IEEE 2nd International Conference on Computing, Applications and Systems (COMPAS 2025)
23–24 October 2025, Kushtia, Bangladesh

Enhancing DeepFake Classification Performance Using a CNN and XceptionNet-Based Pipeline

Nishat Tasnim Susmi
Dept. of ICE, PUST
Pabna, Bangladesh
nishat.210605@s.pust.ac.bd

Manob Chandra Chanda
Dept. of ICE, PUST
Pabna, Bangladesh
manob.210624@s.pust.ac.bd

Md. Sarwar Hosain
Dept. of ICE, PUST
Pabna, Bangladesh
sarwar.ice@pust.ac.bd

Md. Rifat Hossen
Dept. of ICE, PUST
Pabna, Bangladesh
rifat.220614@s.pust.ac.bd

Md. Anwar Hossain
Dept. of ICE, PUST
Pabna, Bangladesh
manwar.ice@pust.ac.bd

Abul Fazal Mohammad Zainul Abadin
Dept. of ICE, PUST
Pabna, Bangladesh
abadin.7@pust.ac.bd

***Abstract*—Deepfake technology can create a video or image that looks real but in reality it is fake.It is a real threat to our society and our digital security. The popularity of generative models like generative adversarial networks (GANs) has made it easier to produce content that we can not differentiate between fake media and authentic media. To combat this, the study introduces a dual-model deepfake detection system that combines a custom lightweight convolutional neural network (CNN) with a transfer learning-based XceptionNet. This framework is trained and tested on the 140K Real and Fake Faces dataset, which contains an similar number of real and synthetic images. The custom CNN is built from scratch, featuring optimized convolutional layers, ReLU activations, max-pooling, and dense layers with dropout for regularization. The XceptionNet model is fine-tuned with additional dense layers for binary classification. Both models follow the same pre-processing steps and are trained with the Adam optimizer using binary cross-entropy loss. The CNN achieves an impressive 97% accuracy, while XceptionNet reaches 91%, highlighting their strong performance and ability to generalize. Metrics like precision, recall and F1-score verify the reliability of both methods. The research suggest that despite limited data and resources, deepfake detection is feasible by utilizing efficient architectures and training techniques. This framework gives a scalable and resource efficient solution for real world deepfake forensics, particularly in environments with limited computational capacity. Future work will investigate ensemble models and real-time detection implementations.**

***Index Terms*—deepfake detection, CNN, XceptionNet, transfer learning, image forgery**

I. Introduction

In recent years, media featuring unprecedented realism known as deepfakes have emerged due to advancements in deep learning and generative models. These technologies manipulate or produce images and audio with neural networks called generative adversarial networks (GANs). Consequently, it has become easy to produce fake images and videos that humans can not differentiate between fake and real images or videos. The deepfake technology has led to increased misuse across various sectors, especially on social media, where false information and fabricated events are circulated, distorting public perception. Therefore, there is a crucial need for reliable and automated methods to detect disinformation methods that are accurate,safeguard privacy and reputation, and help preserve trust in media and digital exchanges. Models such as ResNet and EfficientNet,recognized for their depth, scalability and ability to capture complex visual features, are popular options that could deliver strong performance in manipulated image classification [1]. Recently, attention-based architectures such as the vision transformer (ViT) have been proposed. These models utilize self-attention mechanisms to capture long-range relationships within images, enabling them to detect subtle traces of manipulation spread across the spatial region [2]. Besides deep learning, traditional machine learning algorithms like support vector machine (SVM) , random forests, and bagging which utilize engineered spatial, temporal or frequency features have also been used in spoofing detection.

Convolutional neural network(CNNs) are fundamental for capturing layered patterns within image data, which is need for detecting fine facial abnormalities present in deepfake videos [3].CNNs enables to recognize patterns regardless of their position in the image.This is crucial for identifying facial features that might be located differently in deepfake content [4].This work offers an original take on deepfake image detection, combining the benefits of a custom CNN and XceptionNet model. Unlike most studies that depend on large datasets for good performance, our method can potentially achieve accurate detection with a relatively small dataset. This is especially important since gathering extensive labeled deepfake data is often impractical. Our approach is twofold: first, we develop and train a lightweight CNN from

979-8-3315-5525-2/25/$31.00 ©2025 IEEE

scratch, modified specifically for deepfake detection, providing a strong baseline. Second, we implement a transfer learning pipeline using the pretrained Xception model on ImageNet, adding our own dense layers and retraining on our dataset. This dual-model setup allows us to systematically compare features learned from our dataset with the more robust, generalized features derived from the pretrained network.

