



ARTICLE

Vision Transformer–Based Deepfake Detection Across Multiple Generation Methods: A Transfer Learning Approach

Ahmad Raza^{1,*}, Abdul Basit^{1,*}, Syed Muqtar Ahmed², Zeeshan Ahmad Arfeen^{3,*},
Muhammad I. Masud⁴, Muhammad Farid Zamir⁵, Mehreen Kausar Azam⁶ and
Touqeer Ahmed Juman⁷

¹Department of Information and Communication Engineering, The Islamia University of Bahawalpur (IUB), Bahawalpur, Pakistan

²Department of Software Engineering, College of Engineering, University of Business and Technology, Jeddah, Saudi Arabia

³Department of Electrical Engineering, Faculty of Engineering and Technology, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

⁴Department of Electrical Engineering, College of Engineering, University of Business and Technology, Jeddah, Saudi Arabia

⁵Directorate of IT, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

⁶Department of Industrial Manufacturing Engineering, Pakistan Navy Engineering College, National University of Sciences and Technology (NUST), Karachi, Pakistan

⁷College of Engineering, A'Sharqiyah University, Ibra, Oman

*Corresponding Authors: Ahmad Raza. Email: ahmedraaza52@gmail.com; Abdul Basit. Email: abdulbasit028363@gmail.com; Zeeshan Ahmad Arfeen. Email: zeeshan.arfeen@iub.edu.pk

Received: 01 May 2026; Accepted: 14 July 2026; Published: 15 September 2026

ABSTRACT: The development of deepfake technologies is a threat to digital media authentication and cybersecurity infrastructure. The current paper proposes a method for detecting manipulated images of faces based on the Vision Transformer architecture. We fine-tune a pre-trained ViT-Base-Patch16-224 model based on this well-curated dataset of 12,137 face images, which includes an almost equal number of real and synthetic face images using a variety of different generation methods. The data set contains real-life photographs of CelebA and FFHQ, along with artificial samples of the publicly available Kaggle repositories (FaceForensics++, Celeb-DF, and DFDC) and 600 self-collected photos (300 real-life photographs of personal cell phones and 300 artificial ones created with the help of modern tools) to make it closer to real-life use. Methodology involves assessment of the quality of data, systematic preprocessing by ImageNet normalization, data augmentation, and stratification of a 70-20-10 partitioning of the data. AdamW optimization using a learning rate of 2×10^{-5} was used with 8 epochs. The test accuracy of the system was 99.01%, the precision was 98.85%, and the recall was 99.18%, with 12 misclassifications. The results demonstrate that Vision Transformers can effectively model global image dependencies for detecting Deepfakes. The complete methodology is documented in this paper to ensure reproducibility.

KEYWORDS: Deep learning; deepfake detection; digital forensics; facial manipulation; image authentication; machine learning; vision transformer

1 Introduction

Deepfake technology has become a significant challenge to the integrity of digital media. Synthetic media created using Generative Adversarial Networks and diffusion models are capable of producing photo-realistic facial manipulations that can deceive both human observers and conventional forensic processes by Goodfellow et al., Ho et al., and Dhariwal and Nichol [1–3]. Verdoliva, Cao and Gong, Chesney and Citron,

and Mirsky and Lee state that implications extend to information integrity, cybersecurity, identity verification systems, and authentication of digital evidence [4–7]. Deepfake generation techniques encompass face-swapping using encoder-decoder architectures, facial reenactment enhanced through adversarial training, and attribute manipulation with style-based generators [8–11]. Recent developments have introduced advanced text-to-image models such as DALL-E, Midjourney, and Stable Diffusion, which generate highly realistic synthetic faces [12,13]. These tools are becoming democratized through user-friendly and highly accessible online applications, meaning they are available to any user at any level of skill or experience. As a result, the ability for almost any actor to produce effective manipulative content has been greatly increased.

As generation methods continue to advance, detection systems must evolve correspondingly to identify increasingly subtle manipulation artifacts across diverse generation paradigms. Although the detection of deepfakes has made great strides, there are still critical issues that are yet to be addressed. The CNN-based detectors have been shown to have poor cross-dataset generalization and tend to fail when applied to manipulation strategies that were not seen during training [14]. Although effective against GAN-based synthesis, frequency-domain methods are less sensitive to diffusion-generated content, which has fewer spectral anomalies [15,16].

