may incorporate stronger evaluation protocols, including additional dataset splits to enhance result reliability. Finally, the importance of considering ethical and social aspects in applying these technologies is underscored. Automated deepfake detection carries risks of false positives and biases that may negatively affect individuals or groups, reinforcing the need for transparency, auditing, and regulation. Concluding, this work contributes to advancing knowledge in deepfake detection, confirming the effectiveness of CNNs as a reference method, and emphasizing the

importance of interpretability and ethical reflection in the development of technological solutions for information security and combating digital disinformation.

6. Scientific Integrity and AI Disclosure

Google Gemini large language model was employed solely for the purpose of grammar correction, spelling verification, and punctuation refinement of the English manuscript. No original scientific content was generated by the AI tool, and the authors maintain full responsibility for the technical integrity and conclusions presented in this work.

References

Ankolekar, A., Madappa, R., and Savakis, A. E. (2022). Can simpler be better? review of methods for the detection of gan-generated imagery. In *Defense + Commercial Sensing*.

Arora, T. and Soni, R. (2021). A review of techniques to detect the gan-generated fake images. In *Generative Adversarial Networks for Image-to-Image Translation*, pages 125–159. Elsevier.

Chauhan, R., Popli, R., and Kansal, I. (2023). A systematic review on fake image creation techniques. In *2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 779–783. IEEE.

CIPLAb (2020). Real and fake face detection. `https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection`. Acesso em jul.2025.

Guarnera, L., Giudice, O., and Battiato, S. (2022). The face deepfake detection challenge. *Journal of Imaging*, 8(10):263.

Gupta, A. (2021). Human faces. `https://www.kaggle.com/datasets/ashwingupta3012/human-faces`. Acesso em jul.2025.

Heidari, A., Jafari Navimipour, N., Dag, H., and Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Mining and Knowledge Discovery*, 14(2):e1520.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the quality of gans for image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:8107–8118.

Khoo, B., Phan, R. C.-W., and Lim, C.-H. (2022). Deepfake attribution: On the source identification of artificially generated images. *WIREs Data Mining and Knowledge Discovery*, 12(3):e1438.

Malik, A., Usama, M., Ali, M., and Razzak, I. (2022). Deepfake detection for human face images and videos: A survey. *IEEE Access*, 10:18757–18775.

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., Jiang, J., et al. (2020). Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*.

TAVARES, C. D. M. (2024). Inteligência artificial e deepfakes: Desafios jurídicos e tecnológicos para a integridade do processo democrático. *Revista Justiça Eleitoral em Debate*, 14(1).