International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 03 | March - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Hybrid Deep Learning and Machine Learning Approach for Detecting DeepFake Content

Bandreddi Sai Kumar[1], P.V.V.Vamsi Krishna Murthy[2], G.Satya Bhushan Manikanta [3], D.Phenehas Obededom[4]

[1,2,3,4] *Dept. of CSE-AIML, Sasi Institute of Technology and Engineering, Andhra Pradesh*, India;
Email: saikumar.bandreddi@sasi.ac.in, krishnamurthy.pratti@sasi.ac.in, Bhushan.guttula@sasi.ac.in, obededom.dovari@sasi.ac.in

Abstract

The rapid growth of artificial intelligence and deep learning technologies has led to the widespread creation of deepfake images. These manipulated images pose serious threats to digital authenticity, privacy, and the reliability of online information. Manual identification of such manipulated content has become increasingly difficult due to improvements. This paper proposes a hybrid deep learning approach for detecting deepfake images using DenseNet121, ResNet18, EfficientNet-B4 and a custom Random CNN model. The system is trained on approximately 30,000 images from two datasets. Each model extracts visual features independently and their predictions are combined using an ensemble learning strategy. Experimental results show that the proposed hybrid model achieves an accuracy of approximately 96–98%, demonstrating strong performance in identifying manipulated images. The system is implemented using Python and Flask with a web interface that allows users to upload images and obtain prediction results.

Keywords: Deepfake Detection, Hybrid Deep Learning Models, CNN, Image Manipulation Detection, Ensemble Learning, EfficientNet, DenseNet, ResNet.

1. Introduction

Deepfake technology uses advanced deep learning techniques to generate highly realistic fake images and videos. These manipulated media can be used for misinformation, identity theft, and digital fraud. As deepfake generation methods become more sophisticated, detecting manipulated media has become an important research challenge. Deep learning models such as CNNs have shown strong performance in image classification tasks. However, single-model approaches often struggle to detect complex manipulation artifacts. Therefore, hybrid approaches that combines multiple architectures. In this work, a hybrid framework combining DenseNet121, ResNet18, EfficientNet-B4, and a custom Random CNN model is proposed. The ensemble prediction mechanism improves detection reliability and robustness across multiple datasets.

2. Literature Survey

Deepfake detection has become an important research area due to the rapid growth of artificial intelligence–based image and video manipulation technologies. Deep learning methods, particularly convolutional neural networks (CNNs), have been widely used for identifying manipulated media because they can automatically learn complex visual features from images. Several studies have proposed different deep learning architectures and hybrid models to improve the detection of deepfake images.

Jadhav et al. (2024) proposed a deepfake detection framework using a Video Vision Transformer architecture, which analyzes spatial and temporal features of manipulated media. Their model demonstrated promising performance in detecting deepfake videos by leveraging attention-based learning mechanisms. However, transformer-based models generally require higher computational resources compared to traditional CNN architectures.[3]

Neha and Arora (2023) developed a deep learning–based system for detecting deepfake images using convolutional neural networks. Their approach focused on extracting facial features and identifying inconsistencies introduced during the deepfake generation process. The study showed that CNN-based models can achieve reliable performance when trained on balanced datasets containing real and fake images. However, the model performance may decrease when tested on previously unseen datasets.[7]

Sunil et al. (2024) investigated deepfake detection using data augmentation and layer unfreezing techniques across multiple deep learning models. Their method improved model generalization by exposing the network to various manipulated image patterns during training. The experimental results showed that data augmentation helps improve detection accuracy and model robustness.[14]

Patel et al. (2023) introduced an improved dense CNN architecture designed to enhance deepfake detection performance. Their model utilized dense connectivity to strengthen feature propagation between layers and reduce the vanishing gradient problem. Experimental evaluation demonstrated improved classification performance compared with conventional CNN architectures.[9]

