Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 4, Issue 1, 2026

REAL AND FAKE FACE CLASSIFICATION USING AN ENHANCED MOBILEVIT ARCHITECTURE

Hafza Eman[*1], Nayab Khalid[2], Seemab Karim[3], Anum Aleem[4], Amna Shakeel[5]

[*1,5]*Department of Computer Science, Heavy Industries Taxila Education City (HITEC) University, Taxila Cantt, 47080, Pakistan*
[2]*Department of Computer Science, University of Engineering and Technology (UET) Taxila, UET–HMC Link Road, Taxila, 47050, Pakistan*
[3,4]*Faculty of Computing, Riphah International University Gulberg Green Campus, Block D Islamabad, 44000, Pakistan*

[*1]hafza.eman@hitecuni.edu.pk, [2]nayabkayani333@gmail.com, [3]seemab.karim.724@gmail.com, [4]anum0812@gmail.com, [5]amna.shakeel@hitecuni.edu.pk

DOI: https://doi.org/10.5281/zenodo.18252633

Keywords
Artificial intelligence, Deep learning, Real and Fake Face, Human images, Transfer Learning, Vision Transformer

Article History
Received: 15 November 2025
Accepted: 28 December 2025
Published: 15 January 2026

Corresponding Author: *
Hafza Eman

Abstract

Humans can normally recognize faces, but today's advanced technology and artificial intelligence make it difficult to tell real faces from fake ones. Modern image editing tools and AI techniques can create very realistic fake face images. Because of this, people often struggle to identify whether a face image is real or artificially created. To solve this problem, deep learning techniques are increasingly being used because they provide more accurate and reliable results than human judgment. Although deep learning techniques have been widely explored, Vision Transformer architectures remain underexplored for fake face detection. This paper adopts the MobileViT architecture and enhances it with task-specific modifications to improve fake face detection performance. The proposed approach used the MobileViT architecture, which combines the strengths of convolutional neural networks and Vision Transformers. MobileViT effectively captures both local facial features through convolutional layers and global contextual information through transformer-based attention. This hybrid architecture makes it well suited for fake face detection. Experimental results demonstrate that the proposed MobileViT-based model outperforms baseline models. It achieved a training accuracy of 85.37%, validation accuracy of 83.79% and test accuracy of 83.68%. The study demonstrates that MobileViT architecture significantly improves fake face detection while maintaining computational and memory efficiency. This research has important applications in areas such as identity verification, social media content moderation, cybersecurity, and digital content authentication. Accurate detection of fake faces is critical in these domains, and the proposed MobileViT-based approach provides an effective and reliable solution for distinguishing real and manipulated facial images.

Introduction

Digitally altered images and videos showing people with fake facial expressions have attracted significant public attention and criticism in recent years due to their potential to mislead and manipulate audiences [1]. These manipulated media, commonly known as deepfakes, are artificial intelligence (AI) generated images, audio,

and videos that appear realistic but are not genuine. Deep learning advancements have made creating deepfakes easier and more convincing than ever before. These advancements allow even non-experts to produce content that can deceive human observers. This rapid progress has intensified concerns about the societal impact of synthetic media. These concerns include threats to privacy, trust, and personal security [2]. According to recent surveys, deepfakes now present real-world challenges for both individuals and automated systems that attempts to verify authenticity [3].

Deepfakes pose serious risks, which includes misuse for disinformation, fraud, and non-consensual imagery. One of the earliest known cases of deepfakes occurred in December 2017, when a Reddit user called "Deepfakes" used publicly available AI tools to create fake pornographic videos by replacing real faces with fabricated ones [4]. This incident demonstrated the harmful potential of deepfake technology and foreshadowed later waves of misuse. More recently, lawmakers have moved to address non-consensual deepfake content through legislation such as the Take It Down Act, passed in 2025, which requires online platforms to remove non-consensual intimate imagery within strict timeframes [5].

Deepfaking refers to the use of artificial intelligence to replace a person's face in images or videos with another person's face in a highly realistic way [6]. This type of synthetic media aims to mislead viewers or change the original meaning of the content. Most existing deepfake detection methods depend on feature extraction techniques and machine learning models, which automatically learn important patterns and features from data using advanced neural networks. However, significant challenges remain, such as the rapid improvement of deepfake generation methods, the lack of comprehensive real-world datasets, and the absence of standard benchmarks for evaluating detection systems [7]. Recent surveys emphasize that detection models often struggle when confronted with real or partially manipulated deepfakes outside controlled datasets [8].

Generative Adversarial Networks (GANs) have been central to producing realistic fake media. A GAN consists of a generator that synthesizes fake images and a discriminator that attempts to distinguish real from fake [9]. While GANs have enabled the creation of highly convincing media, humans often find it difficult to detect such content without specialized tools [10]. Consequently, the development of reliable deepfake detection systems remains a critical research challenge. Existing deep learning-based detectors achieve high accuracy in controlled settings, but their performance often drops when applied to unseen datasets or sophisticated deepfake variants [11].

Recent research has focused on transformer-based architectures due to their ability to capture long-range dependencies and contextual information in images [12]. One such architecture, Mobile Vision Transformer (MobileViT), combines the local feature extraction capabilities of convolutional neural networks with the global context modeling power of transformers. This hybrid design allows MobileViT to maintain a lightweight structure suitable for resource-constrained environments while preserving high accuracy in visual tasks [13]. Despite its potential, MobileViT remains largely underexplored in the domain of deepfake detection, with only a few studies evaluating its effectiveness for detecting manipulated media [14].

In this work, we propose a MobileViT-based deepfake detection framework that leverages the architecture's ability to capture both fine-grained texture details and global image context. Our approach aims to improve accuracy against a wide variety of deepfake generation techniques and minimize the computational resources. The use of MobileViT not only addresses computational efficiency but also opens a promising direction for deploying deepfake detection models on devices with limited resources, such as mobile phones and edge devices.

Objectives

The main objective of this research are as follows:

1. To design and implement a MobileViT based model for real and fake face classification to capture both local facial

features and global contextual information.
2. To evaluate the performance of the proposed MobileViT architecture against baseline deep learning models to assess its effectiveness in detecting AI-generated and manipulated facial images.
3. To analyze the efficiency and practicality of the MobileViT-based approach for fake face detection.