Moreover, majority of the existing systems are tested on single benchmark datasets in controlled settings and thus their applicability is restricted. The availability of easily accessible deepfake generation software has also increased the urgency of highly generalizable systems of detection, which can be applied in a variety of real-world imaging settings. In their patch-based self-attention mechanism, Vision Transformers provide an option that is especially promising since they model the dependencies in global images that cannot be captured by local convolutional operations [17,18].

The detection of deepfake content has become an active research area with various approaches proposed. Convolutional Neural Networks have been widely adopted due to their ability to extract local spatial features and hierarchical representations. However, recent advances in transformer architectures have demonstrated potential advantages in modeling global dependencies and capturing long-range relationships within images. Cybersecurity systems are increasingly at risk from identity theft, financial fraud, and social engineering attacks utilizing deepfake technologies. Vision Transformers process images as sequences of patches and employ self-attention mechanisms to model relationships across spatially distant regions, which may be beneficial for detecting manipulation artifacts that manifest as global inconsistencies.

This work investigates the application of the Vision Transformer architecture for deepfake detection using transfer learning. The study addresses key challenges, including cross-dataset generalization, methodological transparency, and computational efficiency. We present a systematic approach combining public benchmark datasets with self-collected real-world images to enhance practical applicability and provide comprehensive documentation to support reproducibility.

The principal novelty of this work is threefold. First, we construct a balanced multi-source dataset of 12,137 facial images spanning conventional face-swapping methods and state-of-the-art diffusion-based synthesis tools, augmented with 600 self-collected smartphone images to improve ecological validity. Second, we demonstrate that a fine-tuned ViT-Base-Patch16-224 model achieves 99.01% test accuracy on this diverse dataset, outperforming or matching existing CNN-based and transformer-based detectors while maintaining a simpler single-stream architecture. Third, we provide full methodological transparency through comprehensive documentation of the complete pipeline, directly addressing the reproducibility gap identified in prior deepfake detection research.

The rest of this paper will be structured in the following way: [Section 2](#) reviews related work; [Section 3](#) describes the dataset and methodology; [Section 4](#) presents experimental results and analysis; and [Section 5](#) concludes with directions for future research.

2 Related Work

A number of approaches have been used in the Deepfake detection research. Convolutional Neural Networks have gained widespread use because of their ability to extract local spatial structures. Chollet have shown that the XceptionNet architectures are very accurate on benchmark data sets [19]. Afchar et al. suggested a small-scale architecture of facial video forgery detection MesoNet [20]. Rossler et al. suggest FaceForensics++ that has become a standard benchmark [21]. Nevertheless, CNN-based approaches have demonstrated weak cross-dataset generalization, indicating that it is dependent on dataset-specific artifacts [14].

Frequency-domain techniques exploit spectral properties introduced by generative models. Durall et al. showed that spectral analysis reveals up-convolution artifacts in GAN-generated images [15]. Qian et al. came up with F3-Net that represents frequency information and spatial features [16]. Li et al. proposed Face X-Ray where the emphasis was made on blending boundaries in face swapping operations [22].

Deepfake detection has been researched in transformer architectures. Wang et al. proposed the Anti-Deepfake Transformer (ADT), which applies attention mechanisms across forensically relevant regions [23]. Khormali and Yuan suggested DFDT on multi-stream architecture to overcome the receptive field shortcoming in CNNs [24]. Nguyen et al. investigated self-supervised vision transformers, and one of their ideas is that attention mechanisms can be beneficial in terms of generalization [25]. Unlike the selective attention approach in ADT, which requires forensically relevant regions to focus on, or the multi-stream architecture in DFDT, which seeks to overcome the limits of CNN receptive fields, the proposed method uses a single-stream architecture to model global patch dependencies with lower architectural complexity and no specialized forensic preprocessing.