Several researchers have also conducted comparative studies of deep learning models for deepfake detection. Khatri et al. (2023) compared multiple deep learning approaches and found that convolutional neural networks outperform many traditional machine learning techniques in identifying manipulated images. Their study highlighted the importance of selecting appropriate architectures and datasets for achieving reliable detection results.[5]

EfficientNet-based models have also gained attention for deepfake detection due to their ability to balance network depth, width, and resolution while maintaining computational efficiency. Research studies have shown that EfficientNet architectures can achieve high accuracy in detecting fake images

Table 1: Comparison of Existing Models

Authors & Year	Model Architecture	Dataset Used	Performance	Result	Limitations
Jadhav et al., 2024[3]	Video Vision Transformer	Deepfake video datasets	High detection accuracy	Effective deepfake detection	High computational cost
Neha & Arora, 2023[7]	CNN-based model	Deepfake image datasets	Improved classification accuracy	Reliable detection	Limited generalizatio n
Sunil et al., 2024[14]	Multiple CNN models	Augmented image dataset	Better performance using augmentatio n	Improved model robustness	Requires large training data
Patel et al., 2023[9]	Dense CNN	Deepfake image dataset	High precision detection	Improved detection accuracy	Model complexity
Khatri et al., 2023[5]	Deep learning comparative models	Deepfake datasets	Comparativ e evaluation	Identified best performing models	Limited dataset diversity
Pan et al., 2020[8]	CNN architecture	Deepfake datasets	Accurate classification	Effective feature extraction	Sensitive to dataset bias
Zhalgasbaye v et al., 2024[15]	EfficientNet-B4	Deepfake image datasets	High accuracy	Efficient detection	Computatio nal resources required
Singh et al., 2024[12]	EfficientNet model	Deepfake images	Strong classification performance	Robust detection	Limited training samples

Das et al., 2025[1]	CNN + EfficientNet ensemble	Facial manipulatio n datasets	Improved ensemble accuracy	Robust detection	Complex model training

3. Methodology of Proposed System

The proposed deepfake detection system follows a deep learning–based approach for identifying manipulated facial images. The primary objective of the methodology is to automatically classify input images as real or deepfake by learning discriminative visual features using multiple convolutional neural network architectures.

Initially, two publicly available datasets were collected for training and evaluation. The first dataset is the DeepFake-vs-Real-20K dataset, and the second dataset is the Deepfake and Real Images dataset obtained from Kaggle. These datasets contain both genuine and manipulated facial images. After combining the datasets, approximately 30,000 images were used for model training and evaluation. The dataset was divided into 80% training data and 20% testing data to ensure proper model learning and unbiased evaluation. Before feeding the images into the deep learning models, a preprocessing stage was performed. This preprocessing stage includes image resizing, normalization, and tensor conversion, which ensures that the images have a uniform format suitable for deep learning models. The proposed system is designed to automatically detect deepfake images using a hybrid deep learning framework that combines multiple convolutional neural network models. The system architecture consists of several modules that work together to process input images and generate prediction results. The uploaded image is then processed by the image validation and preprocessing module. In this stage, the system verifies the image format and resizes the image to the required dimensions. The image is then normalized and converted into tensor format to make it compatible with deep learning models. After preprocessing, the image is forwarded to the feature extraction and classification module. In this module, multiple deep learning models including DenseNet121, ResNet18, EfficientNet-B4, and Random CNN analyze the image independently. Each model extracts deep visual features such as edges, facial textures, and structural inconsistencies that may indicate deepfake manipulation.

Each model produces a prediction score representing the probability that the image belongs to either the real or deepfake class. These individual predictions are then passed to the ensemble prediction module. The ensemble module combines the outputs from all models using probability averaging or voting mechanisms to generate the final classification result. Finally, the result visualization module displays the prediction result to the user through the web interface. The system indicates whether the uploaded image is real or deepfake, along with a confidence score representing the certainty of the prediction. The proposed system achieves an overall detection accuracy of approximately 96–98%, demonstrating the effectiveness of hybrid deep learning models for deepfake image detection.