Literature Review

The concept of face manipulation predates modern digital technologies, with one of the earliest documented cases dating back to 1860, when a portrait of Southern leader John C. Calhoun was altered by replacing his head with that of a U.S. President for propaganda purposes [15]. Early image manipulation relied heavily on manual techniques such as splicing, painting, and copy move operations, often followed by post-processing steps including scaling, rotation, and color adjustments. While these methods required significant skill and effort, they laid the foundation for contemporary manipulation practices. With advancements in computer graphics and machine learning (ML) techniques, image tampering has become increasingly automated and semantically consistent. These advancements have lowered the barrier for creating convincing manipulations and significantly expanding their societal impact [16].

Face-swapping represents a specific and highly impactful form of image and video tampering. The first widely recognized deepfake appeared in 2017, when a Reddit user known as "deepfake" released manipulated celebrity videos created using encoder–decoder architectures [17]. These early methods relied on two autoencoders sharing a latent space which requires extensive training data and substantial computational resources. Despite these limitations, face-swapping rapidly gained popularity and became the foundational deepfake technique. These techniques inspired applications such as FakeApp, FaceSwap, and Deepnude by 2019 [18]. Multimedia manipulation strategies are now commonly categorized into copy move, splicing, deepfake generation, and resampling. These strategies reflect the increasing diversity of attack vectors in digital media [19].

Beyond traditional face swapping, face reenactment techniques such as Face2Face [20] introduced a new paradigm by transferring facial expressions and head movements from a source actor to a target while preserving the target's identity. Face2Face enables real-time facial reconstruction and expression synchronization which results in highly realistic output videos that are difficult to distinguish from authentic content. Unlike simple face replacement, reenactment techniques manipulate subtle facial dynamics which make detection more challenging. This evolution underscores the need for detection methods specifically designed to address facial motion inconsistencies rather than relying solely on global image artifacts [21].

The rapid advancement of deep generative models, particularly Generative Adversarial Networks (GANs) [22] and more recently diffusion models, has dramatically increased the realism of deepfakes. These technologies pose serious threats to digital trust, privacy, and security. International organizations such as UNESCO have identified deepfakes as a major contributor to the "crisis of knowing,". Policy responses, including legislative initiatives such as the TAKE IT DOWN Act [5, 23], further reflect the growing recognition of deepfakes as a societal risk. Surveys published between 2024 and 2025 indicate that modern generative models can now produce images and videos that are nearly indistinguishable from authentic media, even under forensic analysis [6].

Despite significant progress in deep learning based detection, generalization remains one of the most critical challenges. Many detection models achieve high accuracy on constrained benchmark datasets but experience severe performance degradation when applied to unseen datasets or real world content [7]. This issue is exacerbated by diffusion based generative models, which reduce or eliminate many of the visual artifacts traditionally exploited by forensic algorithms. As a result, recent research emphasizes robustness, explainability, and cross-dataset evaluation as

essential components of reliable deepfake detection systems [24].

To address the gap between laboratory benchmarks and real world conditions, new datasets have been introduced. The Deepfake-Eval-2024 [25] benchmark represents a significant advancement by providing a multimodal dataset that reflects how deepfakes are actually circulated online. Such datasets enable more realistic evaluation of detection models and encourage the development of systems that can operate effectively under diverse and uncontrolled conditions.

Early deepfake detection methods focused on handcrafted features and traditional forensic cues such as metadata analysis, Error Level Analysis (ELA), and JPEG compression artifacts. Tools like FotoForensics and MMC employ these techniques, but they are easily bypassed by sophisticated attackers and are ineffective against GAN-generated images [26]. Consequently, researchers shifted toward convolutional neural networks (CNNs), which demonstrated superior performance in capturing texture and frequency-domain artifacts. Tariq et al. [27] pioneered the use of neural networks to detect GAN-generated fake faces by analyzing statistical image components. Subsequent studies, including Wang et al. [26], proposed LBP-Net and ensemble models combining texture-based and deep features.

Several comparative studies have evaluated the effectiveness of popular CNN architectures for fake face detection. Taeb et al. [28] reported that VGG19 achieved the highest accuracy (95%) on the "140K Real and Fake Faces" dataset when combined with data augmentation. Other researchers explored frequency-domain cues, arguing that discriminative information often resides beyond the spatial domain. Kiruthika and Masilamani [29] demonstrated that image quality assessment (IQA) features derived from both spatial and frequency domains can effectively distinguish real and fake faces, even when visual differences are minimal. Similarly, Salman and Abu Naser [30] found ResNet50 to be the most effective architecture after extensive training on large-scale datasets.

Recent research has increasingly adopted transformer-based architectures due to their ability to model long-range dependencies and contextual relationships across facial regions. Attention driven methods have shown improved performance in detecting subtle inconsistencies in high quality manipulations, in face reenactment and diffusion-based deepfakes [10]. However, transformer heavy models often require substantial computational resources which limits their deployment in real world and resource constrained environments.

To address computational constraints, recent studies have explored lightweight detection architectures suitable for edge devices. MobileViT, introduced by Rastegari et al., combines convolutional inductive biases with global attention mechanisms, achieving an effective balance between efficiency and accuracy. While MobileViT [13] has demonstrated strong performance in general vision tasks, its application to deepfake detection remains relatively underexplored. Early investigations suggest that mobile friendly vision transformers could enable scalable, energy efficient detection systems for real time and embedded applications [14].

Ensembling techniques have been proposed to improve robustness and generalization. Silva et al. [31] introduced an explainable hierarchical ensemble of weakly supervised models, demonstrating improved performance across diverse manipulation types. Explainable AI (XAI) approaches are increasingly emphasized to enhance transparency and trust in detection systems such as high stakes applications such as legal and forensic analysis [14, 32].

The existing literature shows that, although deepfake detection methods have improved a lot, their accuracy is still not reliable in real world scenarios. Many models work well on controlled benchmark datasets but fail when tested on new, unseen, or real online content. The rapid progress of generative models, especially diffusion based methods, has made deepfakes more realistic and harder to detect. Also, it reduces the effectiveness of traditional cues and even advanced deep learning models. In addition, many high performing approaches are computationally

expensive and difficult to deploy in practical or resource limited environments. These limitations clearly indicate that current methods are not yet sufficient for robust and dependable deepfake detection. Therefore, the proposed methodology is introduced to address these challenges by improving detection accuracy, generalization, and practical usability in real world conditions.

