JOURNAL OF COMPUTING HORIZONS
Vol. 01, No. 01, 2026
DOI: 10.22610/JCH.V1I1.4990

RESEARCH ARTICLE | PEER REVIEWED

Convolutional Neural Network and Transformer-Based Visual Deepfake Detection: A Comparative Analysis

Muhammad Salman [1], Iqra Tariq [1], Zameer Ahmad [2] *, Naila Batool [3]
[1] *Department of Computer Science, GIFT University, Gujranwala, Pakistan*
[2] *School of Computer Sciences, Universiti Sains Malaysia, Penang, 11800, Malaysia*
[3] *Department of Computer Science, University of Management and Technology, Lahore (Sialkot Campus), Pakistan*
* **Corresponding Author:** zameerahmad.cs@gmail.com

ABSTRACT – This paper explores the performance of several deep learning architectures for visual deepfake detection, including Swin Transformer, Vision Transformer (ViT), Xception, ResNet, and EfficientNet. We evaluate these models on the CelebDF-v2 dataset, assessing their effectiveness in detecting manipulated facial images using performance measures such as accuracy, AUC, and F1 score. Our experiments reveal that the Swin Transformer consistently outperforms other architectures, achieving an F1 score of 91.17% and AUC of 55.61%, substantially higher than the Vision Transformer's 86.85% F1 and 34.19% AUC, showing its superior ability to detect subtle visual artifacts in deepfakes. Additionally, to enhance model interpretability, we employ GradCAM visualizations to highlight the regions of input frames that are most influential to the model's predictions. These results demonstrate the potential of both CNN and transformer-based architectures to enhance deepfake detection accuracy and the value of visual explanations for improving model transparency and reliability in media forensics.

Keywords: *Image forgery; GradCAM; synthetic images; deep-fake.*

I. INTRODUCTION

One of the breakthroughs in the field of AI is image forgery, which is also known as deepfakes. Deepfakes are fake videos, images, or audio created using artificial intelligence (AI). [1] Deepfakes are defined as believable media generated using deep neural networks, and this summarizes the whole story very well. This technology uses advanced machine learning models to manipulate or generate content that looks and sounds real, even though it's entirely fake. While deepfakes have legitimate applications in entertainment, education, and accessibility, their misuse poses significant threats, including misinformation, fraud, manipulating public opinion, and invasion of privacy [2]. These forged videos are generated using AI assistance and can serve multiple purposes. This technology primarily leverages deep neural networks, particularly Generative Adversarial Networks (GANs), to produce highly realistic images, videos, or audio recordings that are virtually indistinguishable from real content [3, 4, 5]. Figure 1 illustrates the processing pipeline of face re-enactment, a technique used in deepfake generation alongside other methods such as face replacement, editing, and synthesis.

Deepfakes have revolutionized the entertainment industry, enabling actors to perform stunts or deliver lines they never actually recorded, and filmmakers to create hyper-realistic visual effects [1, 6, 7]. Deepfakes are improving movie dubbing by lip-synchronizing and matching facial expressions to the translated dialogue of the target language. This allows for a wide reach of audience, yet the expressions remain native to the language, giving the audience special attachment [6, 8]. While the technology behind deepfakes has shown tremendous promise in various creative fields, such as computer-generated imagery (CGI) for movies, the creation of digital doubles for actors, and voice dubbing, it also carries significant risks when used maliciously [6].

The malicious use of deepfakes can be used for unethical purposes, such as defamation, political manipulation, revenge porn, and other forms of misinformation, raising concerns about their impact on privacy, security, and public trust [9]. Deepfakes have reached the tipping point, posing a national security threat, as evidenced during the early phase of the Russo-Ukrainian War [10, 11]. A deepfake of President Zelenskyy emerged ordering the Ukrainian military to surrender to the Russian army [12]. In a 2019 video, Zuckerberg was making controversial statements about Facebook's data practices and user privacy [13]. The video showed Zuckerberg boasting about having "total control of billions of people's stolen data, all their secrets, their lives, their futures". As deepfake technology advances, so does the urgency of detecting and mitigating its malicious uses. The detection of deepfakes has become a critical area of research, particularly due to the rise of fake media in political, social, and personal contexts. Traditional detection methods often struggle with detecting subtle visual inconsistencies, especially as deepfake generation models improve [14, 15].

This study compares the performance of various deep learning models, including Convolutional Neural Networks (CNNs) and transformer-based architectures, in deepfake detection. The aim is to provide solutions to mitigate the risks posed by fake media, curtail misinformation, and ensure the trustworthiness of digital content.

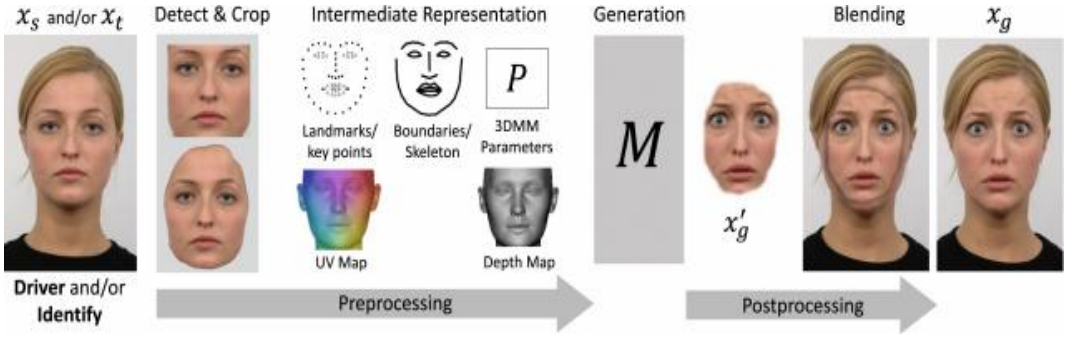


FIGURE 1: The Processing Pipeline for Making Re-Enactment and Face-Swap Deep-Fakes [1]

The rest of the paper is structured as follows: Section II reviews current research and highlights the limitations in deepfake detection. Section III covers the methodology, including architecture, data preprocessing, and evaluation metrics. Section IV, Section V, and Section VI detail the implementation, training, optimization, and results analysis. Section VII concludes with key insights and future research directions.