4. Analysis of Datasets

The performance of the proposed deepfake detection system was evaluated using two publicly available datasets collected from Kaggle: the DeepFake-vs-Real-20K dataset and the Deepfake and Real Images dataset. These datasets contain facial images belonging to two classes: real images and deepfake images. The datasets were selected because they provide a balanced collection of authentic and manipulated images that help the deep learning models learn visual patterns associated with deepfake generation. Compared to many previous studies, which often rely on a single dataset, the use of multiple datasets in the proposed system improves data diversity and reduces overfitting.

After combining both datasets, approximately 30,000 images were used in the proposed system. The dataset was divided into 80% training data and 20% testing data.

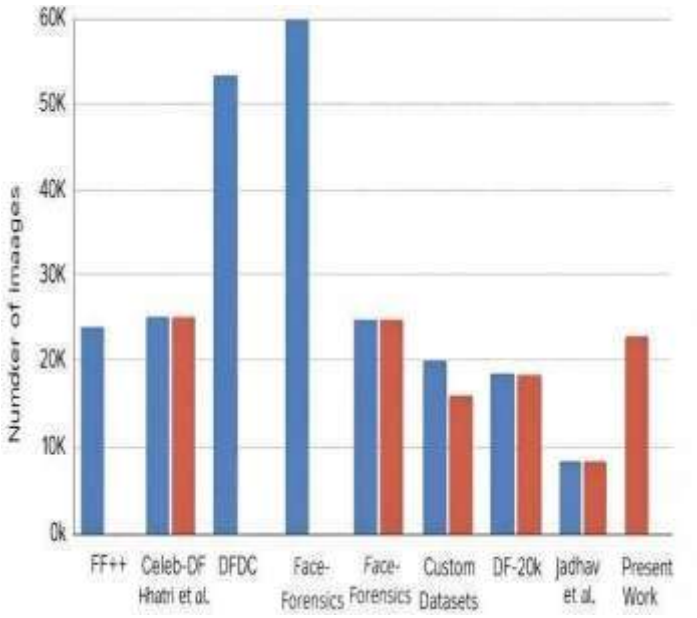

Fig 1.0: Comparison of Datasets Used in Deepfake Detection Studies.

5.System architecture

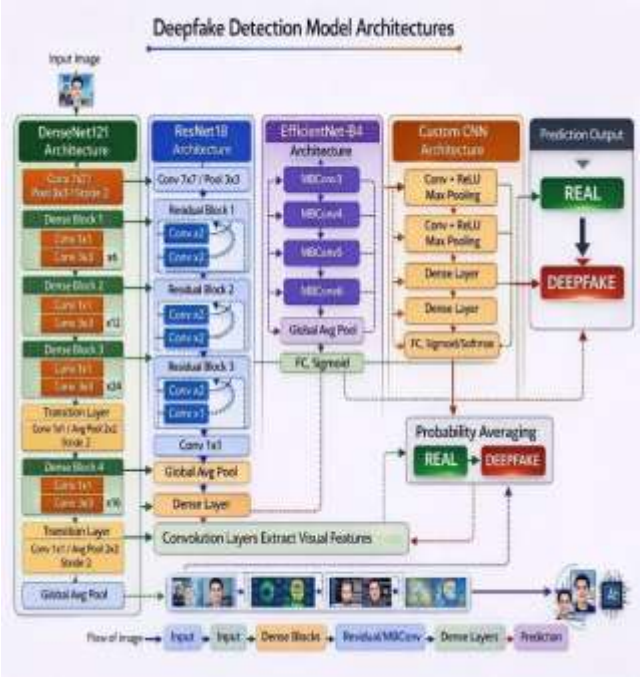

Fig. 1.1: Deepfake Detection Model Architecture using Hybrid Deep Learning Models

The figure illustrates the hybrid deep learning architecture used for deepfake detection. The input image is processed by multiple CNN models including DenseNet121, ResNet18, EfficientNet-B4, and a custom Random CNN. Each model extracts visual features and produces prediction probabilities. The final decision is generated using probability averaging in the ensemble module to classify the image as real or deepfake.

