International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Deepfake Detection: An Exploration of Deepfake Mediums

Rishita Kapile[1], Rajiv Surgoniwar[2], Swati Powar[3]

[1,2,3]MIT-Art, Design and Technology University, Pune

Abstract

The rapid advancement of deepfake technology, powered by artificial intelligence, has intensified the challenge of distinguishing authentic media from sophisticated synthetic manipulations, posing critical risks to digital security and trust. This paper presents a Full-Stack Deepfake Detection Application that leverages a 10-layer Deep Convolutional Neural Network (CNN) to identify deepfake images with exceptional accuracy. The model architecture integrates dilated convolutions to capture intricate spatial artifacts and employs dropout regularization (rate = 0.5) to mitigate overfitting, achieving robust generalization. Trained on a diverse dataset of real and synthetic images, the system attains >99.9% training accuracy and >99% validation accuracy, demonstrating high reliability in detecting state-of-the-art deepfakes. The application is implemented as an end-to-end solution, combining TensorFlow for model training, FastAPI for backend inference, and React.js for an intuitive frontend interface, enabling real- time image analysis. Users can upload media via a web interface and receive instantaneous classification results, supported by a confidence score. This research provides a helpful instrument to fight media scams, theft of identities, and disinformation by handling adaptability, automated processes, and real-time performance. The high accuracy and modular design of the suggested system highlight its potential for use in security- critical settings, boosting confidence in the legitimacy of digital media. Future updates may expand the reach of detection to incorporate asymmetrical attack resilience and audiovisual deep fakes.

Keywords: Deepfake Detection, Convolutional Neural Network (CNN), Dilated Convolutions, Dropout Regularization, Synthetic Media Manipulation, Binary Classification, Real-Time Image Processing, Digital Security, Media Authenticity, Deep Learning, Image Preprocessing, Scalable Systems.

Fig.1. Example of a Deepfake Image.

INTRODUCTION

The invention of machine intelligence (also called AI) has transformed the production of digital media by opening up previously unthinkable opportunities for the synthesis of images and videos. Deepfake the internet, a subset of AI- driven ways of developing synthetic media, has evolved into a double-edged sword among those advances. Whereas it has a possibility of utilization in the creative, virtual reality, and entertainment sectors, its improper use seriously compromises information integrity, security, and trust among the public. Deep Fakes have been used as a weapon for financial fraud, identity theft, propaganda for politics, and disinformation operations. They use GANs, which are generative adversarial networks, and autoencoders to edit or create hyper-realistic content. The need to create dependable, adaptable, as well as efficient systems for identification is currently a global issue as various synthetic media become more and more similar to real content.

Standard methods of discovering deepfakes, including based on regulations algorithms or personal diagnostic testing, rely on having the knack to distinguish visual irregularities such as strange face expressions or pixel-level variances in abnormal setting. nevertheless these strategies have lots of disadvantages, such as Rigid computations are unable to keep up with the fast growth of generative artificial intelligence (AI) systems, while human-driven research is inherently challenging, not objective, and inclined toward errors. In this regard, advanced fake architectures like used StyleGAN3 and DeepFaceLab yield results using less obvious errors, keeping current methods for detection worthless. The variation shows the critical nature of artificial intelligence-powered alternatives which take advantage of deep learning's sophisticated pattern-recognition capacities to keep current in the advancement of the media composed of artificial materials.

The current investigation demonstrates the Full-Stack Deep fake Sensor applied layer, an original structure designed to take on the challenges of current methodologies. The key role of the 10-in layers Deep Convolutional Neural System, or CNN, is for categorizing images as "real" or "fake." The recommended design incorporates convolutions with dilated areas that expand the mathematical model's sensitive field without enhancing interpreting overhead in comparison with normal CNNs.

This makes it possible the model to notice high-frequency imperfections and delicate spatial imperfections that are distinctive to deepfakes. Additionally, dense layers leverage regularization of drops (rate = 0.5) which enhances generalization, giving consistent outcomes across various kinds of datasets and manipulation procedures.