II. RELATED WORK

Deepfake technology has evolved significantly, utilizing GANs to create highly realistic synthetic images, making detection increasingly challenging. This literature review summarizes key contributions and advances in deepfake detection, highlighting novel techniques and frameworks. Graph Neural Networks (GNNs) are a type of deep learning models designed to analyze and process data represented as graphs. [16] demonstrated that by treating images as non-Euclidean data, GNNs can infer relationships between image patches to enhance generalization across various deepfake datasets. The study proposes a dual-branch network that extracts features from both RGB images and color difference images (CDIs), resulting in improved performance over existing methods on benchmark datasets.

The Thumbnail Layout (TALL) method, introduced in [17], presents a novel approach to detecting deepfake videos. TALL efficiently captures spatial-temporal dependencies by transforming video clips into a predefined thumbnail layout, preserving both spatial and temporal information without adding high computational costs. This method is noteworthy for its model-agnostic approach and its applicability across different deepfake manipulation techniques. The AWARE-NET [18] technique looks promising, as it introduces adaptive weighting schemes to better balance model contributions based on their performance, thereby improving generalization. However, there is a significant limitation of this approach: substantial compute power is required, as multiple instances of several architectures are used in its base, which limits its real-world use.

The integration of audio-visual features has also been explored to improve deepfake detection. The study [19] introduced the Predictive Visual-Audio Alignment Self-Supervision (PVASS-MDD) method, which detects deepfakes by analyzing the inconsistencies between visual and audio signals. (PVASS-MDD) is a two-stage process, with the first stage focusing on capturing real visual-audio correspondences, and the second stage targeting deepfake inconsistencies. In a similar vein, the paper [20] proposes a multi-task learning strategy that combines binary classification losses from both audio and video modalities. This study suggested that the combination of audio-visual information provides a more reliable and effective means for detecting deepfakes.

Data augmentation and synthetic data generation have been widely adopted to improve dataset diversity and reduce overfitting in deep learning models. Several surveys highlight the importance of augmentation techniques in enhancing robustness across computer vision tasks [21]. Additionally, fairness-aware augmentation strategies [22] have been explored to address demographic bias in facial datasets. Synthetic data generation methods such as SYNAuG [23] have also been proposed to mitigate class imbalance by augmenting underrepresented categories. Furthermore, recent deepfake detection approaches integrate both local CNN features and global transformer representations to capture complementary information, thereby improving detection performance [24].

Common Forgery Artifact Extraction Network (CFNet) was introduced in [25] with the aim of identifying manipulated facial images. It does so by filtering out common forgery artifacts hidden in low-quality images. Another innovative approach, proposed in [26], is the Contextual Cross-Modal Attention framework, which uses Recurrent Neural Networks (RNNs) to enhance deepfake detection by integrating contextual information from both audio and video.

Catastrophic forgetting in neural networks occurs when a model forgets previously learned information after being trained on new tasks. Addressing the challenge of catastrophic forgetting, [27] introduced an incremental learning framework for face forgery detection. This method uses an adapter-based approach combined with a confidence-based ensemble prediction mechanism to improve generalization and performance across multiple benchmark datasets.

With the rise of ViT and Swin, researchers have employed Transformer-based models for detecting deepfakes. [28] proposed the High-Frequency Fine-Grained Transformer (F2Trans), which focuses on fine-grained forgery cues using a Central Difference Attention (CDA) module and a High-frequency Wavelet Sampler (HWS). To better capture salient characteristics while preserving efficiency, the Shuffle Transformer, for example, combines region-of-interest (ROI) and spatial-encoding (SE) modules [29]. Similarly, other transformer-based approaches, such as ConvNext-Pnet [30], prioritize model interpretability and explainability, thereby enhancing the trustworthiness of deepfake detection systems by providing insights into their decision-making processes.

III. METHODOLOGY

In this study, we build upon the work presented by P. P. Chandran et al., which explores the use of transformers in the audio-visual domain for deepfake detection [31]. It proposed a tri-network architecture that integrates video, audio, and audio-visual modalities and was evaluated on the FakeAVCeleb dataset using ViT and Swin Transformers. To expand on this, we first examined the dataset samples and applied data augmentations. Figure 2a shows real and fake frames extracted from the CelebDF-v2 dataset, while Figure 2b illustrates the transformations and augmentations applied to enhance dataset diversity. We trained and evaluated the following models: Xception, ResNet, EfficientNet, ViT, and Swin Transformers.

Figure 3 provides a detailed representation of the Swin Transformer architecture, showcasing its hierarchical structure, patch partitioning, and self-attention mechanisms. The complete framework pipeline, including training, testing, and evaluation for all models, is depicted in Figure 4 and Figure 5. The full process for deepfake detection, including frame extraction, data augmentation, model training, and evaluation, is detailed in Algorithm 1. The left subfigure illustrates real and fake samples from the dataset, while the right subfigure demonstrates the transformations and augmentations.