The DenseNet121 model uses dense connections between layers, allowing each layer to receive information from all previous layers. This helps the network reuse features and improve information flow during training. The ResNet18 model uses residual connections or skip connections, which help the network learn deeper representations without suffering from the vanishing gradient problem. This allows the model to capture complex image patterns effectively.

The EfficientNet-B4 model uses compound scaling techniques to balance network depth, width, and resolution. This model is designed to achieve high accuracy while maintaining computational efficiency.The Random CNN model is a custom convolutional neural network designed to extract additional visual features from the input images. It contains convolution layers, activation functions, pooling layers, and dense layers for classification.Each model generates a prediction probability indicating whether the image is real or deepfake. These outputs are then passed to an ensemble module, where the predictions are combined using probability averaging.The ensemble approach improves the overall system performance because each model learns different visual characteristics of deepfake images. By combining their predictions, the system produces a more reliable final decision.

Finally, the system outputs the classification result as REAL or DEEPFAKE, along with a confidence score.

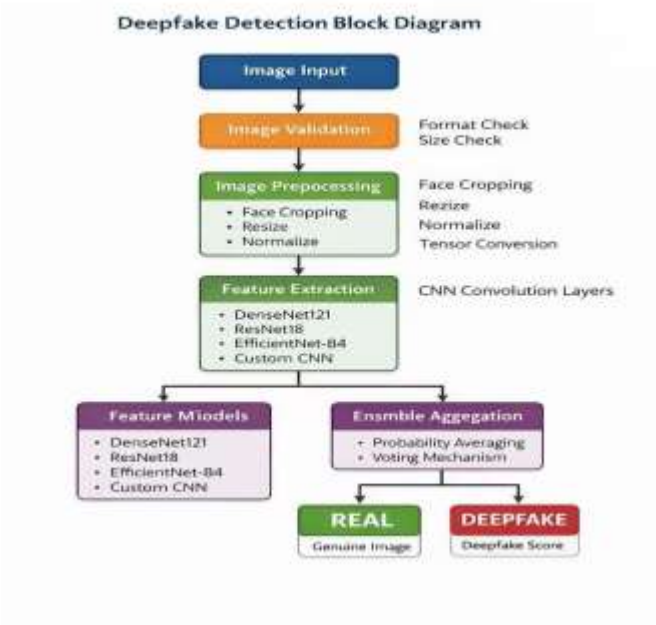

Fig 1.2: Block Diagram of the Proposed Deepfake Detection System

The block diagram shows the overall workflow of the deepfake detection system. The process begins with image input followed by image validation and preprocessing steps such as resizing, normalization, and tensor conversion. The processed image is then analyzed by multiple deep learning models for feature extraction, and their outputs are combined using ensemble aggregation to produce the final prediction.

6. Results and Findings

The performance of the proposed deepfake detection system was evaluated using multiple deep learning models, including DenseNet121, ResNet18, EfficientNet-B4, and Random CNN. Each model independently predicts the probability of an image belonging to either the real or deepfake class. To improve prediction accuracy and reliability, an ensemble learning strategy was implemented.

The ensemble module combines predictions from all models using two methods: probability averaging (soft voting) and majority voting (hard voting). These methods help reduce individual model bias and improve overall classification performance.