The creative thinking of the whole thing goes beyond their algorithmic layout. It uses a full-stack development framework, embracing React.js for an intuitive frontend interface, FastAPI for outstanding performance background inference, and TensorFlow for neural network training. Users may upload media to receive outcomes for classification with

>99% precision for validation in a couple of seconds thanks to these emancipation thus rendering real-time visual analysis conceivable. Scalability and versatility to the latest advances in deepfake production can be guaranteed by the application's scalable building construction, which consists of picture the preprocessing process extraction of attributes, model Inference is, and result visualisation. Recent data emphasizes how important this task is: a 2023 report with Deeptrace demonstrated a 330% yearly surge in malevolent deepfake written material, 96% of which involved monetary scams and non-consensual pornographic material. As illustrated by the numerous political deepfakes that influence public opinion during elections since social networking business entities have no capacity to control fraudulent media.

This type of technology delivers a proactive safeguarding measure against such threats by optimizing

detection and lowering must have on human expertise. This work adds to the global effort to reduce the threats posed by synthetic media by overcoming the gap amongst theoretically deep learning modifications and workable, deployable approaches. In an age of artificial intelligence (AI) erroneous it not only pushes boundaries of deepfake identification technology but also allows individuals, companies, and regulators the ability of safeguarding the digital landscape.

LITERATURE REVIEW

The swift development of counterfeit technology has called for equally quick improvements in detection techniques. This synthesis of the literature identifies enduring challenges, places the updates advocated in this work in context, and illustrates the journey from fundamental manual methods to more complicated AI-driven methods. Rule-based techniques and human skill were crucial elements of early studies to discover deepfakes. Videos were hand examined by criminal investigators for anomalies which involves uneven lighting, strange facial actions (such erratic eye blinking or humming mistakes), or variances in skin textures. Li et al. (2018), for example, designed a technique to detect cosmetic cornerstone contradictions, which are frequently present in early deepfakes as a result of subpar face reenactment algorithms. Although the above technique generated a moderate level of accuracy (~70%), it was not workable for applications of a large scale due to its requirement for step-by-step examination.
Governed by rules mechanisms, which emphasis on unique artifacts added during deepfake fabrication, are popping up as semi-automated alternatives. Agarwal et al. (2019) discovered the fact that generative strategies, such as automated encoders, have problem correctly representing high-resolution details, synthetic images regularly show abnormalities in the higher-frequency portions of the The Fourier range. Similar to the aforementioned, Matern et al. (2019) took pleasure in imperfections in eye reflections by making utilize of their discovery which deepfakes often come in short in their modeling of kerato thoughts. Nevertheless, these approaches had two major downsides:-

A. Overfitting to space artifacts

As computer graphics advanced, procedures that emphasized frequency inconsistencies or face markers became redundant. For instance, StyleGAN technique2 (Karras et al., 2020) greatly decreased obvious frequency abnormalities through the inclusion of gradual improvement and cacophony the procedure.

B. Lack of generalization

Rule-based systems performed poorly on datasets outside their training scope, failing to adapt to new manipulation techniques.
The advent of deep learning marked a paradigm shift, enabling automated feature extraction through hierarchical learning. Convolutional Neural Networks (CNNs) emerged as the dominant architecture due to their ability to capture spatial patterns in images.
Afchar et al. (2018) introduced MesoNet, a shallow 4-layer CNN designed to detect mesoscopic features (e.g., texture irregularities) in facial regions. While achieving 85% accuracy on the FaceForensics++ dataset, its limited depth restricted its capacity to identify subtle artifacts in high- resolution deepfakes. Rossler et al. (2019) benchmarked deeper architectures, notably XceptionNet (a 71-layer CNN adapted from ImageNet), which achieved 95% accuracy on the same dataset. However, XceptionNet's computational complexity (over 20 million parameters) made real-time inference challenging, especially on resource-constrained devices.
To address video-specific challenges, researchers combined CNNs with recurrent neural networks (RNNs). Nguyen et al. (2020) proposed a CNN-RNN hybrid to detect temporal inconsistencies, such as

flickering artifacts or unnatural head movements, in video sequences. While effective, the model's latency (≈2 seconds per frame) hindered real-time deployment.

As deepfakes grew more sophisticated, generalization—the ability to detect unseen manipulation techniques—became a key focus. Li et al. (2021) developed a self-supervised framework that identified universal artifacts across GAN architectures, achieving 88% accuracy on cross-dataset tests. Wang et al. (2022) integrated attention mechanisms into CNNs to prioritize manipulation-prone regions (e.g., eyes, teeth, and hairline). Their model achieved 92% accuracy on Celeb-DF, a dataset of high-quality deepfakes, but struggled with diffusion-based models like DALL-E 2 (Rombach et al., 2022), highlighting the need for adaptive architectures.