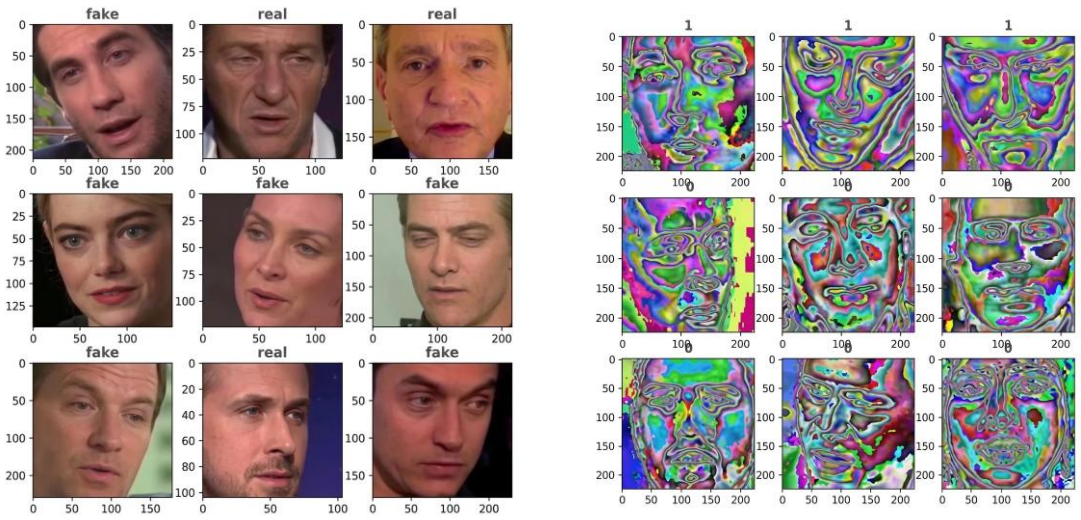


FIGURE 2: **(a)** Real and Fake Samples from the Dataset; **(b)** Transformations and Augmentations Applied

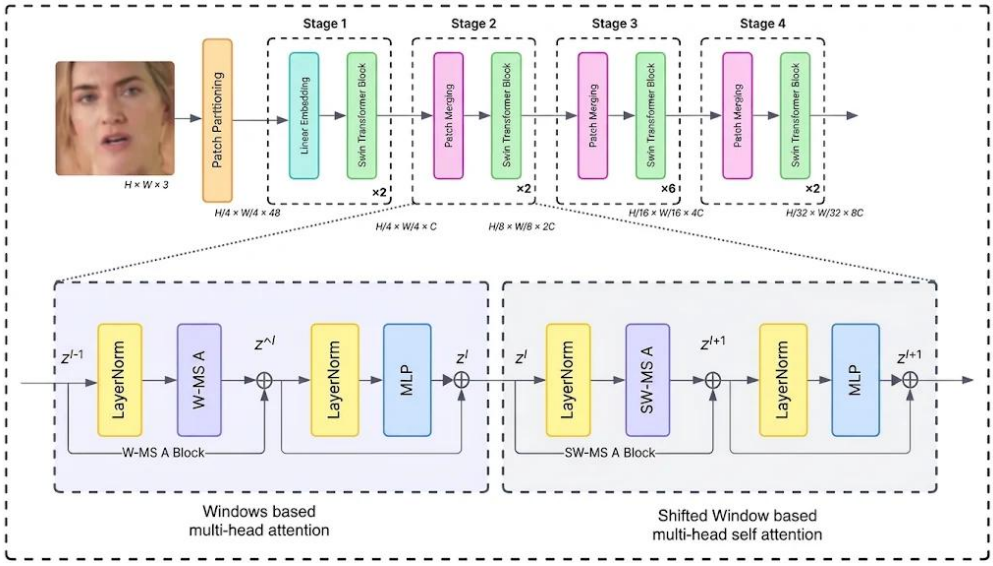


FIGURE 3: Detailed Representation of the Swin Transformer Architecture, Showcasing its Hierarchical Structure, Patch Partitioning, and Self-Attention Mechanisms

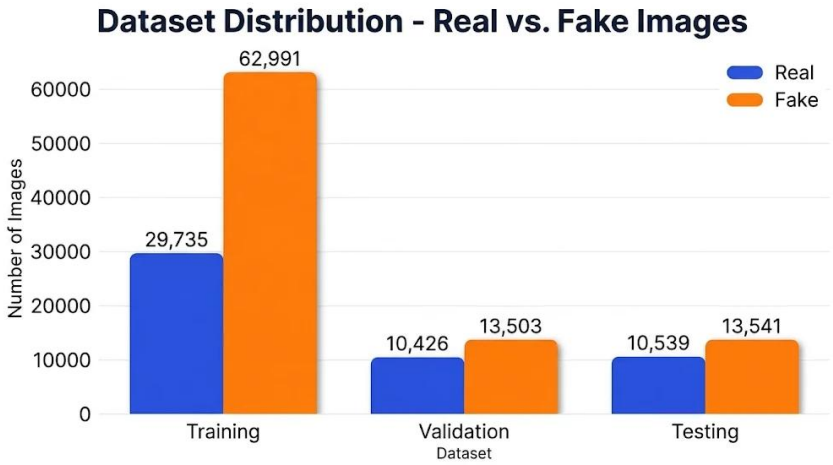


FIGURE 4: Dataset Division into Various Subsets

We trained and evaluated the following models: Xception, ResNet, EfficientNet, ViT, and Swin Transformers. The goal was to compare these architectures in terms of accuracy, robustness, and overall effectiveness in detecting deepfakes across the CelebDF-v2 dataset. Our framework pipeline is depicted in Figure 5, which shows the workflow for training, testing, and evaluating each model. The complete process for deepfake detection used in the study is detailed in Algorithm 1, which outlines the steps involved in frame extraction, data augmentation, model training, and evaluation.