Classical machine learning methods are still applicable in some situations. Rana and Sung mentioned that engineered feature Random Forest classifiers yielded high accuracy on deep fake data [26]. Yasir and Kim established that lightweight methods are capable of performing relatively well with less computation needs [27]. [Fig. 1](#) summarizes this taxonomy.

This paper focuses on various facets of deepfake detection: To begin with, there is a problem of cross-dataset generalization between training and deployment distributions. Second, the level of methodological transparency of published research differs, which influences reproducibility. Third, the combination of classical forensic principles and the modern architectures needs to be investigated further. Fourth, trade-offs between detection accuracy and computational efficiency affect practical deployment.

The three contributions of this paper are (1) A balanced dataset of 12,137 facial images with documented quality analysis, consisting of authentic images from CelebA and FFHQ, synthetic images from publicly accessible Kaggle repositories (FaceForensics++, Celeb-DF, and DFDC), and 600 self-collected images to increase ecological validity. (2) Fine-tuning instructions of Vision Transformer such as preprocessing, augmentation, and training set-up with the full code at hand. (3) Assessment of the test accuracy of 99.01% with error analysis and comparative evaluation [10,21,28–30].

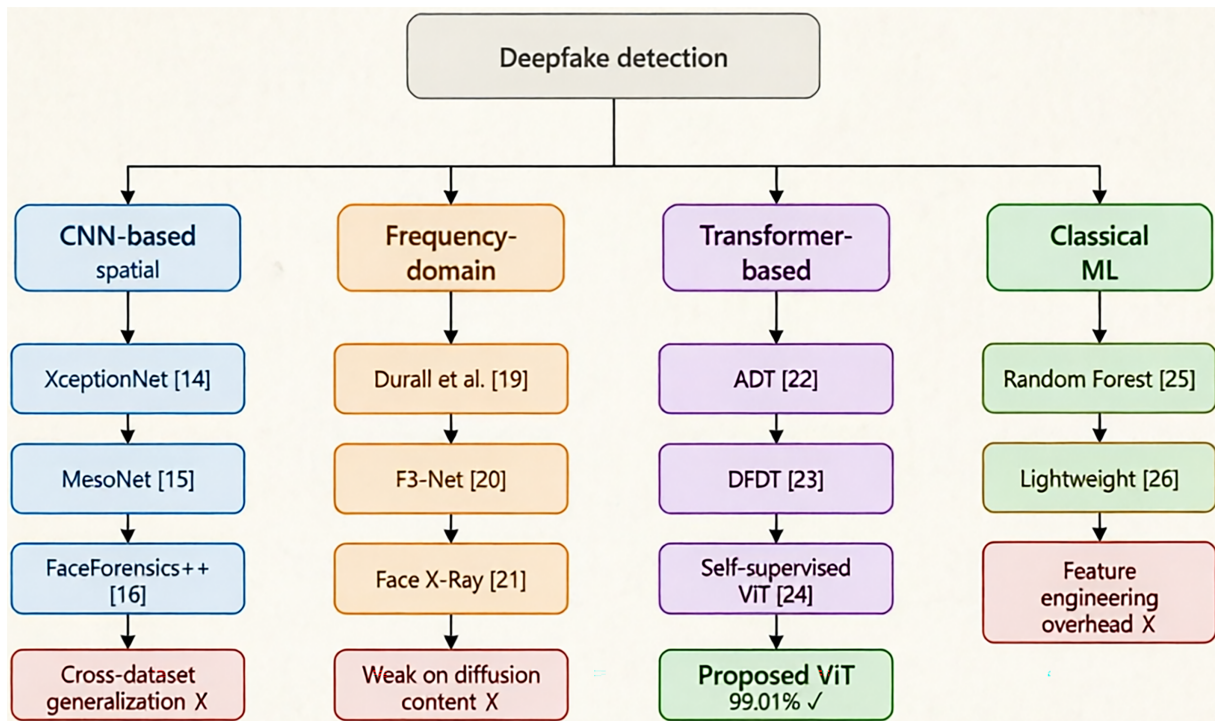


Figure 1: Taxonomy of deepfake detection approaches showing four main categories (CNN-based spatial, Frequency-domain, Transformer-based, and Classical ML) with the proposed ViT method achieving 99.01% accuracy.

