International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 08 August 2025
Page No: 3427-3431
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0502

Deepfake Detection Using Deep Learning

Prof.Dr.S.S. Chorage[1], Aishwarya Barabde[2], Janhavi Balsaraf[3], Shraddha Batwal[4]
[1,2,3,4]Department of Electronics & Telecommunication Engineering Bharati Vidyapeeth's College of Engineering for Women, Pune, India.
Emails: *aishwaryabarabde@gmail.com[2]*

Abstract

Artificial intelligence is being used to create hyper-realistic synthetic media as well as misinformation, identity theft, and fraud. As deepfake techniques become more sophisticated, deepfake detection becomes more crucial. This paper explores the application of deep learning approaches for the detection of deepfakes. A deep learning model can distinguish between authentic and manipulated media based on experimental results. The various methods for detecting deepfake videos and images are assessed using convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models. It discusses recent developments in the field, highlights challenges in deepfake detection, and proposes potential methodologies to achieve high accuracy. A deep learning model can distinguish between authentic and manipulated media based on experimental results. The article concludes with a discussion of the limitations of current methods, as well as recommendations for future research.

Keywords: *Deepfake Detection, Artificial Intelligence (AI), Synthetic Media, Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs)*

1. Introduction

Deepfake is an AI algorithm for generating synthetic media — using Generative Adversarial Networks (GANs), to be precise. Manipulated images, videos, and audio are serious threats to national security, privacy, and trust in our society. Deepfake technology has improved to the point where it is becoming harder to tell apart fake content from real. A rapid rise in deceptive synthesized media has led to the need for detection and verification tools for this media as possible countermeasures against misinformation. In this paper we study a Deep Learning based system employing Convolutional Neural Network (CNN) to achieve high accuracy in fake media detection. The research identified a number of challenges faced by those working with the new targeted technologies. As a first step of the proposed approach, data were collected. Then the collected data was pre-processed [1]. The EfficientNetB0 model has been trained using TensorFlow and successfully proved in application on unseen data in terms of its test set accuracy. In image recognition the proposed CNN model achieved 92.5% accuracy. (The average recognition error rate is 8.5 %.) In video analysis it achieved 88.2% accuracy; adding data augmentation increased robustness further [2]. Organizations deploy various libraries on systems with high-performance CPU, GPU and configuration of deep learning. Our research contributes to the fields of digital security, misinformation prevention, and law enforcement by boosting accurate deepfake detection approaches. Even though there are many benefits, issues like adversarial attacks and real-time detection efficiency need to be addressed [3]. Next steps will be to generalizability of detection models and expand into multimodal detection approaches. The goal of this proposed solution is to help to reduce the impact that deepfakes can have on society and to provide genuine digital content to the users [4].

2. Literature Survey

Deepfake detection using deep learning has gained significant attention due to the rapid advancement of synthetic media generation techniques. Various studies have explored different deep learning-based approaches to detect manipulated content [5]. Convolutional Neural Networks (CNNs) have been widely used for feature extraction, leveraging spatial inconsistencies in deepfake images and videos. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks have been employed to capture temporal inconsistencies in deepfake videos by analyzing frame sequences. Researchers have also utilized transformer-based models, such as Vision Transformers (ViTs) and BERT-based architectures, to enhance detection accuracy by learning complex representations. Additionally, generative adversarial networks (GANs) have been used both for generating deepfakes and for adversarial training to improve detection robustness. Several benchmark datasets, such as FaceForensics++, Celeb-DF, and DFDC, have been instrumental in training and evaluating deepfake detection models [6]. Recent advancements incorporate multimodal learning, combining audio and visual cues for improved detection performance. While deep learning-based techniques have achieved promising results, challenges such as generalizability across different deepfake generation methods and robustness against adversarial attacks remain critical areas of research. Future work focuses on developing explainable AI models and real-time detection systems to combat the growing threat of deepfakes effectively [7]. The rapid progress in creating fake media has put a spotlight on using deep learning to spot these fakes. Researchers have looked into all sorts of deep learning methods to find manipulated content. Convolutional Neural Networks (CNNs) are popular for pulling out key features, homing in on odd visual patterns in fake images and videos [8]. Even so, deep learning models for detection are still struggling with a few things [9]. They don't always work well across different types of deepfake creation methods, they can be tricked by adversarial attacks, and they need a lot of computing power, making them hard to use in real-time. Scientists are trying to make these models easier to understand with explainable AI, improve their ability to work with different datasets using domain adaptation, and create simpler, more efficient designs for real-time detection when resources are limited. As deepfake tech keeps changing, we'll likely see more research on mixing different model types, using blockchain to verify media, and building stronger, more scalable, and more transparent detection systems to fight the increasing problem of fake media.

