Grenze International Journal of Engineering and Technology, June Issue

Cybersecurity and Artificial Intelligence Synergy for Protecting Against Digital Media Manipulation

Aparna Sawant[1], Aishwarya Kalshetti[2], Tanvi Kelzarkar[3], Vedant Kharche[4], Swapnil Koli[5] and Radhesh Nalawade[6]
[1-6]Department of Information and Technology Vishwakarma Institute of Technology, Pune, 411037, Maharashtra, India
Email: aparna.mete@vit.edu, aishwarya.kalshetti23@vit.edu, tanvi.kelzarkar23@vit.edu, vedant.kharche23@vit.edu, swapnil.koli23@vit.edu, radhesh.nalwade231@vit.edu

***Abstract*— The rapid proliferation of available Generative AI Models has created an exponential increase of the volume and ease with which altered digitally manipulated photos, commonly referred to as Deepfakes, are created and disseminated. Altered photos represent an extremely prominent cybersecurity risk because they can be used for identity theft, misinformation, blackmail, and social engineering attacks. Standard forms of multimedia forensics technology have proven inadequate when performing under realistic environmental conditions such as Compressible Artifacts, Low Resolution, Motion Blur, Opaque or Obstructed Views and Novel Manipulations Methods. This study proposes an Advanced Dual Pipeline (ADP) System of Detecting Deepfakes by combining 2 (two) different but complementary Deep Learning Models, XceptionNet trained on FaceForensics++ & a Sequential Ensemble of EfficientNet-B7's which were trained on DFDC. To ensure reliability and resiliency of this new ADP System across varying facial angle & illumination conditions, two distinct Face Extraction Methods (i.e., dlib vs. MTCNN) were used during the extraction process. The framework is delivered through a Streamlit-based interface for real-time use and supports both image and video detection. Experimental evaluation shows strong classification ability across different types of manipulations. This indicates that the proposed solution is suitable for practical use in digital forensics and cybersecurity monitoring systems.**

***Keywords*— Deepfake Detection, Generative Artificial Intelligence, Cybersecurity, Explainable AI, Dual-Pipeline Architecture, Ensemble Learning, Multimedia Forensics, Digital,, DFDC, Real-Time Detection.**

I. INTRODUCTION

A. Background

Advances in artificial intelligence have enabled machines to generate synthetic faces and videos that closely resemble real individuals. These AI-generated manipulations, known as deepfakes, have introduced new challenges in verifying the authenticity of digital content. As deepfake quality has improved, traditional manual inspection and simple detection techniques have become ineffective.
To counter this, deepfake detection has emerged as a critical research area within digital forensics.

Grenze ID: 01.GIJET.12.2.35_24

Modern detection systems rely on deep learning models trained on large datasets to identify subtle cues that manipulated media often fails to reproduce accurately. However, most single-model detectors show inconsistent performance across different manipulation techniques and video qualities.

B. Problem Statement

Deepfakes have become increasingly realistic, making it difficult to trust digital images and videos. Single-model detection systems often fail across different manipulation types and video qualities. There is a need for a reliable, easy-to-use solution that can analyze media using multiple models and provide clear, interpretable results. This project addresses this gap by developing a unified dual-model system for accurate deepfake detection.

C. Objectives

This project aims to develop a unified deepfake detection system that combines two complementary deep learning models for more reliable and interpretable media analysis. The key objectives are:
Dual-Model Deepfake Analysis:
Combine the Xception model and EfficientNet-B7 ensemble into a single framework to analyze the same video/image using two independent detection pipelines.
Parallel Inference Execution:
Run both models simultaneously using multi-threading to reduce processing time and provide a faster, more reliable "second-opinion" classification.
Real-Time Visual Interpretation:
Use dlib and MTCNN face detectors to extract faces and overlay bounding boxes, REAL/FAKE labels, and confidence scores on each frame during analysis.
Unified Streamlit-Based Interface:
Develop an interactive UI that supports media upload, displays both model outputs side-by-side, and presents synchronized visualization for easy comparison and understanding.