Materials and Methods

Dataset

In this research study, the dataset used for experimentation is from Kaggle, a widely recognized platform for diverse datasets. The dataset comprises a total of 2,041 images, including 1,081 real images and 960 manipulated or fake images. The dataset is available at https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection. Fig. 1 shows some real face images and fake face images.

Fig 1: Sample Real and Fake face images from dataset.

To enhance the model's training and generalization capabilities, data augmentation techniques were employed. The data augmentation process involves the creation of augmented versions of the images through various transformations, such as rotation, scaling, and flipping. This augmentation not only expands the dataset size but also introduces variability. Data augmentation helps the model in learning diverse features and patterns inherent in both real and manipulated facial images. These augmentation techniques are used to enhance the diversity of the training dataset. By flipping images horizontally and applying random rotations, the model becomes more robust and better able to handle variations in orientation and position. The specific parameters, such as the degree of rotation and probability of flipping, can be adjusted based on the characteristics of the dataset and the desired augmentation level.

Proposed Methodology

The proposed methodology establishes a deepfake detection system by using pre-trained backbone architecture of MobileViT. Image data is initially processed and augmented before being fed into the backbone to extract complementary, high-dimensional features. These feature vectors are then passed to a classification head for the classification of real and fake faces. The model's final performance is validated on a test set. Figure 2 shows the proposed architecture diagram of the proposed model.

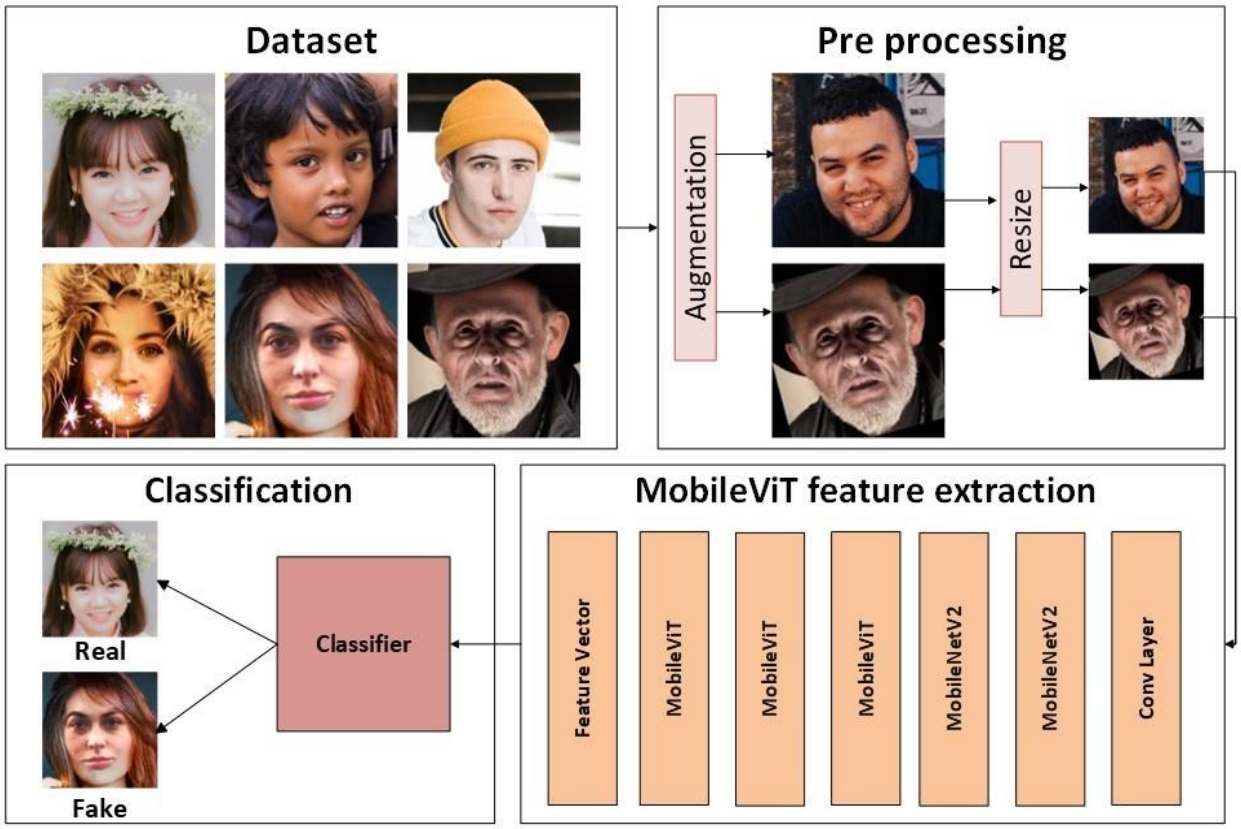

Fig 2. Proposed Architecture Diagram for real and fake face classification.

The proposed framework for real and fake facial image classification is built upon a carefully designed data preparation and augmentation pipeline to ensure robust learning and generalization. Initially, the dataset is organized into two distinct classes, real and fake images. A PyTorch Dataset class is employed to efficiently load and preprocess images in a batch-wise manner. The batch size ensures optimal memory utilization and faster data throughput. During training, data augmentation techniques are applied to increase variability and prevent overfitting. These include random horizontal flips, rotations, affine transformations, as well as brightness and contrast adjustments. Such augmentations simulate diverse real world scenarios, helping the model generalize better to unseen manipulations. For validation and test datasets, a simpler preprocessing approach is adopted, involving only resizing to the standard 224 × 224 input size and normalization, thereby ensuring that evaluation metrics accurately reflect model performance without augmentation-induced bias.

At the core of the framework lies the MobileViT architecture, a hybrid model that combines the strengths of convolutional neural networks (CNNs) with transformer-based self-attention mechanisms. The CNN layers are particularly effective at capturing local spatial features, such as subtle texture inconsistencies and edges, which are often indicative of facial forgeries. The transformer blocks, on the other hand, allow the model to capture long-range dependencies and global contextual information, enabling it to reason across the entire image. This combination ensures that both localized artifacts and broader structural inconsistencies are considered during classification. The model is initialized with ImageNet pre-trained weights, leveraging transfer learning to reduce the dependency on large labeled datasets and accelerate convergence. The original classification head of the MobileViT backbone is replaced with a task-specific multilayer

perceptron (MLP) designed for binary classification. The MLP consists of batch normalization layers, ReLU activation functions, and dropout layers, which collectively enhance regularization, prevent overfitting, and improve the discriminative power of the learned features.