C. Unresolved challenges and future directions

- Real-Time Video Analysis: Most systems focus on images, neglecting the computational demands of video processing.
- Adversarial Attacks: Deepfake generators increasingly incorporate adversarial training to evade detection, as shown by Hussain et al. (2023).
- Adaptation to Emerging Techniques: Diffusion models (e.g., Stable Diffusion) produce fewer artifacts than GANs, necessitating new detection strategies.

METHODOLOGY

Existing approaches to deepfake detection broadly fall into two categories: **(1)** traditional machine learning methods relying on handcrafted feature engineering (e.g., texture descriptors, frequency-domain analysis) combined with classifiers like Random Forest or SVM, and **(2)** state-of- the-art deep learning architectures such as Vision Transformers (ViTs) or Xception CNNs, which automate feature extraction but demand significant computational resources and large datasets. Conventional methods struggle to extend to modern deepfake versions such as diffusion- model-generated material, despite prioritizing interpretability and reducing computing costs. On the contrary, methods based on deep learning have problems with scalability, overfitting on smaller data sets, and latency that makes them challenging for real-world deployment. However, they obtain great accuracy on large datasets.

By using a 10-layer Deep Convolutional Neural Network (CNN) augmented with dropout regularization and dilated convolutions, the proposed approach fills in these gaps. The CNN responds to changing artifacts without human intervention, in contrast to conventional ML techniques, by automated feature extraction. Dropout layers (rate=0.5) provide robustness against overfitting, addressing a major limitation of both models. The dilated convolutions broaden the model's receptive field to capture subtle spatial inconsistencies (such as irregular edge transitions, unnatural lighting gradients) that define contemporary deepfakes.

Transformer-based architectures and shallow CNNs. Through a lightweight FastAPI server and React.js frontend, the design enables immediate inference (<0.5 seconds) while achieving equivalent precision (>99% validation) with much lower computational overhead as as opposed to heavy on resources ViTs or Xception models. Additionally, by combining full-stack adaptability and data enhancement, the suggested approach gets around the prohibitive training requirements of huge transformers and the setup issues of conventional ML frameworks (such as Flask-based latency). Through balancing computing efficiency, automated feature development, and customizable delivery, this strategy performs better than traditional approaches in three crucial areas:

- **Generalization**: Robust performance in contrast to handmade features that are inefficient against

new artifacts, that extend a variety of datasets (e.g., FaceForensics++, Celeb-DF) and developing deepfake types (e.g., diffusion models).

- **Efficiency**: It is suitable for settings with limited computational infrastructure due to its lower training time and resource consumption when compared to ViTs.
- **Deployability**: Smooth introduction of real-time analytic attributes, resolving the old frameworks' scalability and latency issues.

In order to satisfy the demands of quickly changing synthetic media threats, our methodology creates a fresh framework for deepfake detection by finding a balance between interpretability, accuracy, and practicality.

Table I. Model Performance Comparison

Dataset	Different Models		
	Our Model	Baseline CNN	Traditional ML
FaceForensics++	99.84%	96.20%	88.50%
DFDC	99.38%	94.75%	85.30%
Celeb-DF	99.43%	93.60%	82.90%
VDFD(Custom)	99.66%	95.10%	87.40%

D. Data Collection and Preprocessing

The research makes use of well-known deepfake datasets, such as FaceForensics++, Celeb-DF, and DFDC, that include both real and fake photos and videos. Images were normalized to pixel values between 0 and 1 and expanded to a constant resolution of 224×224×3224×224×3 to guarantee standardization. By artificially increasing dataset diversity through the use of data augmentation techniques like random rotations, horizontal flips, and scaling, the capacity of the model to generalize over a broad spectrum of circumstances is improved. Key frames are consistently extracted for video analysis in order to capture the spatial irregularity seen in deepfake content.

E. Model Architecture Design

A 10-layer CNN that is designed to systematically extract features from input data forms the basis of the methodology. While mid-layers detect spatial irregularities (e.g., artificial facial symmetry or illumination gradients), initial convolutional layers detect low-level patterns, such edges and textures. The dilation convolutions used in the last layers expand the field of view without altering the parameters, making it possible to catch minute irregularities such as uneven edge transitions or background mismatches. After dense layers, dropout regularization (rate = 0.5) is used to reduce overfitting and guarantee robustness against unbalanced data. The output layer classifies data as true or fake using a sigmoid activation function.

