International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 12, pp 145-150, December-2024

https://doi.org/10.47001/IRJIET/2024.812022

Video Deepfake Detection Using EfficientNet

[1]Satwika.M, [2]Pranavya.A, [3]Neha.K, [4]Rishika.K, [5]Siva Sankar Namani

[1,2,3,4]Department of CSE (AI & ML), G. Narayanamma Institute of Technology and Science, Hyderabad, India
[5]Assistant Professor, Department of CSE (AI & ML), G. Narayanamma Institute of Technology and Science, Hyderabad, India

***Abstract* - Deepfake detection has become a critical area of research with the growing prevalence of sophisticated face manipulation technologies, which pose severe ethical and security challenges. In this study, we propose an advanced deepfake detection system leveraging the EfficientNetB0 model, a state-of-the-art convolutional neural network (CNN) architecture, to address the challenges of efficiency and accuracy in identifying manipulated media. Our system utilizes video frame extraction and comprehensive data augmentation techniques to preprocess inputs, ensuring enhanced generalization on limited training data. EfficientNetB0, pre-trained on the ImageNet dataset, serves as the backbone for feature extraction, employing its highly efficient architecture with depth wise separable convolutions. Evaluated on the Celeb-DF dataset, the proposed system demonstrates high accuracy and robustness in detecting deepfake content while maintaining computational efficiency, making it suitable for real-world applications. Experimental results validate the effectiveness of this approach, highlighting its potential to contribute significantly to mitigating the adverse impacts of deepfakes.**

Keywords: Convolutional neural networks, CNN, EfficientNet, DeepFake detection.

I. INTRODUCTION

Deepfakes refer to synthetic media content, such as images, videos, or audio, generated using advanced artificial intelligence techniques, predominantly deep learning. They leverage neural networks to alter or fabricate visual and auditory data convincingly. This technology has garnered significant attention due to its potential to manipulate reality, posing threats to authenticity, privacy, and societal trust. The term "deepfake" originates from a combination of "deep learning" and "fake," highlighting the AI-driven approach behind this innovation.

Modern deepfake techniques can be classified into advanced methodologies that manipulate facial features and expressions:

Face2Face: This real-time facial reenactment approach uses 3D facial modeling and texture mapping to replicate a source actor's expressions onto a target individual’s face. By transferring movements while retaining the target's identity, Face2Face enables realistic video manipulation suitable for real-time applications.

Face Swap: As the name suggests, Face Swap technologies replace one person’s facial features with another’s in a seamless manner. It relies on landmark detection, facial segmentation, and advanced neural networks to preserve facial alignment, texture, and lighting, making it a widely used technique for entertainment and malicious activities alike.

Neural Textures: Neural Textures enhance the photorealistic quality of manipulated videos by generating AI-driven texture maps. These maps improve intricate details such as skin patterns, lighting, and dynamic expressions, making the output indistinguishable from real footage. This approach has significantly elevated the visual realism of deepfakes.

Deepfake generation predominantly utilizes Generative Adversarial Networks (GANs), a specialized class of neural networks that consist of two key components: a generator and a discriminator. The generator is responsible for creating synthetic media, such as faces or voices, designed to closely mimic authentic data. Meanwhile, the discriminator assesses the authenticity of the generated content, distinguishing between real and fake inputs, and provides feedback to the generator for iterative improvement. This adversarial training process continues until the generator produces highly realistic outputs that are virtually indistinguishable from genuine data.

This adversarial process continues until the generator produces highly realistic outputs indistinguishable from real data. Variants like StyleGAN and CycleGAN specialize in creating high-resolution content and domain-specific transformations, further pushing the boundaries of deepfake realism. The iterative nature of GANs, combined with large-scale datasets, makes them the preferred architecture for deepfake creation.

Deepfakes have evolved from a technological novelty into a significant societal concern, with their misuse causing profound ethical, economic, and security challenges. One of the most alarming implications is the erosion of trust through disinformation campaigns, where fabricated videos of political

ISSN (online): 2581-3048
Volume 8, Issue 12, pp 145-150, December-2024
https://doi.org/10.47001/IRJIET/2024.812022

leaders announcing false policies or military actions have sown panic and misinformation, undermining confidence in legitimate media. Privacy violations are another critical issue, as deepfake technology has been exploited to create non-consensual explicit content, affecting both celebrities and individuals, resulting in reputational damage, psychological harm, and legal disputes.