The MobileViT backbone itself is organized into multiple stages, starting with a convolutional stem that reduces the spatial dimensions of the input image from 224 × 224 to 112 × 112 while producing 64 feature channels. Subsequent stages consist of a combination of MobileNetV2 blocks and MobileViT transformer blocks. Notably, stages 0 and 1 use MobileNetV2 blocks to extract 128 and 256 output channels, respectively, gradually reducing the spatial dimensions to 56 × 56 and 28 × 28. Stages 2 through 4 employ MobileViT blocks with progressively higher transformer dimensions (144, 192, and 256) and increasing output channels (512, 1024, and 2048), while reducing the spatial dimensions to 28 × 28, 14 × 14, and 7 × 7. Features from the final stage (features[-1] of stage 4) are used as the representation vector, yielding a 2048-dimensional feature vector before the pooling layer. This rich feature representation forms the foundation for the classifier. Feature representations ensure that both high-level semantic and low-level texture features contribute to distinguishing real and fake faces.

The training follows a two-stage optimization process to ensure stable and effective learning. Initially, all parameters of the MobileViT backbone are frozen, and only the newly added classifier head is trained for the first five epochs using a relatively higher learning rate of 1×10^{-5}. This stage allows the classifier to rapidly adapt to the target task without disrupting the pre-trained representations of the backbone. Following this, a progressive unfreezing strategy is implemented. Unfreezing gradually enables gradient updates for selected deeper layers of the backbone while maintaining a lower learning rate for these parameters. This fine-tuning approach ensures that higher-level features are adjusted for forgery detection without catastrophic forgetting of the general visual representations learned from ImageNet. Optimization is performed using the AdamW optimizer, complemented by a ReduceLROnPlateau learning rate scheduler, which dynamically lowers the learning rate when performance plateaus. The model is trained for a total of 30 epochs with a batch size of 16. These hyperparameters shown in Table 1 ensure a balanced trade-off between computational efficiency and convergence stability.

To further enhance robustness and reduce the model's over-reliance on global image patterns, CutMix regularization is incorporated during training. With a fixed probability for each batch, image regions are exchanged between samples, and their corresponding labels are mixed proportionally using a beta distribution. This encourages the model to focus on localized inconsistencies such as texture anomalies, blending artifacts, or subtle distortions. Combined with dropout in the classifier layers, CutMix significantly improves the generalization performance of the model. This makes the model more resilient to novel or previously unseen forgery methods. This combination of sophisticated regularization techniques ensures that the model develops a fine-grained understanding of facial integrity, rather than memorizing spurious correlations in the training data.

Beyond the core training and augmentation strategies, the architectural choice of MobileViT provides several practical advantages for real world deployment. MobileViT models are lightweight, computationally efficient, and optimized for mobile and edge devices. These characteristics make them suitable for scenarios where rapid and resource constrained inference is necessary. Despite their efficiency, the hybrid CNN-transformer design allows them to capture both local and global features, striking a balance between performance and computational cost. Leveraging pre-trained weights ensures that the network starts from a strong initialization. The network continues by reducing the number of required training epochs and enabling reliable performance even with moderately sized datasets.

Finally, the integration of task-specific modifications, including the classifier MLP with dropout, batch normalization, and ReLU

activations, ensures that the model is fully tuned for binary classification while maintaining generalization. The combination of staged optimization, regularization through CutMix, and progressive fine-tuning allows the system to extract meaningful features from real and fake faces efficiently and accurately. Together, these design choices form a comprehensive, robust framework capable of addressing the increasingly sophisticated nature of facial forgeries, while remaining practical for deployment in real-world applications.

Table 1: Hyperparameters used for the model training.

Parameter	Value
Image Size	224 × 224
Epochs	30
Batch Size	16
Optimizer	AdamW
Learning Rate	1×10^{-5}

Results and Discussion:

To evaluate the effectiveness of the proposed MobileViT-based framework, a series of experiments were conducted to measure its ability to distinguish real and fake facial images. The overall objective of these experiments was to assess the model's learning capacity and its generalization to unseen data to ensure reliability in practical deployment. All experimentation was performed using Google Colab with a T4 GPU, which provided sufficient computational resources to train the model efficiently while allowing the testing of different hyperparameter configurations and regularization techniques without significant time constraints.

The performance of the model is assessed through training, validation, and test accuracies, which collectively provide a comprehensive view of its learning behavior and generalization ability. During training, the model achieves an accuracy of 85.37% which indicates that it successfully learns discriminative features from the training dataset and adapts effectively to the task of detecting facial forgeries. The validation accuracy, measured at 83.79%, demonstrates that the model generalizes well to data not seen during training which suggests that overfitting is limited and the learned features are robust. Finally, the model achieves a test accuracy of 83.68%, confirming its capability to maintain consistent performance when applied to completely unseen images, which is critical for real world applications in detecting manipulated facial content. The progression of training, validation, and testing accuracies is illustrated in Figure 3, highlighting stable convergence and effective adaptation.

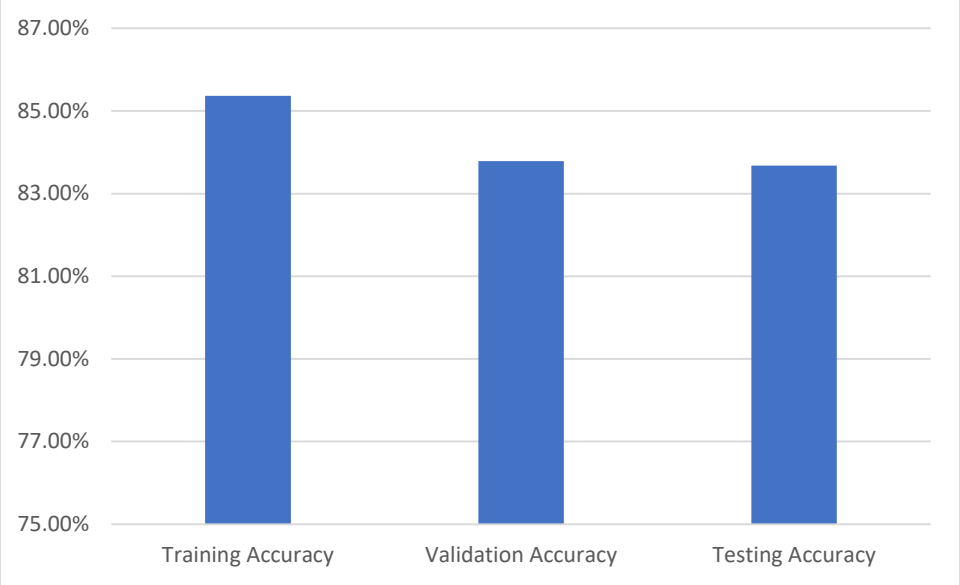

Fig 3: Accuracies achieved for training, validation and testing of proposed model.