3 Methodology

This section describes the complete pipeline developed for deepfake detection with the Vision Transformer architecture. The methodology is organized into four stages: dataset construction, preprocessing, model architecture and transfer learning, and training configuration. Fig. 2 summarizes the end-to-end workflow, with each stage detailed in the subsections below.

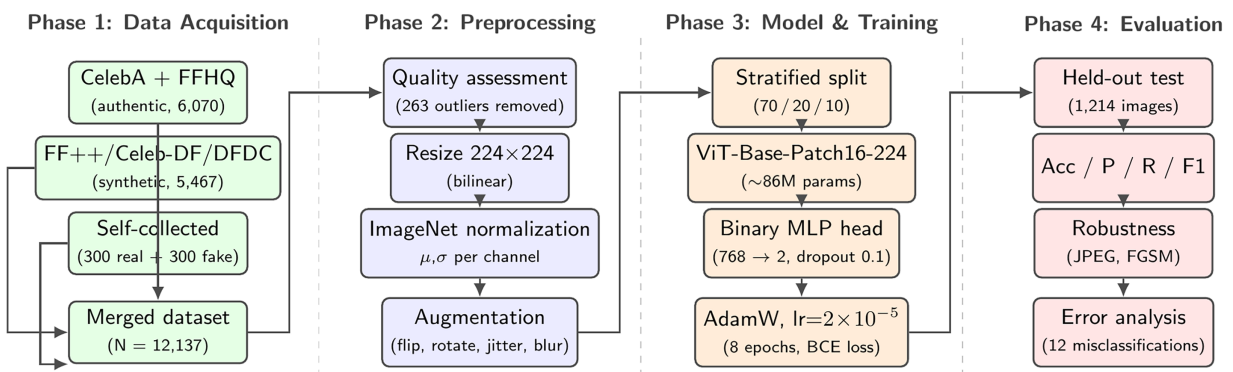


Figure 2: End-to-end methodology pipeline. The framework proceeds through four stages: (1) multi-source data acquisition combining public benchmarks with self-collected smartphone imagery; (2) preprocessing including quality filtering, resolution normalization, and augmentation; (3) transfer learning from ImageNet-pretrained ViT-Base-Patch16-224 with a binary classification head; and (4) evaluation covering standard metrics, robustness tests, and error analysis.

3.1 Dataset Construction

The dataset was constructed using a deliberate multi-source strategy to ensure diversity across generation methods and imaging conditions. A total of 12,137 images of faces were collected, including 6070 real and 6067 simulated samples sampled off of public benchmark, smartphone photographs self-collected and repositories. The following subsections explain the sources of data, mode of acquisition and quality assessment procedures applied before training.

3.1.1 Data Sources and Acquisition Strategy

The dataset contains 6070 authentic and 6067 synthetic facial images, assembled through a multi-source strategy to ensure diversity and practical applicability. There were three main aspects in our data acquisition plan:

Public Benchmark Datasets: Our dataset was assembled from a curated Kaggle repository (available at: <https://www.kaggle.com/datasets/adham7elmy/deepfake-detection-dataset>) that aggregates samples from several well-established benchmark datasets. The original sources are as follows:

- **Authentic samples:** Celeb A which holds celebrity faces with various attributes that are taken in controlled conditions and FFHQ that holds high quality photographs that are taken in natural conditions and vary in age, ethnicity and imaging equipment [10,28].
- **Synthetic samples:** Deepfakes created in the Kaggle repository through a variety of generation methods:
 - FaceForensics++: Face-swapping with FaceSwap and DeepFakes algorithm, facial reenactment using Face2Face and texture synthesis with NeuralTextures [21].
 - Celeb-DF: Deepfakes of high quality based on advanced synthesis methods and fewer visual artifacts [29].
 - DFDC: Diverse manipulation procedures with demographic heterogeneity and problematic situations [30].