Our approach's novelty lies in demonstrating how a combination of architecture design and optimized training strategies such as custom preprocessing,freezing certain pretrained layers and incorporating adaptive dense networks can achieve extremely high classification accuracy even with limited data. Additionally, our method is computationally efficient, showing that robust deepfake detection is possible with fewer resources. These contributions collectively establish a general framework as a viable and scalable approach to deepfake forensics and are especially valuable when data or processing capabilities are constrained.

The whole paper is divided into four sections. Section II reviews related works, emphasizing existing methods and recent developments relevant to our research. Section III explains the methodology in detail and provides a comprehensive overview of the proposed model. Section IV presents the experimental results and offers a comparative analysis of various models. Finally, Section V concludes the paper with a summary of the main outcomes and suggestions for future research.

II. RELATED WORK

This paper aims to evaluate the performance of Xception and CNN to determine their effectiveness for deepfake detection. Various successful methods have been used in this field, but the continual advancement of deepfake creation techniques demands more efficient detection methods. Due to the potential threat of deepfakes, research in deepfake detection has been rapidly increasing.

Recent studies have explored using transfer learning to classify deepfake human face images. Hybrid CNN–Transformer models can capture both local and global features. Nguyen et al. [5] introduced a self supervised ViT that performed well even with limited data and provided better interpretability through attention maps. Wang et al. [6] conducted a comprehensive survey on ViT applications in deepfake detection, examining its architectural strengths and deployment issues. Hybrid CNN–Transformer architectures are also gaining traction datasets [7]. Kafi Anan et al. [8] proposed a weighted ensemble of ResNet-34, DeiT and Xception integrated with wavelet features, delivering 93.2% accuracy and 97.4% AUC.

Advanced CNN-focused frameworks continue to be vital. The authors of [9] utilized EfficientNetV2 on FF++ and FFIW10K, achieving 97.9% validation accuracy . EfficientNet B7, recognized for its efficiency and effectiveness across various tasks, provides benefits in deepfake classification [4]. Soudy et al. [10] created a CViT model that integrates CNNs and ViTs, effectively balancing local and global feature extraction to achieve 97% accuracy on FaceForensics++. The authors [11] trained a self-attention VGG16 neural network using face landmarks to extract spatial features.

This study utilizes CNNs and Xception, highlighting their effectiveness in extracting hierarchical spatial features, along with their computational efficiency and capability to generalize across various manipulation techniques [10]. Unlike heavier models such as VGG16, which are prone to overfitting due to their large number of parameters, Xception's depthwise separable convolutions enable more efficient information processing while preserving essential discriminative features [12]. Xception is ideal for scalable deepfake detection pipelines, being a deeper network than older models like InceptionNet and VGG16. While InceptionNet uses mixed kernels for multi-scale features, Xception isolates correlations, enhancing learning of complex transformations. This reduces parameters, boosts learning efficiency and improves performance, especially in resource-limited settings, due to its deep separable convolutions [1]. Dasgupta et al. [13] improved a lightweight CNN by adding squeeze-and-excitation blocks, demonstrating that even smaller CNN models outperform traditional VGG-based systems on StyleGAN-generated deepfakes.

Additionally, a five-layer CNN from another study achieves a high accuracy of 98% when compared to models like Xception and EfficientNet-B0 [14]. Recent research shows that CNNs outperform traditional methods by effectively capturing complex layered spatial patterns in images. For example, the author compared CNNs with transformers and achieved an accuracy of 88.74% across various benchmarks such as FF++, Celeb-DF, and DFDC [15]. Xi and Chen [16] introduced a swin transformer-based method, achieving 71.29% accuracy on the Real & Fake Face dataset, although it underperformed compared to CNN-based methods in low-data settings. This emphasizes the advantage of CNNs and lightweight transfer learning models like Xception in real-world applications with constrained resources.