6.1 Probability Averaging (Soft Voting)

In the probability averaging method, each model outputs the probability of the image belonging to the **real** or **fake** class. The final prediction is obtained by averaging the predicted probabilities from all models. If there are **N models**, the averaged probability is calculated as:

$$P_{fake} = \frac{1}{N}\sum_{i=1}^{N} P^{(i)}_{fake} \quad \rightarrow \text{Formula 1}$$

$$P_{real} = \frac{1}{N}\sum_{i=1}^{N} P^{(i)}_{real} \quad \rightarrow \text{Formula 2}$$

Where:

- $P^{(i)}$ = probability predicted by model i for fake class
- $P^{(i)}$ = probability predicted by model i for real class
- N= total number of models

For example, if the predicted fake probabilities from four models are: DenseNet121 = 0.70, ResNet18 = 0.60 EfficientNet-B4 = 0.80, RandomCNN = 0.55

The final averaged probability becomes:

$$p_{fake} = (0.70+0.60+0.80+0.55)/2 = 0.6625$$

Similarly,

$$Preal = 1 - Pfake = 0.3375$$

Since the fake probability is higher, the **final prediction is Deepfake**.

6.2 Voting Mechanism (Hard Voting)

Another ensemble strategy used in this work is **majority voting**. In this method, each model directly predicts a class label instead of probabilities. The final class is determined based on the class that receives the highest number of votes.

$$FinalClass = \arg\max \sum_{i=1}^{N} Vote \quad \square \text{ Formula 3}$$

Where:

- $Vote_i$represents the predicted class from model i
- $N$is the total number of models For example:

DenseNet121 → Fake ResNet18 → Fake EfficientNet-B4 → Fake RandomCNN → Real

Number of votes:

Fake = 3

Real = 1

Since the **majority vote is Fake**, the final classification result becomes **Deepfake**

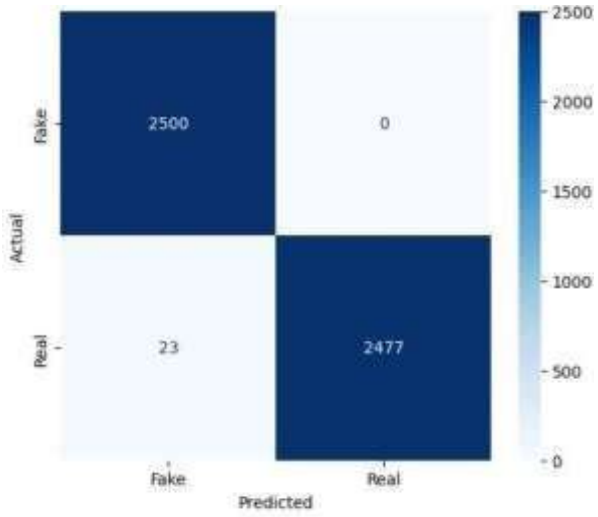

Fig. 1.3: Confusion Matrix of DenseNet121 Model

The confusion matrix illustrates the classification performance of the DenseNet121 model on the testing dataset. The matrix shows the number of correctly classified real and deepfake images along with misclassified samples, demonstrating the effectiveness of the model in detecting manipulated images.

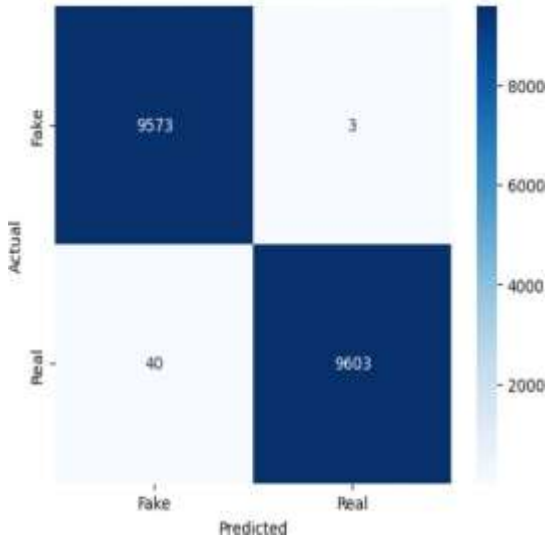

Fig. 1.4: Confusion Matrix of EfficientNet-B4 Model

This confusion matrix represents the performance of the EfficientNet-B4 model on the testing dataset. The model shows strong classification capability in distinguishing real and deepfake images with minimal misclassification.