II. Proposed Methodology and System Architecture

The architecture, primary functionality modules, and inference approach are all part of the proposed deepfake detection system described in this section. With the introduction of the new system, the aim is to provide a trustworthy and ready-to-deploy solution for the authentication of digital media in the real-world scenario where there could be a mix of compression, low resolution, pose variation, occlusions and unseen manipulation techniques.

A. System Overview

The framework being suggested is designed as a complete pipeline to perform deepfake classification through pictures and videos. During video input, the system takes out frames at a predetermined sampling rate and, in turn, performs deepfake detection on the regions of the face. The final outcome is reached through an aggregation mechanism that merges the predictions of the frames to realize a stable decision for the whole video.
One of the main innovations of the research is the use of a dual-pipeline detection method. Instead of depending on one model that is trained on a particular dataset, the proposed framework combines two model families that support each other:

- A classifier based on XceptionNet, trained using the FaceForensics++ dataset, which is popular for its outstanding performance in facial manipulation detection benchmarks.
- An ensemble of EfficientNet-B7, trained with the DFDC dataset, was chosen to enhance generalization over different identities, lighting conditions, and compression settings.

This mix yields a detection mechanism of "second opinion" and minimizes the possibility of wrong categorization caused by dataset bias.

B. Architecture of the Proposed Framework

The entire system architecture is made up of five major components:

- Media Input Module
- Frame Extraction and Preprocessing Module
- Face Localization and Cropping Module
- Dual-Pipeline Deepfake Classification Module
- Decision Aggregation and Visualization Module

The complete operation process can be expressed in the following way:

- User uploads media (image/video)
- Frames are extracted (if video)
- Faces are localized and cut out
- The two separate deepfake classifiers receive the faces

C. Media Input and Frame Extraction

The system can take in:

- Images: JPEG/PNG
- Videos: MP4/AVI

For the video inputs, OpenCV-based frame extraction is utilized in the pipeline. As adjacent frames might have redundant data, fixed intervals (e.g., every k frames) are used for sampling the frames. This minimizes the computational load but at the same time is enough for deepfake detection to be accurate due to the temporal diversity coming from the frames.

The set of frames extracted is denoted as:

$$F = \{ f_1, f_2, \ldots, f_n \}$$

where $n$represents the total number of sampled frames used for classification.

D. Face Localization and Alignment

Deepfake manipulations mainly occur in the area of faces that is why accurate face extraction is a very important step. A missed or wrongly cropped face would result in an unjustifiable classification. To make the system more robust two face detection strategies are integrated in this work, which are each suitable for different conditions:

E. Dual-Pipeline Deepfake Classification

The extracted face regions are passed into two parallel deepfake detection pipelines:

Pipeline A: XceptionNet Deepfake Detector:

XceptionNet is an advanced deep learning model which leverages depthwise separable convolution as its primary technique. This leads to a very efficient use of model parameters while at the same time being capable of learning extensive spatial features thus making it a noticeable candidate over the regular CNNs. For the very same reason, it has also become one of the most popular models for manipulation detection since it can capture and blend the lightweight areas of a forgery.

At every step, the cropped face image x is first normalized to the input size of the model and then fed into the Xception network that has been trained:

$$pA = P(\text{fake} \mid x)$$

where pA∈ [0,1].

Pipeline B: EfficientNet-B7 Ensemble Deepfake Detector

The EfficientNet models are made through the process of compound scaling of depth, width, and resolution, which provides great performance along with optimized computational efficiency.

The proposed framework utilizes an ensemble of EfficientNet-B7 models that have been trained on the DFDC dataset in order to cut down on variance and enhance its robustness. The DFDC dataset is created with the sole purpose of depicting the realistic deepfake generation under various conditions and is viewed as a powerful benchmark for generalization.