In addition to overall accuracy, precision, recall, and F1-score are employed to provide a more nuanced evaluation of the model's class wise performance and its ability to distinguish between real and fake facial images. These metrics offer insight into both the correctness of predictions and the model's sensitivity to each class, which is particularly important in binary classification tasks where imbalances or subtle differences may exist. For the Real class, the model achieves a precision of 0.82 and a recall of 0.84 which indicates that most real images are correctly identified while maintaining a relatively low false-positive rate. This suggests that the model is effective at minimizing the misclassification of fake images as real, which is crucial for applications that require reliable identification of authentic content. For the Fake class, the model attains a precision of 0.85 and a recall of 0.83 which demonstrates robust capability in detecting manipulated or forged images. The slightly higher precision for the Fake class indicates that the model is particularly conservative when labeling an image as fake, reducing the likelihood of incorrectly flagging real images.

The overall balance of the model's performance is further highlighted by the macro-average and weighted F1-score, both of which are 0.84. The macro-average F1-score provides an unweighted evaluation across both classes, reflecting that the model maintains consistent performance irrespective of class distribution. The weighted F1-score accounts for the relative number of samples in each class to ensure that the evaluation is representative even if one class is slightly more prevalent. Together, these metrics demonstrate that the model not only achieves high accuracy but also maintains equitable performance across both real and fake classes. This indicates that it does not favor one class over the other. This balanced performance is critical for real world scenarios, where misclassifying fake images as real or vice versa can have significant implications, such as in security, forensic analysis, and media verification.

These detailed evaluation results, summarized in Table 2, underscore the effectiveness of the MobileViT-based framework in capturing subtle discrepancies between real and manipulated facial images. By using the hybrid convolutional transformer architecture and incorporating strategies such as CutMix regularization, progressive fine-tuning, and robust data augmentation, the model is able to learn rich feature representations that generalize well across different types of facial manipulations. The results indicate that the model can detect forgery artifacts reliably while avoiding over-reliance on superficial cues which makes it suitable for deployment in

practical face verification and forgery detection systems. Overall, the combination of accuracy, precision, recall, and F1-score presents a comprehensive evaluation that confirms the robustness, fairness, and reliability of the proposed framework.

Table 2: Detailed classification results of proposed model.

Metric	Precision	Recall	F1-score
Real	0.82	0.84	0.83
Fake	0.85	0.83	0.84
Macro avg	0.84	0.84	0.84
Weighted avg	0.84	0.84	0.84

The training and validation accuracy curves provide a clear insight into the learning behavior and generalization capability of the proposed MobileViT-based model. As depicted in Figure 4, both training and validation accuracies increase steadily over the course of the 30 training epochs. The validation accuracy closely follows the trajectory of the training accuracy, indicating that the model is consistently improving on unseen data as it learns from the training set. The smooth and gradual rise of these curves demonstrates effective feature extraction and adaptation. The absence of sharp fluctuations or divergences suggests that the regularization strategies such as dropout in the classifier layers, CutMix augmentation, and progressive unfreezing of the backbone successfully prevent overfitting. The accuracy curves highlight the model's ability to maintain stable learning while improving performance on both real and fake facial image classification tasks.