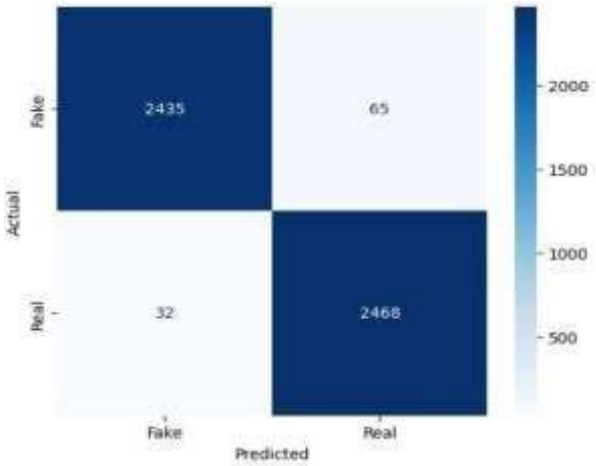

Fig. 1.5: Confusion Matrix of ResNet18 Model

The confusion matrix shows the classification results produced by the ResNet18 model. The results demonstrate that the model effectively identifies deepfake images by learning complex visual features and facial manipulation artifacts.

7. Future Enhancement and Suggestions

Recent research in deepfake detection highlights several directions for improving the effectiveness and reliability of detection systems. One important enhancement suggested in the literature is the use of larger and more diverse datasets. Many existing studies rely on limited datasets, which may not represent all possible deepfake generation techniques. Expanding datasets with images generated using different manipulation methods can help improve the generalization capability of detection models.

Another important improvement proposed in the literature is the integration of advanced deep learning architectures such as Vision Transformers (ViT) and hybrid CNN–Transformer models. These architectures are capable of capturing both spatial and contextual features in images, which can help detect subtle artifacts introduced during deepfake generation. Combining traditional convolutional neural networks with transformer- based models may significantly enhance detection accuracy.

Researchers also suggest improving model robustness and generalization by applying advanced training strategies such as data augmentation, transfer learning, and domain adaptation. These techniques allow models to learn more diverse image patterns and perform

better when tested on unseen datasets or newly generated deepfake content.Additionally, future research may focus on multimodal deepfake detection, where both visual and audio information are analyzed simultaneously. By combining multiple sources of information, such systems can improve detection accuracy and better identify manipulated media.

8. Conclusion

The rapid advancement of artificial intelligence and deep learning technologies has significantly increased the creation and distribution of deepfake images, posing serious challenges to digital media authenticity and information reliability. Detecting such manipulated content has therefore become an important research problem in computer vision and cybersecurity.

In this work, a hybrid deep learning–based deepfake detection system was developed to accurately identify manipulated images. The proposed system integrates multiple convolutional neural network architectures including DenseNet121, ResNet18, EfficientNet-B4, and a custom Random CNN model to extract deep visual features from facial images. By combining the strengths of different models, the system is capable of capturing various manipulation artifacts that may not be detected by a single model.

The system utilizes an ensemble learning approach where predictions from individual models are combined using probability averaging and voting mechanisms to generate the final classification result. The models were trained using approximately 30,000 images collected from multiple datasets, enabling the system to learn diverse image patterns and improve detection reliability.

Experimental results demonstrate that the proposed hybrid framework achieves an overall accuracy of approximately 96–98% in distinguishing real and deepfake images. The confusion matrix analysis further confirms the effectiveness of the system in minimizing misclassification between genuine and manipulated images.

In addition, the integration of a Flask-based web interface allows users to upload images and obtain real-time detection results, making the system practical for real-world applications. Overall, the proposed approach contributes to improving the reliability of digital media by providing an automated, accurate, and scalable solution for deepfake detection.