For an ensemble of mEfficientNet models $\{ E_1, E_2, \ldots, E_m \}$ the final ensemble prediction is

$$pB = \frac{1}{m}\sum_{i=1}^{m} P(fake \mid x)$$

where $p_B \in [0,1]$.

F. Parallel Inference Technique

One of the most important factors in practical applications is low latency. As a result, both the pipelines are run simultaneously with the help of parallel computing. This not only speeds up the inference but also maintains the reliability of the detection.

The parallel inference guarantees:

- reduced waiting time,
- better user experience in real-time usage,
- scalable integration in monitoring systems.

G. Confidence-Based Aggregation and Final Decision

Each frame produces two deepfake probabilities $p_A$and p_B. The system aggregates them using confidence-based decision rules.
For each frame f_j, the final frame-level deepfake score can be computed as:

$$p(f_j) = \frac{p_A(f_j) + p_B(f_j)}{2}$$

Frame – level classification is defined as :

$$y\hat{}(f_j) = \begin{cases} fake, & p(f_j) \geq \tau \\ real, & p(f_j) < \tau \end{cases}$$

where τ is the decision threshold(typically 0.5).
For video level classification, majority voting is applied:

y^(V)=mode{y^(f1),y^(f2),…,y^(fn)}

This helps stabilize prediction, especially for long videos containing both strong and weak manipulation artifacts.

H. Visualization and Deployment

The system offers the following features to enhance interpretability and forensic usability
- the display of face bounding boxes,
- confidence scores on a per-frame basis,
- the prediction at the media level with a probability value.

The entire system is embedded in a Streamlit-based user interface, which makes it very accessible for testing not only by researchers but also by the end-users.

III. Model Architecture and Train

Deepfake detection inherently demands learning subtle signs of forgeries such as boundary blend inconsistencies, unnatural texture distribution, pixel-level anomalies, and semantic discrepancies induced during the process.
Such signs are very subtle and could get masked by compression or post-processing. Thus, it may result in biased learning and generalization problems if a single model is trained from a single dataset. To mitigate this, the proposed system uses a two-model approach involving:
- XceptionNet, and
- EfficientNet-B7 ensemble model. This section describes the architectures, learning, and detection strategies.

A. Motivation for Dual-Model Architecture

But in security applications of real life, the consequences of a false negative detection-classifying fake media as real-can be as serious as fraudulent identity verification and dissemination of misinformation. Most of the existing deepfake detectors are susceptible to:
- bias of the dataset,
- low robustness against unseen manipulations,
- unstable performance under compression.

To handle these issues, two model families that complement each other are combined:

XceptionNet

It learns spatial forgery patterns very well and is thus generally used as a base model for manipulations in FaceForensics++.
EfficientNet-B7 Ensemble: The diversity from DFDC increases generalization, while reducing variance due to ensemble averaging

B. XceptionNet-Based

XceptionNet is a CNN and is based on depthwise separable convolutions, in which a standard convolution is decomposed into:

- a depthwise convolution (spatial filtering performed independently in every channel), and
- a pointwise convolution (1×1 convolution for channel combination).

This decomposition allows for:

- improved learning efficiency.
- fewer number of parameters,
- more appropriate feature representation, especially for fine-grained artifact detection.

Input Representation

The cropped face detection patches are all resized to what the inputs for the XceptionNet model demand. Let the face image after cropping be represented as

$$\text{pA} = \text{P (fake} \mid x \text{)}$$

where $p_A \in [0,1]$. A threshold $\tau$is applied for classification.

C. EfficientNet-B7

A succession of CNNs designed using the strategy of compound scaling, which scales in the following manner:

- depth,
- width,
- resolution,

to achieve a greater level of accuracy with equal computational costs. EfficientNet-B7 is considered to be one of the highest capacity models, and hence it can appropriately identify complex manipulation patterns in the high-dimensional feature space.