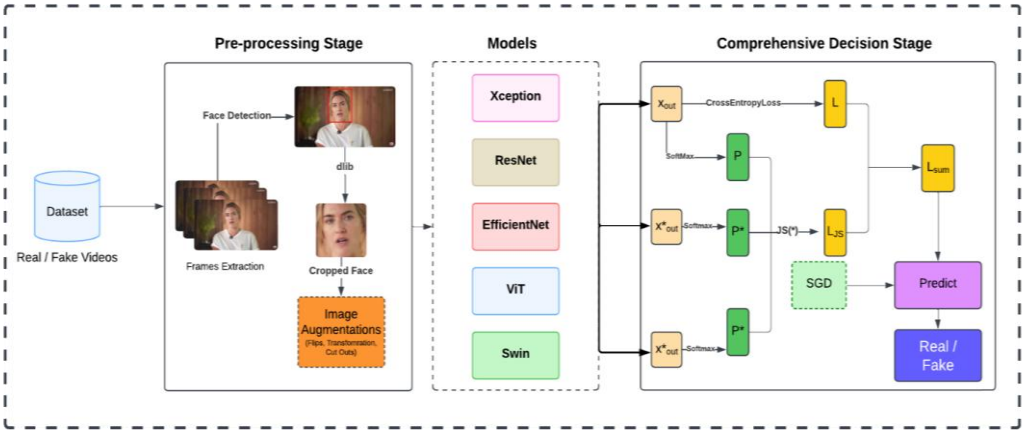


FIGURE 5: Framework Pipeline Used in the Study

A. Dataset

The CelebDF-v2 dataset [32] is a widely recognized benchmark for evaluating deepfake detection models, specifically targeting manipulated videos of celebrity faces. It consists of 890 real videos, sourced from YouTube and paid actors (590 and 300, respectively). By applying the latest forgery methods, it generated 5,639 fake videos from those original videos. The videos in CelebDF-v2 are diverse, containing different facial expressions, lighting conditions, and backgrounds, making it a valuable resource for training and testing deepfake detection models. The fake videos are generated using advanced face-swapping techniques, making the dataset particularly challenging for detection models.

The dataset preparation step took the most time since the authors did not provide predefined guidelines. Another challenge to tackle was that the dataset is imbalanced in nature. Only in CelebDF-v2 is there a 10:1 class imbalance between fake and real. To address the data imbalance problem, we extract 3 times as many frames from the real video as from the fake video. After this, the extracted real frames are half as many as the fake frames, as shown in Figure 4. We conducted experiments by extracting frames at random points throughout the video, rather than selecting a single random point and extracting consecutive frames from it. We divided the dataset into a 70:15:15 ratio for training, validation, and testing the model. Before splitting, the dataset underwent 15 iterations of random shuffling to ensure unbiased distribution. 70% allocated for training includes 70% of both real and fake videos; the same applies to the validation and testing splits. n number of frames are extracted from the video, where for real videos n=48 and n=16 for fake videos to handle imbalanced class distribution.

B. Preprocessing and Augmentation

Faces from the frames are detected and extracted using dlib. The input data for training and evaluating models is a single face cropped image [3, 224, 224]. The extracted faces are augmented, scaled, normalized, and resized before being fed

to the model for training, as illustrated in Figure 2. For augmentations, we used imgaug [33] to apply horizontal flips, affine transformations, and random cut-out on the cropped faces.

In the study, we used a detailed approach to forecast temperature categories using two different datasets. The strategy entailed the use of various machine learning models, namely a hybrid Convolutional Neural Network and Gated Recurrent Unit (CNN-GRU), Random Forest, XGBoost, Transformer, Temporal Fusion Transformer (TFT), and LightGBM. The description of the methodology was based on several important steps: a large amount of data was collected and preprocessed to ensure the data was clean and representative. This involved strategies for handling missing values, feature engineering to create new, relevant attributes, and encoding categorical variables. This was followed by resampling the data to address class imbalance and scaling to improve the model. All the models were trained with regard to particular hyperparameters to maximize their predictive power. Lastly, we evaluated the models using accuracy, precision, recall, and F1-score, and generated confusion matrices to assess how each model performed across temperature categories. This multi-perspective method was used to determine the most appropriate model for predicting temperature and to gain insight into the nature of the data.

ALGORITHM 1: Deepfake Detection Model-Agnostic Approach

Input: Video file path (video_path)
Output: Classification label (Real or Fake)

```
video_path ← "input_video.mp4"
frames ← ExtractRandomFrames(video_path, num_frames=48)
for each frame:
    grayscale_frame ← ConvertToGrayscale(frame)
    faces ← DetectFaces(grayscale_frame)
    for each face in faces:
        cropped_face ← CropFace(frame, face)
    end
end
augmented_faces ← [ ]
for each cropped_face:
    resized_face ← Resize(cropped_face, size←(224, 224))
    flipped_face ← RandomFlip(resized_face)
    normalized_face ← Normalize(flipped_face)
    augmented_faces.append(normalized_face)
end
model ← InitializeModel(num_classes=2)
loss_function ← CrossEntropyLoss()
optimizer ← SGD(model.parameters(), lr=0.001, decay_rate=0.1)
for each epoch:
    for each batch in augmented_faces:
        predictions ← model(batch)
        loss ← loss_function(predictions, labels)
        optimizer.step()
    end
end
probabilities ← Softmax(predictions)
predicted_class ← ArgMax(probabilities)
accuracy ← ComputeAccuracy(predictions, labels)
auc ← ComputeAUC(predictions, labels)
f1_score ← ComputeF1Score(predictions, labels)
logloss ← ComputeLogLoss(predictions, labels)
```

IV. EXPERIMENTAL SETUP

We used PyTorch to facilitate training and evaluating our model. Following previous studies, we set the batch size to 24 images and selected SGD (Stochastic Gradient Descent) as the optimizer, with a learning rate of 1x10-3. The learning

rate is set to decay by a factor of 0.1 every 10 epochs. We chose Cross-Entropy Loss, which is common for binary classification tasks. Accuracy, AUC, and F1 scores are reported here because they provide a balanced overview of precision and recall, which is required in this scenario. For the evaluation stage, we use the scikit-learn library. We use scikit-learn to calculate and report LogLoss, AUC, F1, and Accuracy scores [34]. Our model is trained for 20 epochs, with the final selection being the model with the lowest validation loss for subsequent evaluation and testing. All experimental settings, model configurations, and computational environment details are summarized in Table 1. To facilitate our model implementations and leverage pre-trained weights, we heavily rely on the `PyTorch` Image Models repository by Ross Wightman [35].