F. Feature Extracion and Training

By automating feature extraction, the CNN do deals with the need for manually created descriptors (such HOG and Haralick) that are employed in conventional machine learning techniques. The Adam optimizer and the binary cross-entropy loss function are used for training, and a dynamically modified learning rate is used to maintain convergence. With regularization in batches, the model is trained over more than 50 epochs. To avoid overfitting, early stopping is used to stop training once validation accuracy hits a plateau.

G. Model Validation

Using 5×2-fold cross-validation is used to completely validate performance, guaranteeing robustness across various data splits. To assess the effectiveness of classification, metrics including accuracy, precision, recall, F1-score, and AUC-ROC curves are calculated. In order to evaluate generalization's capacity to adapt to new deepfake variations, such as those produced by dispersion models, it is further tested on unknown datasets (such as Celeb-DF).

H. Full Stack Deployement

Using React.js for front-end interaction and FastAPI for back-end processing, the system is built as an end-to-end solution. After processing uploaded media, the backend uses the trained CNN to do inference in real time and returns the results over RESTful API endpoints. Users can upload photographs, check classification results, and access scores for trust through an easy-to-use interface on the website. Scalability is guaranteed by this architecture, which can handle multiple queries with inference durations of less than 0.5 seconds, which is crucial to halting the propagation of viral deepfakes.

I. Performance Evaluation

By exhibiting >99% validation accuracy across datasets and computational economy, the model's effectiveness is compared to modern techniques. In order to demonstrate robustness to changing threats, robustness is further verified against examples of adversary and content generated through diffusion models.

RESULTS AND DISCUSSIONS

A. RESULTS

The proposed model achieves **>99% accuracy** across all datasets , outperforming baseline CNNs (93.6–96.2%) and traditional ML methods (82.9–88.5%). Precision and recall metrics exceed 99%, with F1-scores demonstrating balanced performance. For instance, on FaceForensics++, the model attains 99.84% accuracy, 99.80% precision, and 99.75% recall, misclassifying only 5 out of 2,000 test samples (Figure 3). Inference times remain consistently below 0.5 seconds, even under high user loads, validating real-time deployability. Training efficiency is notable, requiring **80% less time** than Vision Transformers (ViTs) while maintaining competitive accuracy.

B. Key Findings

Architectural Efficacy: Dilated convolutions enhance detection of subtle spatial artifacts (e.g., irregular textures), while dropout regularization ensures robustness against overfitting, as evidenced by stable training-validation accuracy curves.

Generalization: The model maintains 99.43% accuracy on Celeb-DF, a dataset with high-quality deepfakes, demonstrating adaptability to evolving synthetic media.

Scalability: The FastAPI-React.js integration supports concurrent user requests without latency spikes, addressing a critical gap in existing Flask-based systems.

Computational Efficiency: Training completes in <12 hours on a single GPU, contrasting sharply with ViTs (>24 hours), making the approach viable for resource-constrained environments.

C. Interpretation of Results

The proposed deepfake detection framework achieves >99% accuracy across multiple benchmark datasets , demonstrating its efficacy in identifying synthetic media. This performance stems from two key architectural innovations: dilated convolutions, which expand the receptive field to capture subtle spatial artifacts (e.g., inconsistent lighting gradients, unnatural edge transitions), and dropout

regularization, which prevents overfitting by randomly deactivating neurons during training. For instance, on the Celeb-DF dataset—a challenging benchmark with high-quality deepfakes—the model achieves 99.43% accuracy, outperforming traditional ML methods by 16.5%. This underscores the limitations of handcrafted features (e.g., HOG, Haralick) in adapting to modern generative techniques like diffusion models, which lack traditional GAN artifacts.

The model's real-time inference capability (<0.5 seconds per image) further distinguishes it from computationally intensive architectures like Vision Transformers (ViTs), which require >11 hours for training on similar datasets. This efficiency is critical for real-world deployment, where rapid detection is necessary to mitigate the viral spread of deepfakes on social platforms.

D. Practical Implications

The system's full-stack deployment (FastAPI + React.js) addresses a critical need for scalable, user-friendly deepfake detection tools. Unlike Flask-based systems (e.g., Paper 1), which suffer from latency spikes (>5 seconds), the proposed architecture handles concurrent requests with sub-second inference times, making it suitable for integration into content moderation pipelines (e.g., social media platforms).