The growing threat of DeepFake technologies has spurred considerable efforts to develop effective detection systems. Initially, methods focused on handcrafted features and machine learning approaches like SVM and decision trees. Nonetheless, these traditional methods have proven inadequate against the complexity of contemporary generative models [17]. Çetintaş and Yucel [18] proposed a hybrid DenseNet121 architecture and achieved 89% accuracy on a custom deepfake dataset.However, these models were computationally heavy,lacked scalability and did not perform well under resource constraints.

In contrast to complex ViT or DenseNet combinations, our lightweight CNN and Xception models focus on efficient training and inference with minimal computational load. Whereas previous ensemble and hybrid methods can overfit or demand significant resources,our approach prioritizes simplicity and optimization. Trained on the 140K Real and Fake Faces dataset,these models strike a balance between accuracy and efficiency, making them suitable for practical use.

III. PPROPOSED MODEL

The proposed approach to detect deepfake images using different deep learning architectures is described in this section. Once the data was cleaned, it was run through, among other models, CNN and XceptionNet. Importantly, all models were run using the same experimental setup, a crucial step to ensure a fair performance comparison. The assessment used accuracy, confusion matrix and several more amongst others.The general workflow from pre-processing of the data to evaluation of the model is shown in Fig 1.

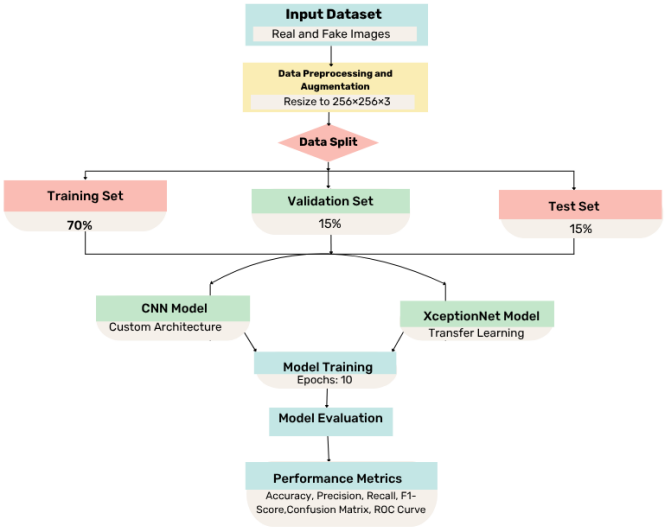


Fig. 1. Workflow architecture of the proposed deepfake detection system

A. Dataset and Preprocessing

A binary classification method was applied in using the 140k Real and Fake Faces of Kaggle dataset [19] which includes 70,000 real faces and 70,000 fake faces. Each of the images was resized to 256x256x3 and was scaled to the interval [0,1]. ImageDataGenerator was used to augment data with techniques of horizontal flipping, rotation, zooming and shifting to improve generalization.Dataset was divided into three parts: training, validation and testing data.The labels were coded as 0 when they were real and 1 when they were fake.CNN and Xception models were each given preprocessed batches to train and test.

B. Model Architecture

The traditional sequential architecture of the custom CNN developed in the current study involves the sequence of layers with a purposeful design to identify and learn specific discriminatory characteristics in terms of facial images.The main aspects of the model are the following:

1) Input Layer: Input images are 256×256×3; the input layer of the network accepts RGB images of dimension 3. All photographs were resized and scaled to ensure that pixel values would fall between the areas of [0,1] before being fed into the model adaptation model which helps accelerate the process of training.

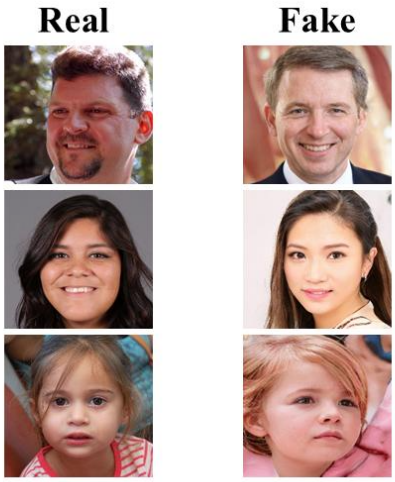


Fig. 2. Sample real and deep-fake images from the dataset

2) Convolutional Layers: The architecture has three convolutional layers where they learn filters and introduce them to the input image and the feature maps. Its first layer has 32 filters, the second and third layers have 64 filters each with kernel size 3×3. It is these layers that are meant to identify local structures like edges, textures, shapes, etc., used in the image. The more the depth level, the more abstract features of DeepFake classification are taken by the model.