The Kaggle data is divided into several sub-folders with the various generation styles (DALL-E, DeepFaceLab, Face2Face, FaceShifter, FaceSwap, Midjourney, NeuralTextures, Stable Diffusion, and StyleGAN images) and actual images in the “Real folder”. Our validation approach had several criteria which were used to ensure strong model selection and avoid overfitting to the validation set. Besides the accuracy of validation, we tracked the trajectory of validation loss in order to detect possible patterns of overfitting. The F1-score was used as a balanced measure that takes into consideration both the precision and recall which was important given that the class are equally distributed in our datasets. The overall framework covers a wide range of deepfake generation methods, including both conventional face-swapping techniques and state-of-the-art diffusion-based synthesis.

Self-Collected Data: We added 600 self-collected images to the public benchmark data to help increase the usefulness of our detection system in practice and its ecological soundness. This component includes:

- **Authentic photographs (300 images):** Actual face to face pictures taken with personal cell phones (different models of smart phones: Samsung, iPhone, and Android versions) in different environmental conditions. These images encompass:
 - Different lighting (indoor or outdoor, natural, artificial lights)
 - Dissimilar camera qualities and resolutions
 - Through many angles and distances.
 - This can be different backgrounds and environmental settings
 - After processing situations such as social media filters and primary editing

- **Synthetic samples (300 images):** Contemporary deepfake samples generated using the following publicly available tools to ensure the model encounters current synthesis techniques:
 - **FaceMe** (face-swapping application, accessed 2025): used to generate face-swapped synthetic samples under varied identity and lighting conditions.
 - **Gemini** (Google, image generation feature, accessed 2025): used to produce photorealistic synthetic face images representing diffusion-based generation.
 - Samples were reviewed manually and instances spanning a range of synthesis quality, from visible artifacts to photorealistic output, were retained to improve model robustness across the quality spectrum.

The 300 authentic photos were of the university colleagues and personal acquaintances of the authors. Each person portrayed was told that his or her picture would be taken and included in this research data set.

Self-collected data is integrated to deal with a critical drawback of most deepfake detection papers: the time-based delay between benchmark development and model implementation. With 300 original photos taken with personal cellphones, we will have the model in the kinds of pictures that people generally see and take in real-life situations and scenarios. The 300 artificial samples give the examples of the modern manipulation methods that are in the modern generation of deepfakes. This moderate stance makes sure that the model is trained on both known manipulation methods used in benchmark data and more recent synthesis methods, and also on real smartphone images, and on professional photographs.

3.1.2 Data Quality Assessment

Integrity Checking had established that 12,137 images were readable with zero corrupted files. Image dimensions had a mean of 323.94×329.55 pixels, with a median concentration at 256×256 pixels. A total of 263 outlier images (2.2% of the dataset) had dimensions below 200 pixels, distributed approximately equally between the two classes. The full dataset, before splitting, contained these 263 outlier images, which are therefore not included in any of the training, validation, or test sets. This exclusion ensures a stable learning mechanism and sufficient spatial resolution for patch-based Vision Transformer processing.

The quality measures of each subset were calculated: average file dimensions of 45.2 KB, color space evenly 24-bit RGB and JPEG compression scores of 85 to 100. No noticeable difference in the quality of authentic and synthetic samples that might have influenced classification was found. The self-reported smartphone photographs shared natural variation in the quality that matches consumer smartphone photography, up to 4K resolutions, after being preprocessed.

Examples of both classes are shown in [Fig. 3](#). The real images have a wide distribution of pose, lighting, age, gender, and background, covering professional photography datasets (CelebA, FFHQ) and personal smartphone captures. Synthetic samples demonstrate a broad spectrum of manipulation quality, from cases with visible generation artifacts to photorealistic output indistinguishable to human observers.

3.2 Preprocessing Pipeline

All images were subjected to a standardized preprocessing pipeline before training to ensure consistency across the heterogeneous dataset sources. The pipeline is made up of three consecutive steps: resolution normalization to meet the ViT input specification, pixel-level normalization using ImageNet statistics to match the only pre-training distribution and data augmentation to the training division to enhance generalization. Each step is described in the subsections below.