Economically, deepfakes pose cybersecurity threats, such as financial fraud through impersonation of executives to authorize unauthorized transactions, and voice cloning in sophisticated spear-phishing attacks, jeopardizing organizational security. On a broader scale, deepfakes threaten national and global security, with the potential to fabricate videos of world leaders announcing military aggression or resignations, which could incite geopolitical instability and crises.

II. LITERATURE SURVEY

Recent studies in deepfake detection have focused on leveraging advanced deep learning methods to identify forged media by analyzing manipulation traces and extracting discriminative features. Li et al. (2018) identified a unique anomaly in GAN-generated videos, noting the absence of human eye blinking, and developed a method leveraging this observation for detection. Sabir et al. proposed a spatio-temporal RCNN framework, which effectively captures manipulation inconsistencies across video frames by extracting temporal and spatial features. However, the method is computationally expensive and limited to video-based detection. Guera and Delp extended this idea by combining CNNs for feature extraction with RNNs to model temporal dependencies, which improved detection in video content but came with significant computational overhead. Wang et al. (2020) shifted the focus to static image detection by identifying common artifacts in CNN-generated images, utilizing a ResNet-50 architecture pre-trained on ImageNet. Their method achieved high accuracy but struggled to generalize beyond specific image contexts, demonstrating a narrower scope.

Further, Yang et al. explored head pose inconsistencies in deepfake videos and classified them using Support Vector Machines (SVMs). While this approach highlighted subtle discrepancies in head movements, its performance was limited by dataset availability and errors in pose estimation. Chen and Yang developed a CNN-based manipulated face detector designed to identify tampered faces under varying conditions, such as illumination and pose, but faced challenges in achieving robustness across diverse datasets. Durall et al. introduced a novel approach by shifting from the spatial to the frequency domain, employing the power spectrum as a forensic feature to detect deepfakes, especially in low-resolution media. Their work demonstrated promising results in uncovering subtle manipulation traces but faced limitations in low-light scenarios. These methodologies, evaluated on prominent datasets like FaceForensics++, Celeb-A, and ForenSynths, underscore the rapid advancements in deepfake detection while highlighting ongoing challenges in scalability, robustness, and computational efficiency.

Despite significant advancements in deepfake detection, numerous challenges persist, limiting the effectiveness and scalability of current methods. A key issue is the lack of generalization across datasets, with many models performing well on specific datasets like FaceForensics++ or Celeb-A but struggling with unseen or diverse data. Computational complexity is another major concern, as methods leveraging deep neural networks, such as CNNs and RNNs, often require substantial computational resources, making them less viable for real-time applications. Robustness to variations in pose, lighting, and resolution further complicates detection, as many algorithms fail to adapt to such inconsistencies in manipulated media. Additionally, low-resolution deepfakes and subtle manipulations in the spatial or frequency domain pose significant detection challenges. The rapid evolution of generative models also means that detection techniques can quickly become outdated, necessitating continuous updates. These limitations collectively highlight the need for scalable, robust, and computationally efficient solutions in the fight against deepfake-generated media.

III. PROPOSED SYSTEM

In this work, we propose a novel system for deepfake detection using a Convolutional Neural Network (CNN), specifically leveraging the EfficientNetB0 model, on the Celeb-DF dataset. The system utilizes EfficientNetB0, a state-of-the-art pre-trained model, as the backbone for feature extraction, followed by classification for detecting real vs. deepfake videos. The primary goal is to classify the authenticity of video frames, with the system optimized for both efficiency and accuracy in detecting manipulated content.

System Overview

The proposed deepfake detection system is structured in two main phases:

1. Data Preprocessing and Augmentation:

Video Frame Extraction: The first step involves extracting individual frames from video files. Since deepfake videos often show subtle inconsistencies in each frame, extracting frames is crucial for isolating these artifacts and analyzing them.

Data Augmentation: Given the limited availability of labeled deepfake data, data augmentation techniques are employed to artificially expand the training dataset. This includes rotation, shifting, shearing, zooming, and flipping, which helps the model generalize better and prevents overfitting.

2. Feature Extraction and Classification using EfficientNetB0:

EfficientNetB0 Model: The EfficientNetB0 model, pre-trained on the ImageNet dataset, serves as the backbone of the proposed system. EfficientNetB0 is known for its efficiency in terms of both model size and performance, providing an excellent trade-off between accuracy and computational resources.

Fine-tuning: We use EfficientNetB0 without its top (classification) layers. The feature extraction part of the model is frozen to retain the pre-trained weights, while the top layers are customized for the binary classification task (real vs. deepfake).