3) Activation Function: After every convolutional layer, there is a ReLU (Rectified Linear Unit) activation layer, thereby providing non-linearity to the model. ReLU can be expressed as f(x) = max(0,x) and allows the network to learn and comprehend complicated representations,removing the occurrence of negative activation of neurons but maintaining the positive ones.

4) Pooling layers: In order to gradually diminish the space and computational complexity dimensions, a MaxPooling2D operation is performed after each convolutional block, and its pool size is 2×2. Pooling allows keeping the most noticeable features and removing duplicated spatial information and improves translation invariance and avoids direct overfitting.

5) Fully Connected layers: The dense layers perform the task by transforming the feature maps after convolutional and pooling into a 1D vector. There is a dropout layer (rate = 0.2) introduced to minimize overfitting, where during training 20 percent of neurons are randomly disabled. The last layer comprises a dense layer that contain one neuron and sigmoid activation function.It presents a probability score that can either show that the input image is real or fake.

The hierarchical features are effectively learnt in this structured CNN model and the model performs excellently as a classifier with relatively simple computational complexity.

Table I and Fig. 3 represents the layers of the proposed

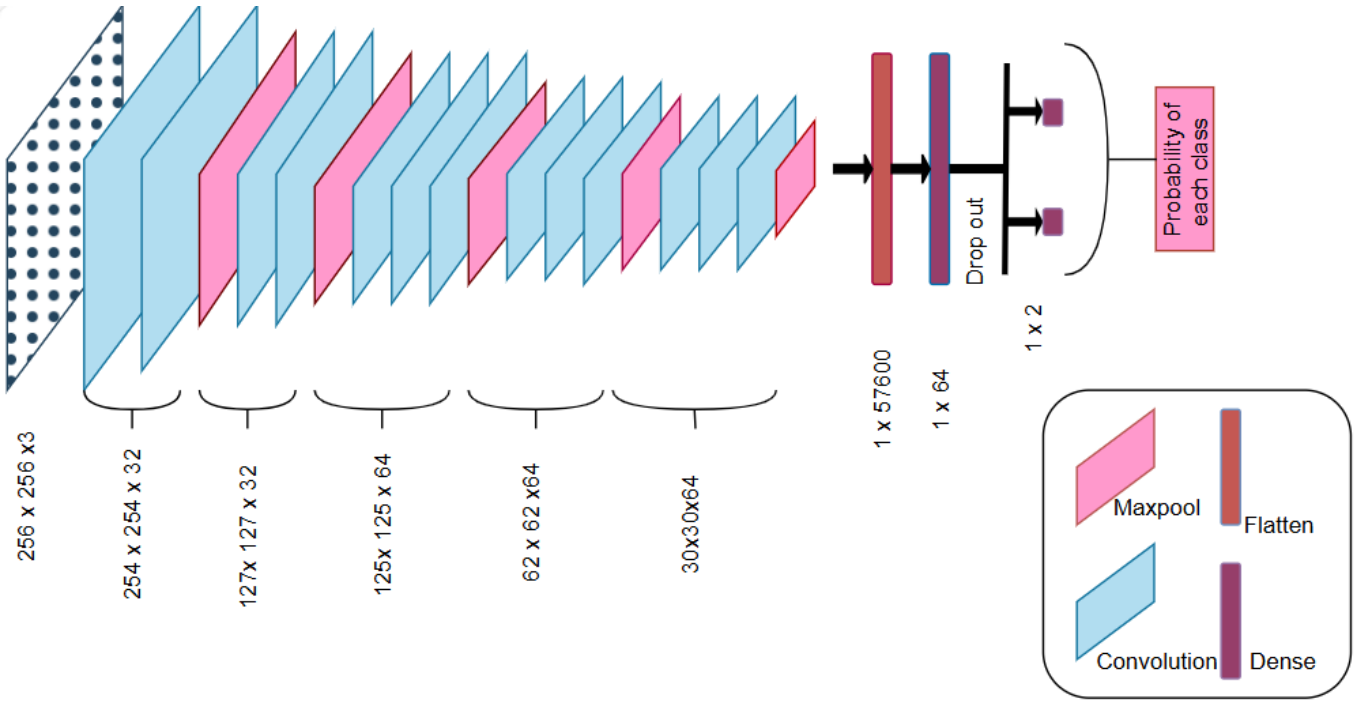


Fig. 3. Architecture diagram of the custom CNN model

TABLE I
SUMMARY OF TRAINING PARAMETERS

Layer Type	Output Shape	Parameters
Input Layer	(256, 256, 3)	0
Conv2D (3×3, 32)	(254, 254, 32)	896
MaxPooling2D (2×2)	(127, 127, 32)	0
Conv2D (3×3, 64)	(125, 125, 64)	18,496
MaxPooling2D (2×2)	(62, 62, 64)	0
Conv2D (3×3, 64)	(60, 60, 64)	36,928
MaxPooling2D (2×2)	(30, 30, 64)	0
Dropout (0.2)	(30, 30, 64)	0
Flatten	(57600)	0
Dense (64, ReLU)	(64)	3,686,464
Dropout (0.2)	(64)	0
Dense (1, Sigmoid)	(1)	65

model. A CNN with five convolutional layers and two dense layers serves as the model network for the 140k dataset.

C. Training Strategy and Optimization Techniques