9. References

[1] Das, S., et al. (2025). Enhanced deepfake detection using CNN and EfficientNet-based ensemble models for robust facial manipulation analysis. In *Proceedings of CE2CT 2025*. https://doi.org/10.1109/CE2CT64011.2025.10939946

[2] Dheeraj, J. C., Nandakumar, K., Aditya, A. V., Chethan, B. S., & Kartheek, G. C. R. (2021). Detecting deepfakes using deep learning. In *Proceedings of RTEICT 2021* (pp. 651–654). https://doi.org/10.1109/RTEICT52294.2021.9573740

[3] Jadhav, S., Narale, D., Kore, R., Shisode, U., & Kulange, A. (2024). Unmasking the illusion: A novel approach for detecting deep fakes using video vision transformer architecture. In *Proceedings of ACOIT 2024*. https://doi.org/10.1109/ACOIT62457.2024.10939366

[4] Jolly, V., Telrandhe, M., Kasat, A., Shitole, A., & Gawande, K. (2022). CNN-based deep learning model for deepfake detection. In *Proceedings of ASIANCON 2022*. https://doi.org/10.1109/ASIANCON55314.2022.9908862

[5] Khatri, N., Borar, V., & Garg, R. (2023). A comparative study: Deepfake detection using deep learning. In *Proceedings of Confluence 2023*. https://doi.org/10.1109/Confluence56041.2023.10048888

[6] Mishra, A., Bharwaj, A., Yadav, A. K., Batra, K., & Mishra, N. (2024). Deepfakes—Generating synthetic images and detecting artificially generated fake visuals using deep learning. In *Proceedings of Confluence 2024*. https://doi.org/10.1109/Confluence60223.2024.10463337

[7] Neha, & Arora, B. (2023). Deep learning-based model for deepfake image detection: An analytical approach. In *Proceedings of ICIMIA 2023* (pp. 1019–1027). https://doi.org/10.1109/ICIMIA60377.2023.10426561

[8] Pan, D., Sun, L., Wang, R., Zhang, X., & Sinnott, R. O. (2020). Deepfake detection through deep learning. In *Proceedings of BDCAT 2020* (pp. 134–143). https://doi.org/10.1109/BDCAT50828.2020.00001

[9] Patel, Y., et al. (2023). An improved dense CNN architecture for deepfake image detection. *IEEE Access, 11*, 22081–22095. https://doi.org/10.1109/ACCESS.2023.3251417

[10] Petmezas, G., et al. (2024). Video deepfake detection using a hybrid CNN-LSTM-transformer model for identity verification. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042- 024-20548-6

[11] Sadiq, S., et al. (2023). Deepfake detection on social media: Leveraging deep learning and FastText embeddings. *IEEE Access*. 10.1109/ACCESS.2023.3308515

[12] Singh, S., Sarala, P., Chandra, S. K., & Kumar, M. D. (2024). Deepfake image detection using EfficientNet model. In *Proceedings of OTCON 2024*. https://doi.org/10.1109/OTCON60325.2024.10687774

[13] Siva Rama Lingham, N., Devakanth, J. J. M. A., Raj, G., Gayathri, K., Janani, R., & Dhanapal, R. (2024). Development of deepfake detection techniques for protecting multimedia information. In *Proceedings of ICAAIC 2024*. https://doi.org/10.1109/ICAAIC60222.2024.10575155

[14] Sunil, R., Mer, P., Diwan, A., & Parmar, P. (2024). Deepfake detection in digital images using data augmentation and layer unfreezing across various deep learning models. In *Proceedings of TENCON 2024*. https://doi.org/10.1109/TENCON61640.2024.11049067

[15] Zhalgasbayev, A., Aiteni, T., & Khaimuldin, N. (2024). Using EfficientNet-B4 for deepfake image detection. In *Proceedings of IEEE AITU 2024*. https://doi.org/10.1109/IEEECONF61558.2024.10585385