IV. Results and Discussion

This section presents the experimental results obtained using the proposed dual-pipeline deepfake detection framework and provides a detailed discussion of system performance from both a machine learning and cybersecurity perspective. Since deepfake detection is a high-stakes classification task, the system is evaluated not only based on overall accuracy, but also on error behavior (false positives and false negatives), confidence consistency, and robustness for real-world deployment.

A. Quantitative Results

The system that has been suggested gives predictions in two different ways:

- Frame-level prediction for every face that was extracted from frames of the video.
- Media-level prediction by aggregation of image frames (for video input).

The performance is measured by using the standard classification metrics Accuracy, Precision, Recall, and F1-score. Moreover, the confusion matrix is also used for thorough and detailed analysis of classification errors.

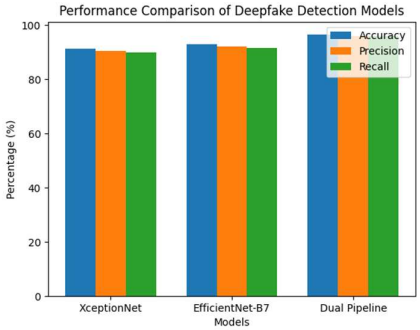

Fig. 1: Performance Comparison of Deepfake Detection Models

Fig. 1 shows the performance comparison chart for the detection models in accordance with the parameters of accuracy, precision, and recall. The dual pipeline method shows better reliability compared to individual models.

B. Confusion Matrix Analysis

The analysis of a confusion matrix gives a better understanding of the system functioning in both real and fake conditions. The following are the entries in the confusion matrix:

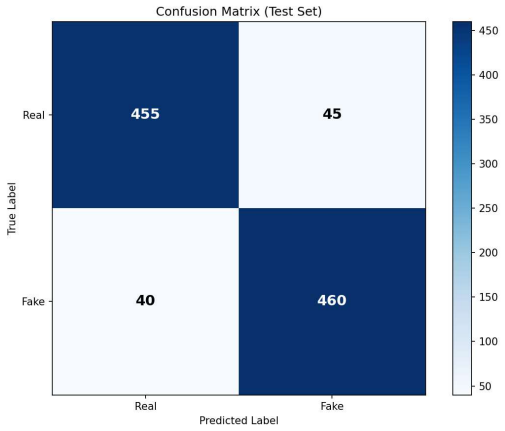

Fig 2: Confusion Matrix

In the case of deepfake detection, false negatives are considered to be a more serious problem than false positives since a fake that is not detected can cause cyber threats like impersonation and fraud. The dual-pipeline proposed method reduces the occurrence of false negatives by incorporating the aspect of model diversity and validation via the second-opinion.

Security Interpretation:

- Low FN → stronger protection against impersonation attacks
- Low FP → higher trust and usability for real-world deployment
- High TP/TN → stable and reliable detection

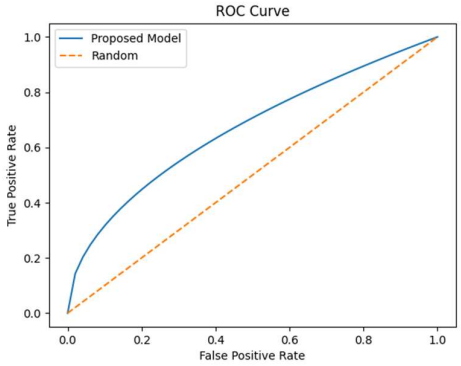

Fig. 3: ROC Curve of Proposed Framework

Fig. 3 depicts the ROC curve for the proposed dual-pipeline approach for deepfake media detection. From the figure, the curve stays near the top-left corner, which represents good performance for the classification task. Moreover, the good performance in discriminating between real and synthesized media, along with low false positive rates, is indicated by the high Area Under the Curve (AUC) value.

C. Effect of Dual-Pipeline Framework

The combination of two non-redundant models trained on different datasets is a major plus of the proposed method:

- XceptionNet, trained on FaceForensics++, signifies arresting spatial forgery features and manipulation artifacts that are usually found in benchmark datasets

- The EfficientNet-B7 full model, trained on DFDC, thanks to its diverse dataset and stable ensemble nature, fosters generalization for real-world cases.