The optimizer was selected in order to achieve a compromise between speed and precision. This allowed the model to effectively learn the complex visual representations that were necessary for performing well in a DeepFake classification task.

Training Configuration: The CNN model was trained using 10 epochs,binary cross-entropy loss and Adam optimizer with a batch size of 32. Binary Cross-Entropy was taken as the loss function because it is optimal for a binary problem. Overfitting was prevented by using early stopping for obtaining the best generalization possible. All the training parameters used in this process are shown in Table II.

TABLE II
SUMMARY OF TRAINING PARAMETERS

Parameter	Value
Epochs	10
Batch Size	32
Learning Rate	0.001
Optimizer	Adam
Early Stopping	Patience = 3
Loss Function	Binary Cross-Entropy

IV. RESULTS AND DISCUSSION

Here, the results of training the model are shown and the effectiveness of the model are compared with prevalent state-of-the-art deepfake detection methods.

Performance Evaluation in Training: Fig. 4(a) and 4(b) display the training process of the proposed CNN model for 10 epoch. Both training and validation losses also decrease consistently from about 0.42 and 0.26 in the first epoch to about 0.045, and 0.085 in the last epoch for training and validation respectively. Such a steady decrease in loss values is an indication that the model is not overfitting. Similarly, the accuracy curves steeply increase during the first epochs and seem to stabilize at approximately 97.8% training accuracy

TABLE III
PERFORMANCE COMPARISON OF PROPOSED AND EXISTING MODELS

Study's	Method	Data	Accuracy	F1-Score	Precision	Recall
[20]	LightFFDNet v2	140k-real-and-fake-faces	0.71	0.74	0.78	0.71
[1]	Hybrid Dataset Utilizing CNN	140k-real-and-fake-faces	0.88	0.88	0.89	0.86
[21]	CNN	140k-real-and-fake-faces	0.95	0.95	0.93	0.98
Proposed	CNN	140k-real-and-fake-faces	0.97	0.97	0.97	0.97
Proposed	XceptionNet	140k-real-and-fake-faces	0.91	0.91	0.92	0.91

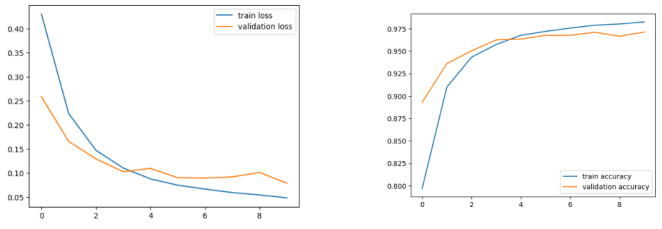
and 97.2% validation accuracy. The closeness between these values demonstrates the generalization of the model to new data. The general training behavior confirms the stability of the chosen CNN architecture and the validity of the training approach use.