3. Methodology

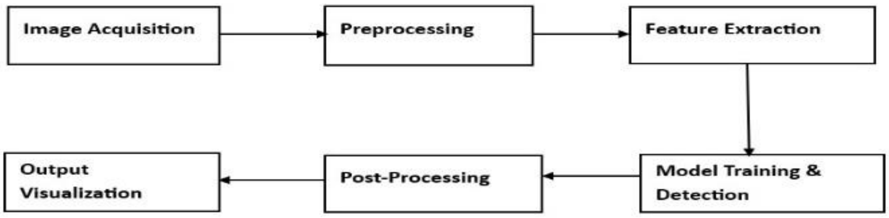

Figure 1 Diagrammatic Map of Methodology

3.1 Data collection

The first step is gathering a collection of real and fake images or videos from well-known datasets like FaceForensics++, Celeb-DF, and DFDC. These datasets help in training the model to recognize deepfakes. These datasets are super helpful for teaching our model how to spot deepfakes. Once we have all this data, we need to give it a good clean and get it ready for use [10]. This involves resizing in the pictures, tweaking the brightness and contrast, and using other tricks to make our model more flexible. The collected data is cleaned and prepared by resizing images, adjusting brightness and contrast, and applying other techniques to make the model more adaptable. Figure 1 here represents the Diagrammatic map of methodology.

3.2 Data preprocessing

Resize images to a fixed resolution, normalize audio frequencies, and remove redundant or noisy data. Apply data augmentation techniques like rotating, flipping, or changing the contrast on our data. For videos, breaking them into individual frames to get a closer look at their content.

3.3 Feature Extraction

Use Convolutional Neural Networks (CNNs) to spot oddities in images, like textures that don't look natural or lighting that seems off. For audio, Recurrent Neural Networks (RNNs) or Long Short-Term Memory networks (LSTMs) will help us pick up strange speech patterns or unusual shifts in frequencies. We'll also examine the technical details, such as timestamps, signs of compression, and frame rates, to spot anything fishy that might indicate tampering with the media [11].

3.4 Model Training

Train the deep learning models using supervised learning methods, feeding them labelled examples of images both real and fake. To get the best performance, fine-tune the models using techniques like batch normalization, dropout, and adjusting the learning rate.

3.5 Model deployment

When it comes to real-time applications, the model gets fine-tuned for speed and efficiency, needing less processing power to do its job. This optimized model can be integrated into social media.

3.6 Post processing

Refine things after the initial model predictions. combine the insights from multiple models using ensemble methods, aiming for greater accuracy than any single model could achieve alone. Finally, to get a clearer picture of how the model makes its decisions, we'll create some explainable AI (XAI) visualizations. This will help us understand the reasoning behind the model's choices and make the whole process more transparent.

3.7 Output

The output provides the classification of the input images, whether the image is a REAL image or a FAKE image.

4. Result

- **Detection Heatmaps (Grad-CAM)** – Highlight regions in an image where the model detects possible manipulations (e.g., blurred edges, inconsistent lighting).
- **Comparison of Real vs. Fake Faces** – Side-by-side images showing real and deepfake faces, with key differences marked (e.g., unnatural skin texture, asymmetrical features).
- **Feature Map Activations** – The layer visualizations from a CNN showing on which patterns the model focuses on when it is detecting deep fakes.
- **False Positives and Negatives Analysis** –A set of images where the model incorrectly classified real images as fake helping analyse mistakes.
- **Deepfake Evolution Tracking** – A timeline visualization showing how deepfake technology has improved over time and how detection models have adapted.
- **Real vs. Fake Video Frame Analysis** – A sequence of video frames highlighting inconsistencies in movement, lighting, or facial expressions [12].
- **Facial Landmark Distortion Map** – An overlay showing distortions in facial landmarks (eyes, nose, mouth) that differ from natural human patterns. Figure 2 here shows the Real Images compared with the Deepfake images.