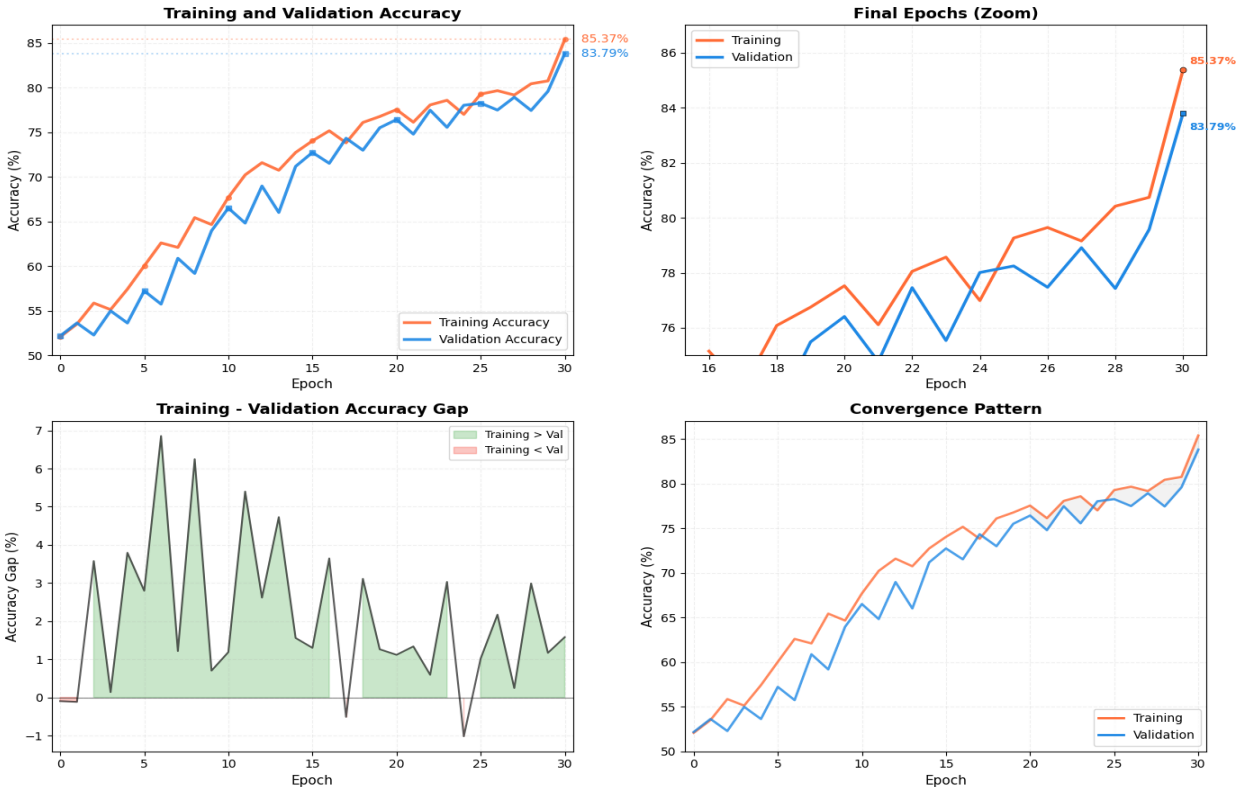

Fig 4: Training and Validation Accuracy curves, Training – Validation Accuracy gap and Convergence Pattern.

The training and validation loss curves, shown in Figure 5, further support these observations by illustrating the reduction in prediction errors over time. Both curves decrease consistently across the epochs, with the validation loss closely tracking the training loss throughout the training process. This steady decline indicates stable convergence and suggests that the model is effectively minimizing the classification error without becoming overly specialized to the training data. The close alignment between training and validation loss reinforces that the applied techniques and including careful learning rate scheduling with ReduceLROnPlateau, staged optimization, and robust data augmentation enable the model to generalize well. These loss curves, together with the accuracy curves, provide a comprehensive understanding of the model's learning dynamics and confirm that it achieves a reliable balance between accurate training performance and strong generalization on unseen samples.

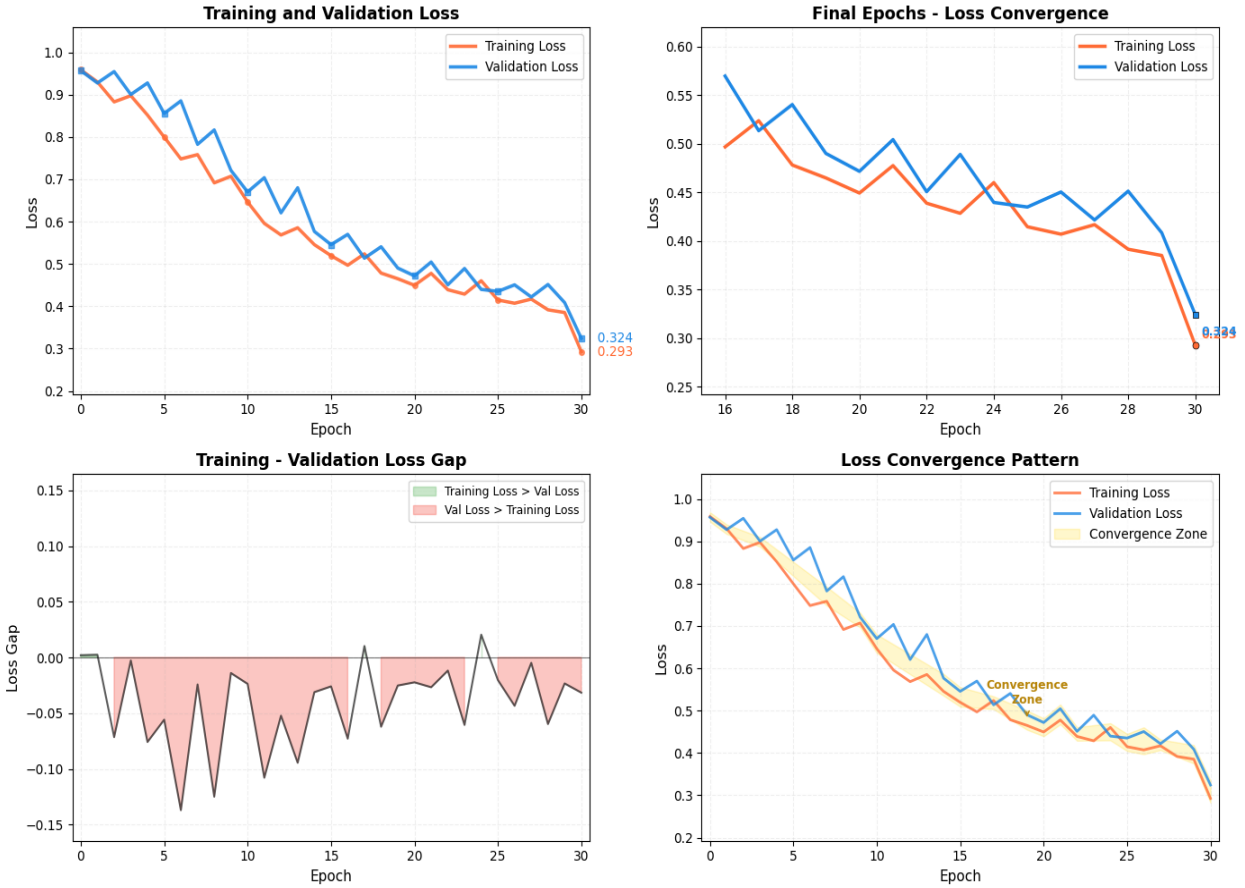

Fig 5: Training and Validation Loss curves, Training – Validation Loss gap and Loss Convergence Pattern.

The confusion matrix provides a detailed view of the model's class-wise prediction behavior, offering insight beyond overall accuracy metrics. As illustrated in Figure 6, the proposed MobileViT-based model correctly identifies 84.25% of real images and 83.15% of fake images, demonstrating a well-balanced performance across both classes. Misclassification rates remain relatively low, with 15.75% of real images incorrectly predicted as fake and 16.85% of fake images misclassified as real. These errors are understandable given the inherent difficulty of deepfake detection, where manipulated images often contain subtle visual artifacts that are

challenging to discern even for state-of-the-art models.

The near-symmetry of the confusion matrix reflects that the model does not favor one class over the other, indicating minimal class bias and consistent discriminative capability. Such balanced performance is particularly important for real-world applications, where both false positives (misclassifying real faces as fake) and false negatives (failing to detect manipulated images) can have significant consequences. By correctly capturing this equilibrium, the model demonstrates its ability to generalize effectively across diverse samples while maintaining sensitivity to the nuanced differences between authentic and manipulated facial features. The confusion matrix reinforces the reliability and fairness of the proposed framework for binary face forgery detection tasks.