The CNN model also performed well in classification, obtaining 9836 true positive, 9596 true negatives, while generating 164 false positive and 404 false negative. This indicates a very reliable capability to tell apart real and fake images.In comparison, XceptionNet reported 9440 true positives and 8831 true negatives, but had higher misclassifications—560 false positives and 1169 false negatives.. This shows how the custom CNN model is more accurate and is a better representation of the balance between precision and recall. XceptionNet takes advantage of transfer learning, though it might need additional subsequent fine-tuning. In general, the confusion matrix confirms the CNN architecture is a reliable model for detecting deepfakes on the 140k dataset. Fig. 5(a) CNN Confusion Matrix and Fig. 5(b) XceptionNet Confusion Matrix.

Table III compares the performance of our models CNN and XceptionNet with others Deepfake detection methods. As stated, the designed CNN model achieved the highest accuracy of 0.97, F1-score of 0.97, precision of 0.97 and recall of 0.97, thus showing superior classification performance against the other models. XceptionNet has a competitive 0.91 score in all metrics but it is slightly lower.

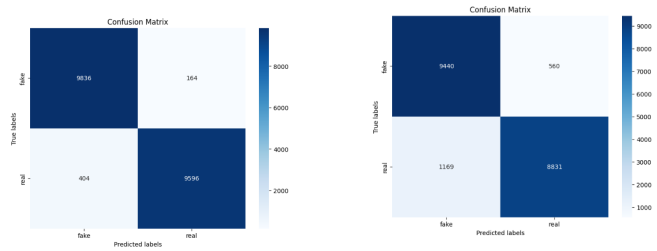
Among existing models, the existing CNN network show decent performance (accuracy = 0.951, F1-score = 0.956), while the hybrid dataset utilizing CNN by Mallet et al. [1] achieved moderate results. The LightFFDNet V2, proposed by Jabbar et al. [20] performed poorly with all metrics below 0.78.This comparison highlights the superiority of the proposed models, which is lightweight and very efficient when trained and tested on 140k real-and-fake-faces dataset.These results confirm the advantages of optimized architectures as opposed to ensemble or transformers based architectures.

These results verify the benefits of optimized architectures over more complex ensembles or transformer-based models,especially in resource-limited settings.The custom CNN we proposed achieved top classification scores 97% across accuracy, precision, recall and F1-score and maintained low false positive and negative rates, ensuring high reliability for



(a) Loss Curve (b) Accuracy Curve

Fig. 4. Training and validation curves for CNN model: (a) loss and (b) accuracy.



(a) CNN (b) XceptionNet

Fig. 5. Confusion matrices of the proposed models.

real-world use. Although XceptionNet also performed well, it exhibited slightly higher misclassification rates,suggesting that additional fine-tuning could bring its performance closer to our CNN's consistency. Both models significantly outperformed traditional architectures, such as VGG16 and certain hybrid deep learning models, demonstrating the usefulness of our approach keeping a balance between efficiency and accuracy. Additionally, the CNN's simplicity and lower computational demands make it suitable for deployment on edge devices or systems with low resources like mobile apps or embedded forensic tools. These findings emphasize the value of designing purpose-built, lightweight architectures that can generalize effectively without extensive datasets or powerful infrastructure, marking an important step toward scalable, accessible deepfake detection.

V. CONCLUSION

This study introduced a dual-model approach for detecting deepfake images, utilizing a custom, lightweight CNN and a

transfer learning-based XceptionNet. Trained and tested on the 140K Real and Fake Faces dataset, the CNN achieved an outstanding classification accuracy of 97%, surpassing the fine-tuned XceptionNet, which reached 91%. The CNN's superior accuracy, combined with its efficiency and low resource requirements, highlights the benefits of a tailored architecture, especially when data and processing power are limited.Also, the stable results for key metrics accuracy, precision, recall, and F1-score, are indicative of the stability of the complete system. This research demonstrates that carefully optimized lightweight models can rival or surpass more complex, resource-intensive models. Future work will incorporate ensemble techniques and aim for real-time deployment to facilitate scalable, immediate deepfake detection in real-world scenarios.