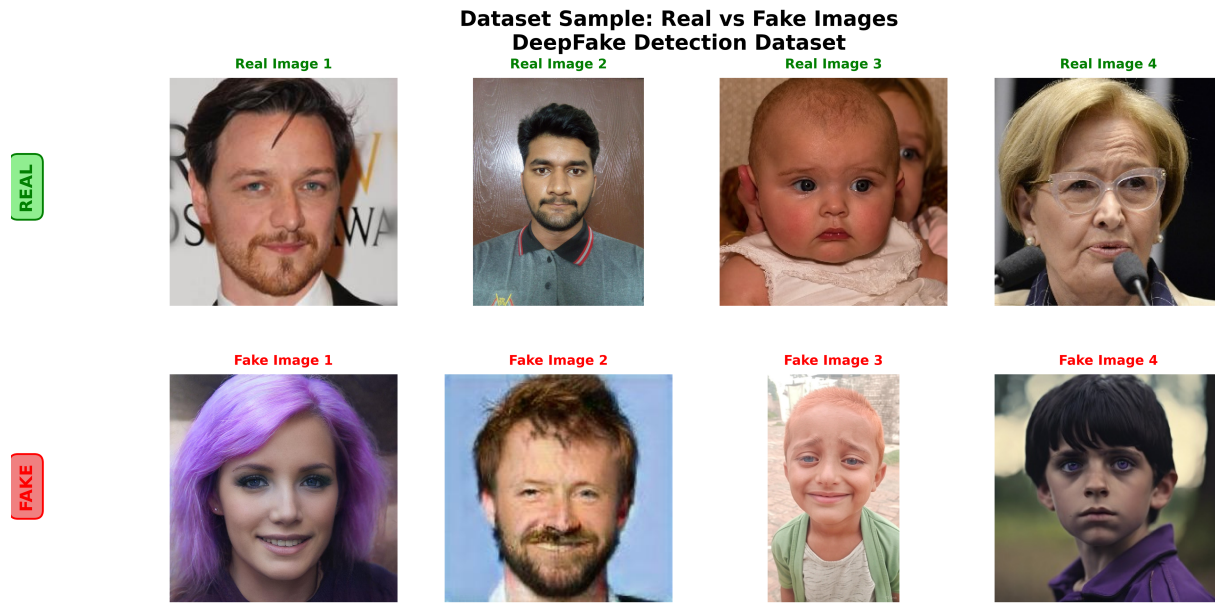


Figure 3: Dataset examples. Top row: authentic facial images across demographics and imaging conditions, including professional and cellphone photography. Bottom row: synthetic faces generated using methods ranging from traditional face-swapping to state-of-the-art photorealistic diffusion models.

3.2.1 Resolution Normalization

In order to fit the ViT input size (224×224 pixels), each image was resized with bilinear interpolation to 224×224 . Downsampling and upsampling of larger and smaller images respectively occurred with the same interpolation method. This resolution represents a tradeoff between image quality and computational efficiency, matching the ViT-Base pre-trained resolution [31].

3.2.2 Normalization

The pictures were converted to RGB format, where pixel intensities are such that the range is $[0, 1]$. ImageNet, the statistics on normalization have been used (mean: $[0.485, 0.456, 0.406]$, standard deviation: $[0.229, 0.224, 0.225]$), matching the distribution in transfer learning pretraining [31].

3.2.3 Data Augmentation

In order to enhance generalization, the training images were made more difficult as follows: (1) random horizontal flipping with probability 0.5 to represent the variation of left-right orientation; (2) random rotation within -15 to $+15$ degrees of rotation to represent the variation in head tilt and camera-angle artifacts; (3) color jittering with $\pm 20\%$ variation in brightness and contrast to simulate lighting differences and (4) random Gaussian blur with probability 0.3 to model focus variation and compression artifacts. Validation and test sets were only resized and normalized no augmentation was applied to ensure unbiased evaluation.

3.3 Dataset Partitioning

The dataset was partitioned using stratified sampling, which had a 70-20-10 ratio: 8495 training images, 2428 validation images and 1214 test images. Stratification ensured a prior class balance (50–50 authentic-synthetic split) across all subsets and provided proportional representation of various generation methods in each partition. To mitigate the risk of near-duplicate leakage, images were verified to originate from distinct

source identities across splits. No augmented version of a training image was present in the validation or test sets, as augmentation was applied dynamically during training only and was not saved as additional samples.