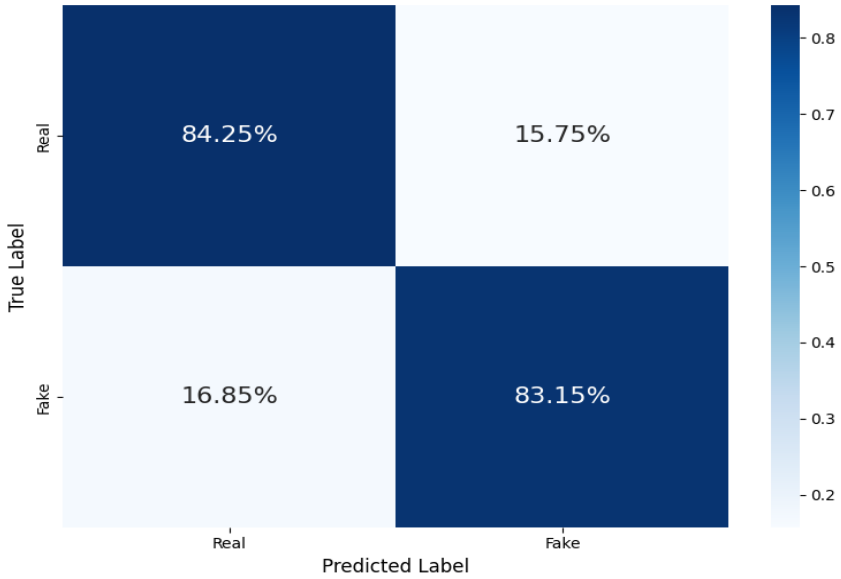

Fig 6: Confusion matrix representing model performance.

Table 3 presents a comparative analysis of the proposed model against several state-of-the-art deepfake detection approaches reported in recent studies. Earlier CNN based models such as VGG16 and ResNet50 reported accuracies of 62.60% and 72.63%. These accuracies shows moderate performance. The integration of GAN with ResNet50 improved accuracy to 82.98% which highlights the benefit of advanced data augmentation techniques. More recent methods, such as Swin Transformer and open source deepfake detectors, achieved accuracies of 71.29% and 69.00%, respectively, but still lag behind the top-performing models. In comparison, the proposed MobileViT-based model achieves the highest accuracy of 83.68%. The proposed model outperformed all the compared approaches. This demonstrates that the proposed method provides better detection capability while maintaining efficiency which makes it a strong alternative to existing deepfake detection models.

Table 3: Comparative Analysis of Proposed model with state-of-the-art models.

Citation	Year	Technique / Model	Accuracy (%)
[33]	2024	VGG16	62.60
[33]	2024	ResNet50	72.63
[33]	2024	ResNet50+GAN	82.98
[34]	2025	Swin Transformer	71.29

[25]	2025	Open source deepfake detectors	69.00
Ours	**2026**	**MobileViT**	**83.68**

Despite the strong performance of the proposed MobileViT-based framework, several limitations should be acknowledged. First, the model is trained and evaluated on static images, which may limit its ability to fully capture temporal inconsistencies or subtle motion artifacts that are present in video-based deepfake detection scenarios. Second, although extensive data augmentation and regularization techniques were employed, the framework may still struggle with highly sophisticated manipulations or adversarially generated forgeries that closely mimic real facial textures. Third, the experiments were conducted using a single GPU platform (Google Colab T4), and while the model is lightweight and efficient, scaling to very large datasets or real-time applications may require further optimization or deployment strategies. Finally, the current framework relies solely on visual information, potentially overlooking complementary information such as audio inconsistencies or multimodal correlations, which could further enhance detection robustness in practical applications.

The experimental results demonstrate that the proposed MobileViT based framework achieves reliable and consistent performance for real versus fake face detection. By using the combination of a lightweight hybrid transformer convolution architecture, transfer learning from pre trained weights, progressive fine tuning, and robust regularization techniques such as dropout and CutMix, the model effectively discriminates between authentic and manipulated facial content. These findings confirm that MobileViT, despite being relatively underexplored in deepfake detection, is a promising and efficient backbone for this task. The framework not only achieves strong performance on standard evaluation metrics but also provides a flexible foundation for future improvements. These improvements include multimodal integration, temporal modeling for video deepfakes, or further architectural refinements to enhance robustness against increasingly sophisticated manipulations.

Conclusion:

This paper presented a MobileViT-based deep learning framework for real and fake face detection. By using the hybrid architecture of MobileViT, which combines convolutional feature extraction with transformer-based global context modeling, the proposed approach effectively captures both local and global features present in manipulated facial images. Experimental results demonstrate that the model achieves strong and consistent performance, with training, validation, and test accuracies exceeding 83%, along with balanced precision, recall, and F1-scores across both classes. These findings confirm that MobileViT, despite being relatively underexplored in deepfake detection research, is a capable and efficient backbone for this task. The results further highlight the effectiveness of the proposed training strategy, including transfer learning, progressive unfreezing, CutMix regularization, and adaptive learning rate scheduling. Together, these techniques contribute to stable convergence and improved generalization, as shown in Figure 7 which clearly shows higher accuracy when compared with state-of-the-art methodologies. Despite these promising results, certain limitations remain. The proposed approach focuses solely on image-based deepfake detection and does not consider temporal inconsistencies present in video data. Also, while MobileViT is computationally efficient, further optimization may be required for deployment on ultra-low-power devices. Future work will explore extending the proposed framework to video-based and multimodal deepfake detection by incorporating temporal information and audio-visual cues. Evaluating the model on larger and more diverse datasets, as well as integrating explainable AI techniques to improve interpretability, are also promising directions. These enhancements aim to further improve

robustness, transparency, and real-world applicability of MobileViT-based deepfake detection systems.