REFERENCES

[1] J. Mallet, L. Pryor, R. Dave, and M. Vanamala, "Deepfake detection analyzing hybrid dataset utilizing cnn and svm," *arXiv preprint arXiv:2302.10280*, 2023.

[2] W. Lu, L. Liu, B. Zhang, J. Luo, X. Zhao, Y. Zhou, and J. Huang, "Detection of deepfake videos using long-distance attention," *IEEE transactions on neural networks and learning systems*, vol. 35, no. 7, pp. 9366–9379, 2023.

[3] K. Omar, R. H. Sakr, and M. F. Alrahmawy, "An ensemble of cnns with self-attention mechanism for deepfake video detection," *Neural Computing and Applications*, vol. 36, no. 6, pp. 2749–2765, 2024.

[4] A. M. Kalemullah, P. Prakash, and V. Sakthivel, "Deepfake classification for human faces using custom cnn," in *2024 7th International Conference on Circuit Power and Computing Technologies (ICCPCT)*, vol. 1, pp. 744–750, IEEE, 2024.

[5] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Exploring self-supervised vision transformers for deepfake detection: A comparative analysis," in *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10, IEEE, 2024.

[6] Z. Wang, Z. Cheng, J. Xiong, X. Xu, T. Li, B. Veeravalli, and X. Yang, "A timely survey on vision transformer for deepfake detection," *arXiv preprint arXiv:2405.08463*, 2024.

[7] D. Wodajo, S. Atnafu, and Z. Akhtar, "Deepfake video detection using generative convolutional vision transformer," *arXiv preprint arXiv:2307.07036*, 2023.

[8] K. Anan, A. Bhattacharjee, A. Intesher, K. Islam, A. Assaeem Fuad, U. Saha, and H. Imtiaz, "Hybrid deepfake image detection: A comprehensive dataset-driven approach integrating convolutional and attention mechanisms with frequency domain features," *arXiv e-prints*, pp. arXiv–2502, 2025.

[9] L. Deng, H. Suo, and D. Li, "Deepfake video detection based on efficientnet-v2 network," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 3441549, 2022.

[10] A. H. Soudy, O. Sayed, H. Tag-Elser, R. Ragab, S. Mohsen, T. Mostafa, A. A. Abohany, and S. O. Slim, "Deepfake detection using convolutional vision transformers and convolutional neural networks," *Neural Computing and Applications*, vol. 36, no. 31, pp. 19759–19775, 2024.

[11] S. Asha, P. Vinod, and V. G. Menon, "A defensive framework for deepfake detection under adversarial settings using temporal and spatial features," *International Journal of Information Security*, vol. 22, no. 5, pp. 1371–1382, 2023.

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[13] S. Dasgupta, J. Mason, X. Yuan, O. Odeyomi, and K. Roy, "Enhancing deepfake detection using se block attention with cnn," in *2024 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*, pp. 1–6, IEEE, 2024.

[14] J. B. Awotunde, R. G. Jimoh, A. L. Imoize, A. T. Abdulrazaq, C.-T. Li, and C.-C. Lee, "An enhanced deep learning-based deepfake video detection and classification system," *Electronics*, vol. 12, no. 1, p. 87, 2022.

[15] V. L. Thing, "Deepfake detection with deep learning: Convolutional neural networks versus transformers," in *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 246–253, IEEE, 2023.

[16] A. J. Xi and E. Chen, "Classifying deepfakes using swin transformers," *arXiv preprint arXiv:2501.15656*, 2025.

[17] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deepfake videos from phoneme-viseme mismatches," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 660–661, 2020.

[18] D. Çetintaş and Z. Yücel, "Deepfake detection using fine-tuned cnn architectures," *Celal Bayar University Journal of Science*, vol. 21, no. 1, pp. 121–128, 2025.

[19] Xhlulu, "140k real and fake faces." https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces, 2020.

[20] G. Jabbarlı and M. Kurt, "Lightffdnets: Lightweight convolutional neural networks for rapid facial forgery detection," *arXiv preprint arXiv:2411.11826*, 2024.

[21] J. Sharma, S. Sharma, V. Kumar, H. S. Hussein, and H. Alshazly, "Deepfakes classification of faces using convolutional neural networks," *Traitement du Signal*, 2022.