TABLE 1: Details of Experimental Setup, Model Configurations, and Training Parameters

Component	Setting / Value
Model Architectures	Xception, ResNet-50, EfficientNet-B0, Vision Transformer (ViT-B/16), Swin-B
Pretrained Weights	All Models Initialized with ImageNet Pretrained Weights
Input Image Size	224 × 224
Optimizer	Stochastic Gradient Descent (SGD)
Learning Rate	3×10^{-3}
Scheduler	StepLR
Scheduler Step Size	10 Epochs
Scheduler Gamma	0.1
Loss Function	CrossEntropyLoss
Batch Size	32
Number of Epochs	20
Data Shuffling	Enabled for Both Training and Validation Loaders
Data Augmentation	RandomHorizontalFlip, Cut-Out Augmentations (Random Cut-Out / Face Cut-Out)
Normalization	ImageNet Mean/Std + Additional (0.5, 0.5, 0.5) Normalization
Face Detection	dlib frontal face detector
Landmark Model	shape_predictor_81_face_landmarks.dat
Model Checkpointing	Best Model Selected Based on Lowest Validation Loss
Hardware	Intel Xeon CPU, NVIDIA RTX 4080 (24GB VRAM), 64 GB RAM
Operating System	Windows 10
Environment	Conda Virtual Environment
Frameworks Used	PyTorch, timm, torchvision, dlib

V. RESULTS AND DISCUSSION

This section presents a performance evaluation of CNN- and transformer-based models for the deepfake detection task. Table 2 presents the training and validation results in terms of Accuracy and LogLoss. In contrast, Table 3 shows the evaluation results of different architectures in terms of F1, AUC, and Accuracy, providing insights into the comparative strengths and weaknesses of the models.

TABLE 2: Performance Metrics for Different Architectures on Training and Validation Subsets

Architectures	Training		Validation	
	Accuracy	LogLoss	Accuracy	LogLoss
Xception	95.78%	0.14540	90.57%	0.27840
ResNet	99.13%	0.02267	95.14%	0.18701
EfficientNet	97.97%	0.05340	92.70%	2.45800
ViT	99.46%	0.01386	97.12%	0.13374
Swin	99.02%	0.04799	99.59%	0.18500

TABLE 3: F1, AUC, and Accuracy of Evaluated Architectures for Deepfake Detection

Architecture	F1	AUC	Accuracy
Swin	91.17%	55.61%	89.58%
ViT	86.85%	34.19%	85.96%
ResNet	77.06%	74.40%	67.81%
EfficientNet	76.52%	78.71%	65.85%
Xception	76.93%	86.75%	66.27%

Figure 6 illustrates the performance comparison across different architectures. Subfigure **(a)** presents the training and validation accuracy, highlighting the generalization capability of each model. Subfigure **(b)** shows the corresponding log-loss values, indicating the confidence and calibration of the predictions. Subfigure **(c)** depicts the gap between

training and validation accuracy, providing a direct measure of potential overfitting. The results demonstrate that while CNN-based models exhibit varying degrees of overfitting, Transformer-based architectures maintain more consistent performance across training and validation sets, suggesting better generalization.

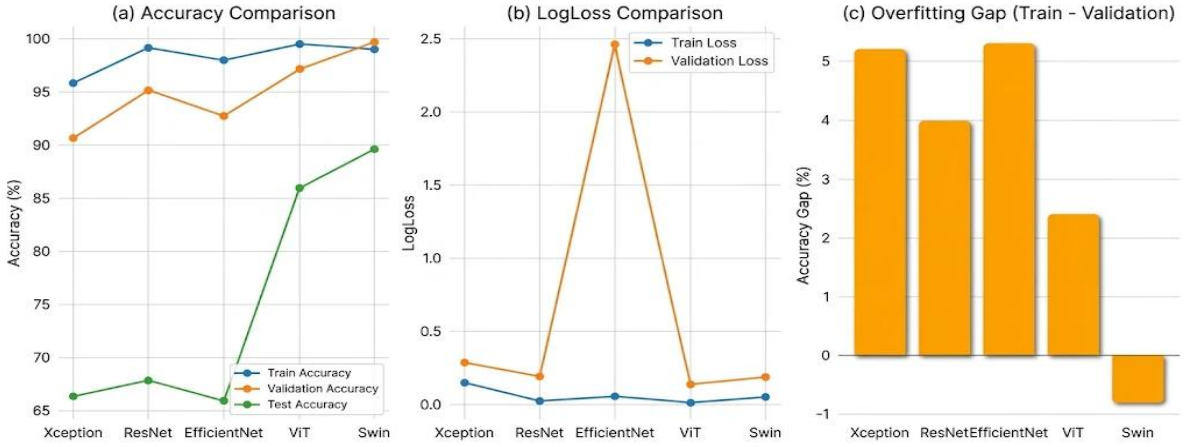


FIGURE 6: **(a)** Training vs. Validation Accuracy **(b)** Training vs. Validation Log Loss **(c)** Overfitting Gap Across Architectures; The results show that Transformer-based models exhibit more stable performance and smaller generalization gaps compared to CNN-based models.

Figure 7 heatmap visualization provides an intuitive comparison of the performance metrics (F1, AUC, and accuracy) across the different architectures. The Swin Transformer, with its high F1 score and accuracy, shows concentrated attention on regions critical for deepfake detection, while models like ResNet and Xception, despite having higher AUC, focus differently, emphasizing class separation. This visualization aids in understanding the varying strengths of each model in detecting subtle deepfake artifacts.