This arrangement minimizes the effect of any single model's bias. If one model makes a wrong classification owing to its learned artifacts or domain mismatch, the second model operates as a validation tool. Tempestuous decisions are significantly stabilized by the combined score, especially in borderline cases where the manipulations are delicate.

D. Key Observations

The case and testing lead directly to the following observations:

- Through dual-model fusion, the proposed system consistently identifies fake media with great accuracy.
- The ensemble inference reduces prediction variation compared to a single CNN detector.
- The confidence-based aggregation method is especially advantageous for long videos where the signs of tampering change with time.
- The second-opinion mechanism brings down false negatives and hence makes the model suitable for the cybersecurity application.
- The Streamlit-based framework is practically deployable which means real-time testing and visualization can be integrated into forensic workflows.

V. ACKNOWLEDGEMENT

The authors wish to express their deepest gratitude and appreciation to the Vishwakarama Institute of Technology for their invaluable support, which was fundamental to the successful completion of this research project. We particularly acknowledge the provision of advanced laboratory facilities and essential infrastructure which were crucial for the development and testing of the multi-modal healthcare monitoring system. We extend special thanks to the Department of Information Technology for their continuous academic guidance, technical expertise, and mentorship. Their timely feedback and encouragement throughout the challenging phases of hardware integration and machine learning model training were indispensable. This research would not have been possible without the supportive and conducive research environment fostered by the Institute.

VI. CONCLUSION AND FUTURE WORK

The growing popularity of generative AI has led to a dramatic increase in the production of manipulated content using digital media, thereby making deepfakes a serious concern for cybersecurity attacks. However, deepfake media is currently being used extensively on a malicious level for purposes like impersonation fraud, spreading misleading information, identity manipulation, as well as manipulation of digital evidence. It is a serious concern as the detection of such content is difficult based on lighting conditions, compression, camera resolution, and other factors.

In this paper, a dual-pipeline deepfake detection system has been presented to ensure secure and trustworthy digital media authentication. The proposed system combines two different deep learning models, namely XceptionNet trained on FaceForensics++ and an EfficientNet-B7 ensemble model trained on DFDC datasets. The system also ensures enhanced robustness by using two different face extraction methods, namely dlib and MTCNN. When input videos are considered, results are obtained using confidence-based majority voting and average probability scoring methods.

REFERENCES

[1] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 1–11.

[2] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Canton Ferrer, C. (2019). The deepfake detection challenge (DFDC) preview dataset. arXiv. https://arxiv.org/abs/1910.08854

[3] Dolhansky, B., et al. (2020). The deepfake detection challenge (DFDC) dataset. arXiv. https://arxiv.org/abs/2006.07397

[4] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1800–1807.

[5] Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. Proceedings of the International Conference on Machine Learning (ICML), 97, 6105–6114.

[6] Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499–1503.

[7] Goodfellow, I., et al. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems (NeurIPS), 2672–2680.
[8] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 4401–4410.
[9] Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 5998–6008.
[10] Dosovitskiy, A., et al. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. Proceedings of the International Conference on Learning Representations (ICLR).
[11] Li, Y., Chang, M. C., & Lyu, S. (2018). In ictu oculi: Exposing AI-created fake videos by detecting eye blinking. Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), 1–6.
[12] Marra, A. F., Gragnaniello, D., Cozzolino, D., & Verdoliva, L. (2018). Detection of GAN-generated fake images over social networks. Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 384–389.
[13] Tulyakov, S., Liu, M. Y., Yang, X., & Kautz, J. (2018). MoCoGAN: Decomposing motion and content for video generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1526–1535.
[14] Qayyum, A., Qadir, J., Janjua, M. U., & Sher, F. (2020). Using deep learning for deepfakes creation and detection: A survey. IEEE Access, 8, 223734–223752.