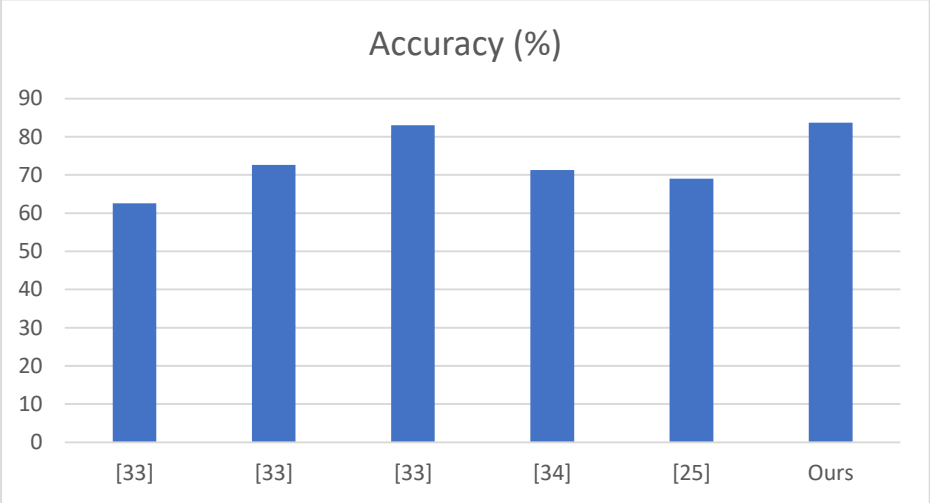

Fig 7: Comparison of proposed model with existing state-of-the-art methods.

References

Deepfakes and the crisis of knowing. 2025; Available from: https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing.

Huang, K. AI Deepfake Security Concerns. 2024; Available from: https://cloudsecurityalliance.org/blog/2024/06/25/ai-deepfake-security-concerns.

Luan, T.A.O., A Survey on Deepfake Detection Technologies. International Journal of Emerging Technologies and Advanced Applications, 2025. 2: p. 1-9.

Regan, G. A Brief History of Deepfakes. 2025; Available from: https://www.realitydefender.com/insights/history-of-deepfakes.

TAKE IT DOWN Act. 2025; Available from: https://en.wikipedia.org/wiki/TAKE_IT_DOWN_Act.

Gong, L.Y. and X.J. Li, A Contemporary Survey on Deepfake Detection: Datasets, Algorithms, and Challenges. Electronics, 2024. 13(3): p. 585.

Kaur, A., et al., Deepfake video detection: challenges and opportunities. Artificial Intelligence Review, 2024. 57(6): p. 159.

Nuria Alina Chandra, R.M., Lin Qiu, Arnab Karmakar, Hannah Lee, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Sejin Paik, Changyeon Lee, Jongwook Choi, Aerin Kim, Oren Etzioni, Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024. 2025.

Liu, B., et al., A Review of Deepfake and Its Detection: From Generative Adversarial Networks to Diffusion Models. International Journal of Intelligent Systems, 2025. 2025(1): p. 9987535.

Lal, K., S. Shiwani, and G.C. Gandhi, Deepfake video deception detection using visual attention-based method. Scientific Reports, 2025. 15(1): p. 40089.

Sen, M.P., M.A. Porwal, and M.V. Shrivastava, Deepfake Detection Techniques: A Comparative Study.

Singh, S. and A. Dhumane, Unmasking digital deceptions: An integrative review of deepfake detection, multimedia forensics, and cybersecurity challenges. MethodsX, 2025. 15: p. 103632.

Rastegari, S.M.a.M., MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. 2022.

Abdulrahman, A.A.M., Deepfake Image Detection Using Explainable AI and Deep Learning. 2025: Rochester Institute of Technology.

Artaius, J. Abraham Lincoln vs John Calhoun: the original deepfake photo of a US president. 2024; Available from: https://www.digitalcameraworld.com/features/abraham-lincoln-vs-john-calhoun-the-original-deepfake-photo.

Anderson, G., Image Manipulation. 2020: City University of New York John Jay College of Criminal Justice.

Tolosana, R., et al., Deepfakes and beyond: A survey of face manipulation and fake detection. Information Fusion, 2020. 64: p. 131-148.

Chesney, B. and D. Citron, Deep fakes: A looming challenge for privacy, democracy, and national security. Calif. L. Rev., 2019. 107: p. 1753.

Ferreira, S., M. Antunes, and M.E. Correia, A dataset of photos and videos for digital forensics analysis using machine learning processing. Data, 2021. 6(8): p. 87.

Thies, J., et al. Face2face: Real-time face capture and reenactment of rgb videos. in Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

Sabir, E., et al., Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI), 2019. 3(1): p. 80-87.

Huang, K., et al., Generative AI Security. Future of Business and Finance, 2024.

Congress., U.S., TAKE IT DOWN Act: Tools to Address Known Exploitation by Immobilizing Technological Deepfakes on Websites and Networks Act. 118th Congress. 2024.

Cirillo, L., A. Gervasio, and I. Amerini, Explainability-driven adversarial robustness assessment for generalized deepfake detectors. EURASIP Journal on Information Security, 2025. 2025(1): p. 23.

Chandra, N.A., et al., Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. arXiv preprint arXiv:2503.02857, 2025.

Wang, Y., V. Zarghami, and S. Cui. Fake face detection using local binary pattern and ensemble modeling. in 2021 IEEE International Conference on Image Processing (ICIP). 2021. IEEE.

Tariq, S., et al. Detecting both machine and human created fake face images in the wild. in Proceedings of the 2nd international workshop on multimedia privacy and security. 2018.

Taeb, M. and H. Chi, Comparison of deepfake detection techniques through deep learning. Journal of Cybersecurity and Privacy, 2022. 2(1): p. 89-106.

Kiruthika, S. and V. Masilamani, Image quality assessment based fake face detection. Multimed. Tools Appl, 2023. 82: p. 8691-8708.

Salman, F.M. and S.S. Abu-Naser, Classification of real and fake human faces using deep learning. 2022.

Silva, S.H., et al., Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. Forensic Science International: Synergy, 2022. 4: p. 100217.

Al Barsh, Y.I., et al., MPG prediction using artificial neural network. International Journal of Academic Information Systems Research (IJAISR), 2020. 4(11).

Safwat, S., et al., Hybrid Deep Learning Model Based on GAN and RESNET for Detecting Fake Faces. IEEE Access, 2024. 12: p. 86391-86402.

Aprille J. Xi, E.C., Classifying Deepfakes Using Swin Transformers. 2025.