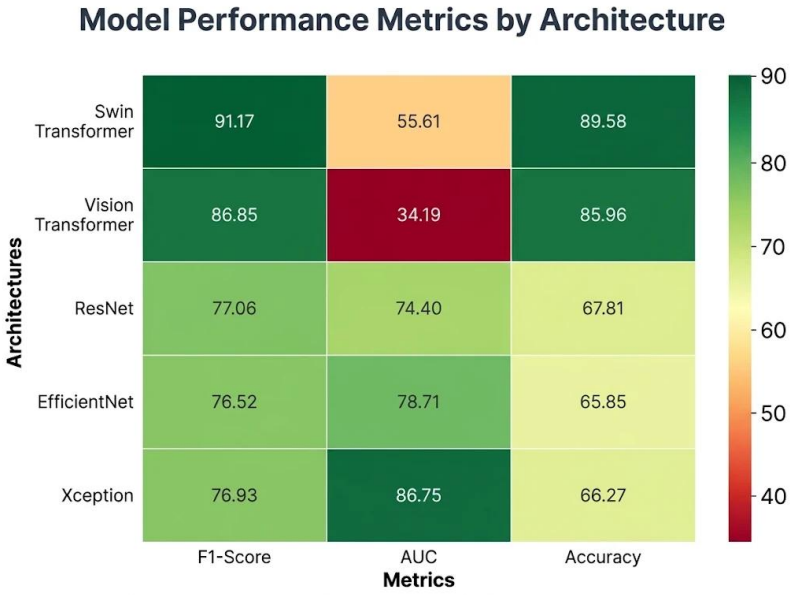


FIGURE 7: Heatmap Visualization of Evaluation Results

Additionally, Grad-CAM visualizations are shown in Figure 8. These visualizations indicate the regions of the input face frames that contribute most to the model's predictions, providing insight into the model's decision-making process. These visualizations are of utmost importance for interpretability: why the model reached the decision to classify this image or video as forged or real. When comparing CNN-based architectures (Xception, ResNet, and EfficientNet) with transformer-based models (ViT and Swin), transformer models generally achieve higher validation accuracy and lower log loss, indicating better generalization. While CNNs like ResNet achieved high training accuracy, they exhibited a slight drop in validation performance, suggesting some overfitting. Xception is good at capturing local artifacts and fine details; ResNet is good at capturing features across multiple scales due to its residual connections.

Overfitting is likely due to CNNs' reliance on local receptive fields, which may struggle to capture global dependencies in complex tasks, as their strength lies in capturing local features and patterns in an image. In contrast, transformers such as ViT and Swin excel at capturing global dependencies via the self-attention mechanism, enabling them to maintain strong performance on both training and validation datasets [36, 37]. The global perspective allows

transformers to focus on relevant features across the entire image, leading to more robust generalization and demonstrating that transformers are better suited for handling complex patterns and larger datasets. Overall, transformers appear to outperform CNNs in terms of validation accuracy and consistency, as evidenced by Tables 2 and 3.

VI. LIMITATIONS AND FUTURE WORK

One limitation of this study is that we did not evaluate cross-dataset generalization. Our decision to focus on CelebDF-v2 was intentional: it allowed for a fair comparison of architectural performance under controlled conditions. But of course, deepfake detection models must eventually face the messier question of robustness across datasets, DFDC, FaceForensics++, DeeperForensics-1.0 – all of which introduce their own distributional quirks and manipulation techniques. We didn't take up that challenge here, but we recognize its importance and plan to address it in future work.

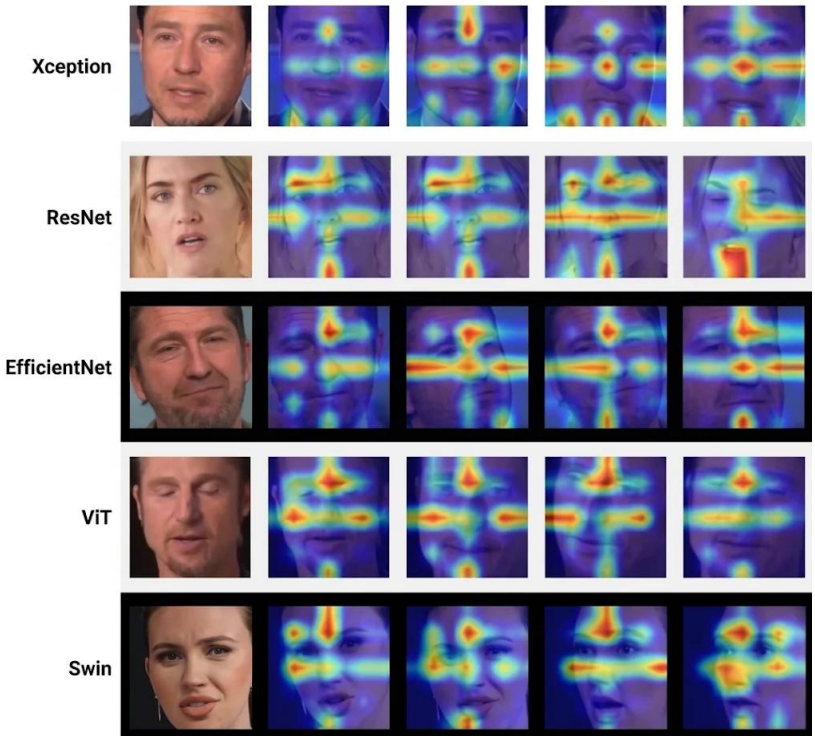


FIGURE 8: Grad-CAM Visualizations Highlighting the Areas of Focus for the Model During Evaluation

Class imbalance posed another difficulty. Our approach, extracting more frames from real videos, was a practical response, not a principled one. We considered oversampling and class-weighted loss, but neither proved straightforward: oversampling risked redundancy; weighting alone didn't offset the skew. We suspect this is a case where theoretical elegance gives way to empirical stubbornness, and we intend to test these strategies more systematically in the future.

Finally, we didn't run ablation studies on individual augmentation techniques, largely due to resource constraints, both time and compute. Instead, we trained on a composite of augmentations without isolating their effects. This limits our ability to say which augmentations matter most, cut-outs, flips, affine transforms, etc., though we plan to unpack that in future iterations.

The results of this study extend beyond the specific dataset and offer useful insights into how CNN- and Transformer-based models perform in deepfake detection. The observed trends show that Transformer-based models are generally better at handling complex patterns and may work more effectively in real-world scenarios. These findings can help researchers and practitioners choose suitable models for similar computer vision problems, especially when dealing with subtle image manipulations. The comparison approach used in this study can also be applied to other areas such as image forensics, biometric verification, and multimedia security.