Global Average Pooling (GAP): After extracting features using EfficientNetB0, a Global Average Pooling (GAP) layer is applied to convert the 2D feature maps into a 1D feature vector. GAP helps reduce the dimensionality while retaining the essential spatial information.

Fully Connected (Dense) Layer: The extracted features from the GAP layer are passed through a dense layer with a sigmoid activation function, outputting a probability score that classifies each frame as either real or deepfake.

3. Training and Evaluation:

Loss Function and Optimizer: The model is trained using binary cross entropy as the loss function, appropriate for a binary classification task. The Adam optimizer is employed with a learning rate of 0.0001, which helps fine-tune the model effectively.

Model Evaluation: After training, the model is evaluated on a separate test set of frames, measuring its performance using accuracy, loss, and other metrics to assess its capability in distinguishing real from fake content.

Detailed System Architecture

Input Layer: The system takes input video frames resized to 128x128x3 dimensions (height, width, and color channels).

EfficientNetB0 Backbone: The backbone is EfficientNetB0, which consists of the following key blocks:

Stem: Initial convolution followed by batch normalization and Swish activation.

MBConv Blocks with Squeeze-and-Excitation (SE): These blocks involve depthwise separable convolutions and utilize the Squeeze-and-Excitation technique to adaptively recalibrate channel-wise feature responses.

Global Average Pooling (GAP): Converts the final feature maps into a 1D vector of features that represent the image content.

Fully Connected (Dense) Layer: After feature extraction, the model uses a dense layer with sigmoid activation for binary classification (real vs deepfake). This layer outputs a probability value (0 for real, 1 for deepfake).

Output: The output of the network is a single scalar representing the probability of a frame being real or deepfake.

Advantages of the Proposed System

Efficiency: By using EfficientNetB0 as the backbone, the model benefits from a highly optimized architecture, achieving superior accuracy with fewer parameters compared to other deep learning models.

Data Augmentation: The inclusion of data augmentation techniques ensures better generalization by increasing the diversity of the training data.

Pre-trained Weights: Using pre-trained weights from ImageNet helps the system leverage learned features that are useful in detecting even subtle discrepancies in deepfake content.

Scalability: The system is scalable to handle large datasets, and EfficientNetB0's lightweight nature makes it suitable for real-time applications.

Experimental Setup

Dataset: The Celeb-DF dataset is used, which contains deepfake videos of celebrities. This dataset includes both real and fake video frames, allowing for binary classification.

Evaluation Metrics: The model's performance is evaluated using:

- Accuracy: The percentage of correctly classified frames.
- Precision, Recall, and F1-Score: These metrics give a more detailed view of the model's ability to detect real and deepfake content.
- Confusion Matrix: A matrix that shows the true positives, false positives, true negatives, and false negatives.

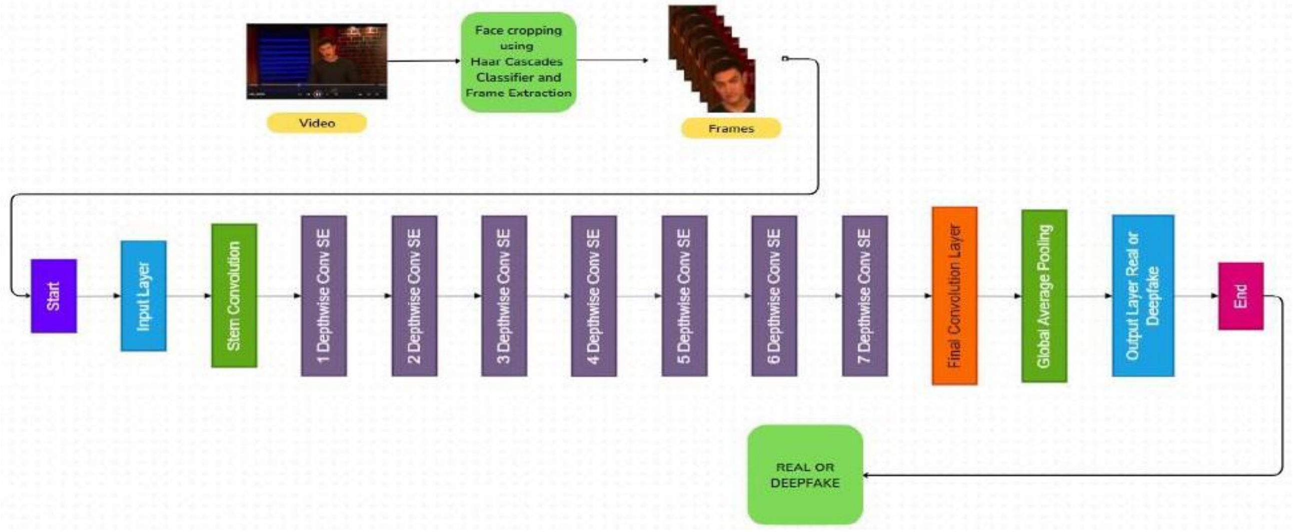


Figure 3.1: Architecture of the Proposed System

Figure shows the architecture of the proposed system.