Furthermore, the model's modular design allows seamless adaptation to new datasets. For example, extending the framework to video deepfakes requires only incremental changes (e.g., frame aggregation), as opposed to overhauling handcrafted feature pipelines.

E. Conclusion

This work establishes a **deployable, efficient, and accurate** deepfake detection framework. By integrating dilated convolutions, dropout regularization, and full-stack engineering, the methodology bridges the gap between theoretical robustness and real-world applicability. Future extensions will target video deepfakes and adversarial robustness, further solidifying its utility in combating synthetic media threats.By harmonizing architectural innovation (dilated convolutions, dropout) with full-stack engineering, the methodology advances the field toward practical, scalable solutions for combating synthetic media. Future work will focus on temporal modeling and adversarial robustness to further solidify its utility in evolving digital landscapes.

REFERENCES

1. Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). Attention augmented convolutional networks. In CVPR (pp. 3286– 3295).
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In ECCV (pp. 213–229).
3. Dzanic, T., Shah, K., & Witherden, F. (2020). Fourier spectrum discrepancies in deep network generated images. NeurIPS, 33, 3022– 3032.
4. Gao, P., Zheng, M., Wang, X., Dai, J., & Li, H. (2021). Fast convergence of DETR with spatially modulated co-attention. In CVPR (pp. 3621–3630).
5. Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. In CVPR (pp. 5039–5049).
6. Jiang, L., Li, R.,Wu,W., Qian, C., & Loy, C. C. (2020).
7. Deeperforensics1.0: A large-scale dataset for real-world face forgery detection. In CVPR (pp. 2889–2898).
8. Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In CVPR (pp. 4401– 4410).

9. Parmar, N., Vaswani, A., Uszkoreit, J., Kaiser, L., Shazeer, N., Ku, A., & Tran, D. (2018). Image transformer. In ICML (pp. 4055–4064).
10. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In CVPR (pp. 1–11).
11. Malik, A., Kuribayashi, M., Abdullahi, S.M., & Khan, A.N. (2022). DeepFake detection for human face images and videos: A survey. IEEE Access, 10, 18757–18775.
12. Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, QuocViet Pham, Cuong M. Nguyen (2022) Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, 103525, doi.org/10.1016/j.cviu.2022.103525.
13. Kaur, A., Hoshyar, A. N., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: Challenges and opportunities.
14. Tolosana, R, Vera-Rodriguez, R., Fierrez, J. et al (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inf Fusion, 64:131–148.
15. M. S. Rana, and A. H. Sung, "DeepfakeStack: A Deep Ensemble-based Learning Technique for Deepfake Detection," 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud), New York, USA, 2020, pp. 70–75.
16. D. Guera, and E. Delp, "Deepfake video detection using recurrent neural networks," in IEEE International Conference on Advanced Video and Signal Based Surveillance, 2018.
17. M. N. Murti and V. S. Devi, "Feature Extraction and Feature Selection, Introduction to Pattern Recognition and Machine Learning," IISc Lecture Notes Series, June 2015, pp. 75-110
18. R. M. Haralick, "Statistical and structural approaches to texture," I Proceedings of the IEEE, vol. 67, no. 5, pp. 786-804, May 1979 M. T. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, "Forensics and Analysis of Deepfake Videos," 11th International Conference on Information and Communication Systems, Jordan, 2020, pp. 053–058.
19. X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo, "Identity-Driven DeepFake Detection," arXiv preprint arXiv:2012.03930, 2020.
20. L. Bondi, E. D. Cannas, P. Bestagini, and S. Tubaro, "Training Strategies and Data Augmentations in CNN-based DeepFake Video Detection," arXiv preprint arXiv:2011.07792, 2020.
21. Z. Hongmeng, Z. Zhiqiang, S. Lei, M. Xiuqing, and W. Yuehan, "A Detection Method for DeepFake Hard Compressed Videos based on Super-resolution Reconstruction Using CNN," Proceedings of the 4th High-Performance Computing and Cluster Technologies Conference & 3rd International Conference on Big Data and Artificial Intelligence, Association for Computing Machinery, New York, USA, pp. 98–103.
22. J. Han, and T. Gevers, "MMD Based Discriminative Learning for Face Forgery Detection," 15th Asian Conference on Computer Vision, Kyoto, Japan, 2020, pp. 121–136.
23. H. Dang, et al., "On the Detection of Digital Face Manipulation," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020, pp. 5780–5789.