However, these results should be tested on more datasets and in real-world settings to confirm their reliability. This is an important direction for future work. Even with these limitations, we believe the study offers a solid foundation for architectural comparison in deepfake detection. Much remains undone, but some things had to be left undone to get anything done at all.

VII. CONCLUSION

The paper compared CNN and transformer-based models for deepfake detection. While the results did not achieve state-of-the-art performance, they aligned with the broader trend in the literature, where transformer models generally outperform CNNs [38, 39, 40, 41]. Interestingly, despite lower AUC scores, the transformer models achieved higher F1

scores than CNNs, which had relatively higher AUC scores. Future work will focus on cross-dataset evaluations to assess and enhance the generalization capabilities of these models using an ensemble learning approach.

DECLARATIONS

Author Contributions: All authors contributed equally to this work.

Funding: The authors received no external funding for this research.

Competing Interests: The authors declare no competing interests.

Data Availability: The datasets analyzed in this study are publicly available from their respective sources. Further information regarding dataset sources and access is provided in the manuscript.

REFERENCES

[1] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," ACM Comput Surv, vol. 54, no. 1, p. 7:1-7:41, Jan. 2021, doi: 10.1145/3425780.

[2] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," IEEE Access, vol. 10, pp. 18757–18775, 2022, doi: 10.1109/ACCESS.2022.3151186.

[3] J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, "A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications," IEEE Trans. Knowl. Data Eng., vol. 35, no. 4, pp. 3313–3332, Apr. 2023, doi: 10.1109/TKDE.2021.3130191.

[4] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative Adversarial Networks: An Overview," IEEE Signal Process. Mag., vol. 35, no. 1, pp. 53–65, Jan. 2018, doi: 10.1109/MSP.2017.2765202.

[5] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "DF-GAN: A Simple and Effective Baseline for Text-to-Image Synthesis," presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16515–16525. Accessed: Nov. 30, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2022/html/Tao_DF-GAN_A_Simple_and_Effective_Baseline_for_Text-to-Image_Synthesis_CVPR_2022_paper.html

[6] G. Murphy, D. Ching, J. Twomey, and C. Linehan, "Face/Off: Changing the face of movies with deepfakes," PLOS ONE, vol. 18, no. 7, p. e0287503, Jul. 2023, doi: 10.1371/journal.pone.0287503.

[7] teresa_myers, “Deepfake in Movies | Facts, Fiction & Future of Film Industry,” Facia.ai. Accessed: Dec. 03, 2024. [Online]. Available: https://facia.ai/blog/deepfake-in-movies-facts-fiction-future-of-film-industry/

[8] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar, "A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild," in Proceedings of the 28th ACM International Conference on Multimedia, in MM '20. New York, NY, USA: Association for Computing Machinery, Oct. 2020, pp. 484–492. doi: 10.1145/3394171.3413532.

[9] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting World Leaders Against Deep Fakes," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, IEEE, 2019, pp. 38–45. Accessed: Nov. 30, 2024. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2019/html/Media_Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.html

[10] A. Ali, K. F. Khan Ghouri, H. Naseem, T. R. Soomro, W. Mansoor, and A. M. Momani, "Battle of Deep Fakes: Artificial Intelligence Set to Become a Major Threat to the Individual and National Security," in 2022 International Conference on Cyber Resilience (ICCR), Oct. 2022, pp. 1–5. doi: 10.1109/ICCR56254.2022.9995821.

[11] "NSA, U.S. Federal Agencies Advise on Deepfake Threats," National Security Agency/Central Security Service. Accessed: Nov. 30, 2024. [Online]. Available: https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3523329/nsa-us-federal-agencies-advise-on-deepfake-threats/http%3A%2F%2Fwww.nsa.gov%2FPress-Room%2FPress-Releases-Statements%2FPress-Release-View%2FArticle%2F3523329%2Fnsa-us-federal-agencies-advise-on-deepfake-threats%2F

[12] "Deepfake presidents used in Russia-Ukraine war," Mar. 18, 2022. Accessed: Mar. 02, 2025. [Online]. Available: https://www.bbc.com/news/technology-60780142

[13] R. M. O'Sullivan Donie, "A deepfake video of Mark Zuckerberg presents a new challenge for Facebook | CNN Business," CNN. Accessed: Nov. 30, 2024. [Online]. Available: https://www.cnn.com/2019/06/11/tech/zuckerberg-deepfake/index.html

[14] M. Mirza and S. Osindero, "Conditional Generative Adversarial Nets," Nov. 06, 2014, arXiv: 1411.1784. doi: 10.48550/arXiv.1411.1784.

[15] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," Mar. 29, 2019, arXiv: 1812.04948. doi: 10.48550/arXiv.1812.04948.

[16] H. She, Y. Hu, B. Liu, J. Li, and C.-T. Li, "Using Graph Neural Networks to Improve Generalization Capability of the Models for Deepfake Detection," IEEE Trans. Inf. Forensics Secur., vol. 19, pp. 8414–8427, 2024, doi: 10.1109/TIFS.2024.3451356.

[17] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, and R. He, "TALL: Thumbnail Layout for Deepfake Video Detection," in 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France: IEEE, Oct. 2023, pp. 22601–22611. doi: 10.1109/ICCV51070.2023.02071.

[18] M. Salman, I. Tariq, M. Zulfiqar, M. Jalal, S. Aujla, and S. Fatima, "AWARE-NET: adaptive weighted averaging for robust ensemble network in deepfake detection," in International Conference on Energy, Power, Environment, Control and Computing (ICEPECC 2025), Feb. 2025, pp. 526–533. doi: 10.1049/icp.2025.1162.