IV. RESULTS

The proposed EfficientNetB0-based deepfake detection system achieved significant performance on the Celeb-DFdataset. On the test set, the model obtained an accuracy of 99.27%, a precision of 100.0%,Recall of 98.85% and an F1-Score of 99.42%. The confusion matrix reveals that the model correctly classified 810 real videos and 1,380 deepfake videos, with no false positives and only 16 false negatives. This demonstrates the model's exceptional ability to accurately distinguish between real and deepfake videos, as reflected in its high precision and recall. The absence of false positives ensures reliability in detecting real videos, while the minimal false negatives highlight the system's robustness in identifying deepfakes. Overall, the confusion matrix supports the effectiveness of the proposed system in deepfake detection.

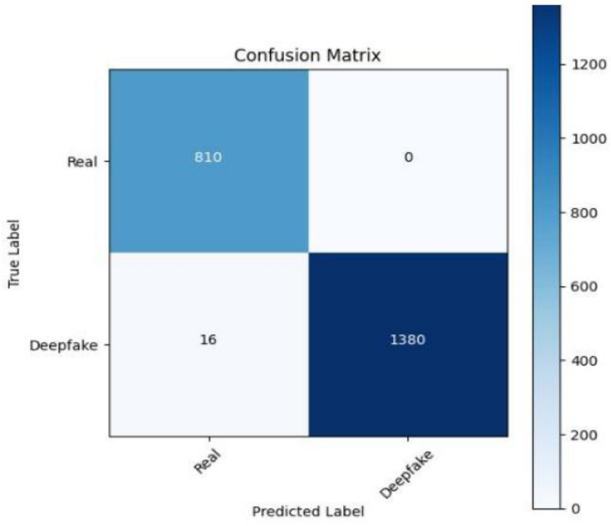
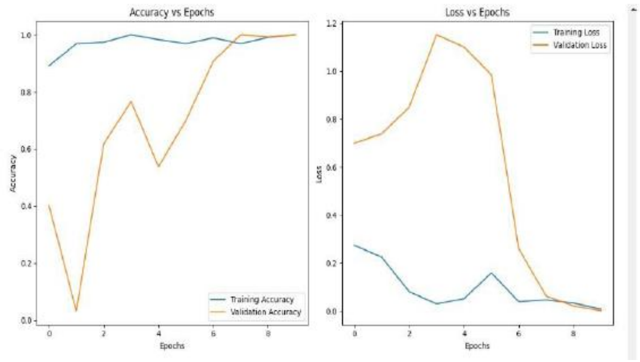


Figure 4.1: Confusion matrix of predictions

The model's performance was evaluated over 10 epochs, and the results are summarized in two graphs: Accuracy vs. Epochs and Loss vs. Epochs. The Accuracy vs. Epochs graph (on the left) shows a steady improvement in training accuracy, starting at approximately 0.9 and approaching 1.0 as training progresses. The validation accuracy, on the other hand, exhibits some fluctuations in the initial epochs, where it starts low and even drops to near 0.0 at one point. However, it rises significantly after the second epoch and converges closely with the training accuracy, achieving nearly 1.0 by the final epoch. This indicates that while the model experienced some instability during early training, it eventually learned to generalize well on the validation dataset.

The Loss vs. Epochs graph (on the right) provides further insight into the model’s training behavior. The training loss decreases consistently throughout the epochs, starting from a value of approximately 0.3 and reaching near 0.0 by the final epoch. In contrast, the validation loss initially rises, peaking around the third epoch, which may indicate early signs of overfitting. However, as the training progresses, the validation loss decreases sharply and aligns closely with the training loss near the final epochs. This sharp decline suggests that the model successfully mitigated overfitting and achieved strong generalization performance.

Overall, the convergence of both accuracy and loss curves for the training and validation datasets demonstrates that the model effectively learned the underlying patterns in the data. The final results reflect high accuracy and minimal loss, indicating that the model performs well and generalizes effectively to unseen validation data.