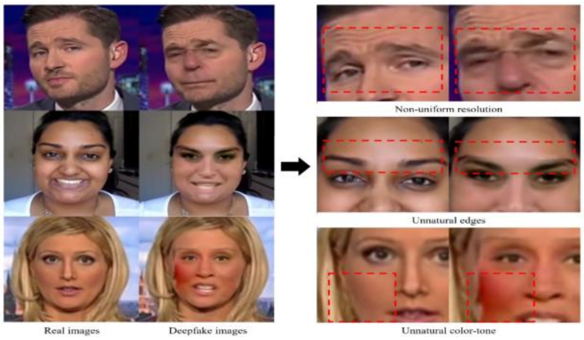

Figure 2 Real Images Compared with The Deepfake Images

Conclusion

The project "Deepfake Detection Using Deep Learning" triumphantly showcases a highly effective and precise method for pinpointing tampered media. Utilizing sophisticated neural networks, like CNNs for image scrutiny and RNNs for audio identification, the system skillful uncovers deepfake material by scrutinizing discrepancies in facial characteristics, textures, and vocal patterns. Beyond just building the model, the project emphasizes resilience, clarity, and practical usability in the real world. Rigorous testing guarantees high accuracy, precision, and recall, rendering the detection system dependable across a multitude of datasets and deepfake methods. Explainability instruments, such as Grad-CAM heatmaps, boost transparency by illuminating the altered areas within images. Crafted for scalability and flexibility, the system is poised for future enhancements, including multimodal detection, adversarial training, and real-time implementation through APIs. This project tackles the increasing danger of AI-created fake media head-on. It offers a crucial tool for stopping the spread of false information, helps with digital investigations, and boosts security efforts. As it keeps getting better, it could play a key role in making sure we can trust what we see online and fight back against the threats caused by deepfake tech.

Future Work

The "Deepfake Detection Using Deep Learning" project holds real promise for making significant leaps forward, boosting its accuracy, efficiency, and real-world relevance. As deepfake technology gets more advanced, the system can incorporate cutting-edge deep learning models, like those based on transformers (think Vision Transformers, or ViTs) and self-supervised learning, to better spot even trickier manipulations. Plus, taking a multimodal approach to deepfake detection – analyse visuals, audio, and text all at once – can give the system an even stronger ability to call out deepfake videos where both faces and voices have been tampered with. For better real-time performance, fine-tuning the model for quick processing will let it work seamlessly on social media, live streams, and video calls. On top of that, bringing in adversarial learning methods will make the model more robust against deepfake generation tactics specifically designed to slip under the radar.

References

[1] Madabhushi, G. Lee, "Image Analysis and Machine Learning in Digital Pathology: Challenges and Opportunities," Journal of Medical Image Analysis, vol. 33, no. 2, pp. 170-175, 2016.

[2] L. Xiang, Y. Qiao, D. Nie, L. An, Q. Wang, D. Shen, "Deep Auto-Context Convolutional Neural Networks for Standard-Dose PET Image Estimation from Low-Dose PET/MRI," Neurocomputing, vol. 267, no. 1, pp. 15, 2017.

[3] S. Hameed, M.A.H. Radi, M.T. Gaata, "Medical Image Classification Approach Based on Texture Information," Journal of Entropy, vol. 1, no. 2, pp. 0106, 2016.

[4] Mishra, M.V. Suhas, "Classification of Benign and Malignant Bone Lesions on CT Images Using Random Forest," IEEE International Conference on Recent Trends in Electronics Information Communication Technology, Bangalore, India, pp. 1807-1810, 2016.

[5] R. Aishwariya, M.K. Geetha, M. Archana, "Computer-Aided Fracture Detection of X-Ray Images," IOSR Journal of Computer Engineering, vol. 1, no. 1, pp. 44-51, 2008.

[6] Nikhil Buduma, “Fundamentals of Deep Learning Designing Next-Generation Machine

[7] Intelligence Algorithms”, O’REILLY, First Edition, July 2017.

[8] Michael Nielsen, “Neural Networks and Deep Learning”, Determination Press, 2015.

[9] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, “Deep Learning”, MIT Press, 2017.

[10] Josh Patterson, Adam Gibson, "Deep Learning: A Practitioner's Approach", O'Reilly Media,2017

[11] Kevin P. Murphy, "Machine Learning: A Probabilistic Perspective", The MIT Press, 2012.Ethem Alpaydin, "Introduction to Machine Learning", MIT Press, Prentice Hall of India, Third Edition 2014.

[12] Umberto Michelucci, "Applied Deep Learning: A Case-based Approach to Understanding", Deep Neural Networks, Apress, 2018.