[19] Y. Yu, X. Liu, R. Ni, S. Yang, Y. Zhao, and A. C. Kot, "PVASS-MDD: Predictive Visual-Audio Alignment Self-Supervision for Multimodal Deepfake Detection," IEEE Trans. Circuits Syst. Video Technol., vol. 34, no. 8, pp. 6926–6936, Aug. 2024, doi: 10.1109/TCSVT.2023.3309899.

[20] S. Muppalla, S. Jia, and S. Lyu, "Integrating Audio-Visual Features For Multimodal Deepfake Detection," in 2023 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA, USA: IEEE, Oct. 2023, pp. 1–5. doi: 10.1109/URTC60662.2023.10534969.

[21] T. Kumar, R. Brennan, A. Mileo, and M. Bendechache, "Image Data Augmentation Approaches: A Comprehensive Survey and Future Directions," IEEE Access, vol. 12, pp. 187536–187571, 2024, doi: 10.1109/ACCESS.2024.3470122.

[22] T. Kumar, A. Mileo, and M. Bendechache, "Saliency-based metric and FaceKeepOriginalAugment: a novel approach for enhancing fairness and Diversity," Multimed. Syst., vol. 31, no. 3, p. 153, Mar. 2025, doi: 10.1007/s00530-025-01741-5.

[23] M. Ye-Bin, N. Hyeon-Woo, W. Choi, N. Kim, S. Kwak, and T.-H. Oh, "SYNAuG: Exploiting synthetic data for data imbalance problems," Pattern Recognit. Lett., vol. 193, pp. 115–121, Jul. 2025, doi: 10.1016/j.patrec.2025.04.013.

[24] M. Javed, Z. Zhang, F. H. Dahri, and T. Kumar, "Enhancing multimodal deepfake detection with local–global feature integration and diffusion models," Signal Image Video Process., vol. 19, no. 5, p. 400, Mar. 2025, doi: 10.1007/s11760-025-03970-7.

[25] H. Wu, X. Wang, R. Wang, J. Xiang, and L. Ren, "Common Forgery Artifact Driven Deepfake Face Detection," in 2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD), Tianjin, China: IEEE, May 2024, pp. 1585–1590. doi: 10.1109/CSCWD61410.2024.10580312.

[26] V. S. Katamneni and A. Rattani, "Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization," in 2024 IEEE International Joint Conference on Biometrics (IJCB), Buffalo, NY, USA: IEEE, Sep. 2024, pp. 1–11. doi: 10.1109/IJCB62174.2024.10744480.

[27] C. Gao, Q. Xu, P. Qiao, K. Xu, X. Qian, and Y. Dou, "Adapter-Based Incremental Learning for Face Forgery Detection," in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of: IEEE, Apr. 2024, pp. 4690–4694. doi: 10.1109/ICASSP48485.2024.10446543.

[28] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu, "F2 Trans: High-Frequency Fine-Grained Transformer for Face Forgery Detection," IEEE Trans. Inf. Forensics Secur., vol. 18, pp. 1039–1051, 2023, doi: 10.1109/TIFS.2022.3233774.

[29] C. Yang, C. Zhu, C. Deng, and Z. Xiao, "ShuffleSR: Image Deepfake Detection Using Shuffle Transformer," in 2023 IEEE 6th International Conference on Knowledge Innovation and Invention (ICKII), Sapporo, Japan: IEEE, Aug. 2023, pp. 639–642. doi: 10.1109/ICKII58656.2023.10332600.

[30] H. Ilyas, A. Javed, and K. M. Malik, "ConvNext-PNet: An interpretable and explainable deep-learning model for deepfakes detection," in 2024 IEEE International Joint Conference on Biometrics (IJCB), Buffalo, NY, USA: IEEE, Sep. 2024, pp. 1–9. doi: 10.1109/IJCB62174.2024.10744484.

[31] P. P. Chandran, S. Sriram, and C. Babu, "Multimodal Forgery Detection in Videos Using a Tri-Network Model," in 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India: IEEE, Mar. 2024, pp. 1507–1514. doi: 10.1109/ICACCS60874.2024.10717044.

[32] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," Mar. 16, 2020, arXiv:1909.12962. doi: 10.48550/arXiv.1909.12962.

[33] A. Jung, aleju/imgaug. (Nov. 30, 2024). Python. Accessed: Nov. 30, 2024. [Online]. Available: https://github.com/aleju/imgaug

[34] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Mach. Learn. PYTHON.

[35] R. Wightman, rwightman/timm. (Nov. 26, 2024). Python. Accessed: Nov. 30, 2024. [Online]. Available: https://github.com/rwightman/timm

[36] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," ACM Comput Surv, vol. 54, no. 10, p. 200:1-200:41, Sep. 2022, doi: 10.1145/3505244.

[37] K. Han et al., "A Survey on Vision Transformer," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 1, pp. 87–110, Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

[38] V. L. L. Thing, "Deepfake Detection with Deep Learning: Convolutional Neural Networks versus Transformers," in 2023 IEEE International Conference on Cyber Security and Resilience (CSR), Venice, Italy: IEEE, Jul. 2023, pp. 246–253. doi: 10.1109/CSR57506.2023.10225004.

[39] K. N. Ramadhani, R. Munir, and N. P. Utama, "Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention," IEEE Access, vol. 12, pp. 8932–8939, 2024, doi: 10.1109/ACCESS.2024.3352890.

[40] D. W. Deressa, P. Lambert, G. Van Wallendael, S. Atnafu, and H. Mareen, "Improved Deepfake Video Detection Using Convolutional Vision Transformer," in 2024 IEEE Gaming, Entertainment, and Media Conference (GEM), Jun. 2024, pp. 1–6. doi: 10.1109/GEM61861.2024.10585593.

[41] S. A. Khan and D.-T. Dang-Nguyen, "Deepfake Detection: Analyzing Model Generalization Across Architectures, Datasets, and Pre-Training Paradigms," IEEE Access, vol. 12, pp. 1880–1908, 2024, doi: 10.1109/ACCESS.2023.3348450.