(a) **(b)**

Figure 4.2: (a) Accuracy vs Epochs graph, (b) Loss vs Epochs graph

V. CONCLUSION AND FUTURE WORK

In this study, we implemented a deep learning-based approach for video deepfake detection using the EfficientNet-B0 architecture. The model demonstrated strong performance, with both training and validation accuracies. For future work, we aim to improve the model's generalization capability by incorporating a more advanced classification strategy. Specifically, we plan to integrate a Long Short-Term Memory (LSTM) network or explore other powerful classifiers that can better capture temporal relationships and subtle patterns within video sequences. This enhancement could enable the model to process sequential frames more effectively and detect inconsistencies that are characteristic of deepfake videos. Furthermore, we intend to expand the dataset with a wider variety of deepfake examples to enhance the model's robustness across diverse scenarios and video qualities.

REFERENCES

[1] Agarwal S, Farid H, Fried O, and Agrawala M (2020) Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA.

[2] Amerini I, Galteri L, Caldelli R, and Del Bimbo A (2019) Deepfake Video Detection through Optical Flow Based CNN. In: IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South).

[3] Ding X, Raziei Z, Larson E.C. et al. (2020).

[4] Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer C (2019) The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv:1910:08854.

[5] Fei J, Xia Z, Yu P, Xiao F (2020), Exposing AI-generated videos with motion magnification. Multimedia Tools Appl. 80(20), 30789–30802.

[6] Hashmi M.F., Ashish B.K.K., Keskar A.G., Bokde N.D., Yoon J.H., Geem Z.W. (2020) An Exploratory Analysis on Visual Counterfeits Using Conv-LSTM Hybrid Architecture. IEEE Access 8, 101293–101308.

[7] Hsu C.-C., Zhuang Y-Xiu., Lee C.-Y. (2020) Deep fake image detection based on pairwise learning. Applied Sciences 10(1), 370.

[8] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, Isao Echizen (2019) Multi-task Learning For Detecting and Segmenting Manipulated Facial Images and Videos. In: IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS).

[9] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian (2018) Not made for eachother–Audio-Visual Dissonance-based Deepfake Detection and Localization. In Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05.

[10] Li, Haodong; Li, Bin; Tan, Shunquan; Huang, Jiwu: Identification of deep network generated images using disparities in color components. Signal Process. (2020).

[11] Li Y, Lyu S (2018) Exposing Deep Fake Videos By Detecting Face Warping Artifacts. In: IEEE Conference Computer. Vision Pattern Recognition.

[12] Li, Y.; Chang, M.; Lyu, S. (2018) In Ictu Oculi Exposing AI Created Fake Videos by Detecting Eye Blinking. In: IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, Hong Kong.

[13] Matern F, Riess C, Stamminger M (2019) Exploiting Visual Artifacts to Expose Deep fakes and Face Manipulations. In: IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA.

[14] Matthews T.F., Cootes J.A., Bangham S.C., Harvey R. (2002) Extraction of visual features for lip reading. IEEE Trans. Pattern Anal. Mach. Intell. 24(2), 198–213.

[15] zingxing Tan, Quoc V. Le (2019) EfficientNet: Re-thinking Model Scaling for Convolutional Neural Networks. Cornell University.

[16] Minh Dang L., Hassan S.I., Im S., Moon H. (2019) Face image manipulation detection based on a convolutional neural network. Expert Syst. Appl. 129, 156–168.

[17] Montserrat D.M. et al. (2020) Deepfakes Detection with Automatic Face Weighting In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA.

[18] Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. (2018) FaceForensics: A large-scale video dataset for forgery detection in human faces.

[19] Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. (2019) FaceForensics++: learning to

detect manipulated facial images. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV).

[20] Sabir E, Cheng J, Jaiswal A, Almageed W.A., Masi I, Natarajan P. (2019) Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. IEEE Conf. Comput. Vision Pattern Recogn. 3, 80–87.

[21] Torfi A, Iranmanesh S.M., Nasrabadi N., Dawson J. (2017) 3D Convolutional Neural Networks for Cross Audio-Visual Matching Recognition. IEEE Access 5, 22081–22091.

Citation of this Article:

Satwika.M, Pranavya.A, Neha.K, Rishika.K, & Siva Sankar Namani. (2024). Video Deepfake Detection Using EfficientNet. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 8(12), 145-150. Article DOI https://doi.org/10.47001/IRJIET/2024.812022

****