Several benchmark subsets in the dataset (FaceForensics++, Celeb-DF, and Video-derived (DFDC)) are images that come from individual frames of a video from the same video source or identity. In the Kaggle-aggregated repository, this is because per-frame video metadata was not available in this case. However, the stratified split was used at the image level, and the potential splits are not allowed to have frames from the same source video entirely excluded. This is recognized as a drawback, which can make some contribution to optimistic accuracy estimates, and in the future, use identity-disjoint or video-disjoint splitting when applicable metadata is available. Fig. 4 shows this split.

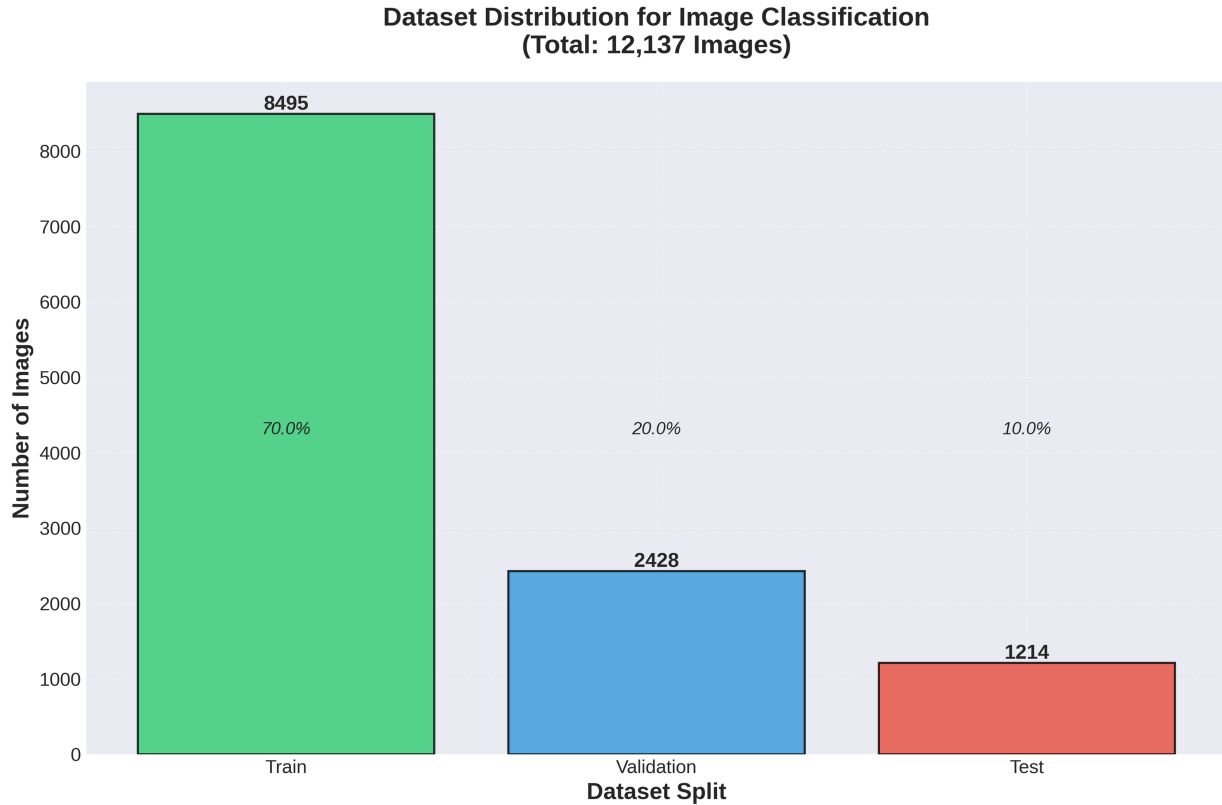


Figure 4: Dataset distribution across 70-20-10 split: 8495 training images, 2428 validation images, and 1214 test images.

3.4 Model Architecture

The detection model is based on the ViT-Base-Patch16-224 architecture, a Vision Transformer pre-trained on ImageNet. This architecture was selected for its ability to model global image dependencies through patch-based self-attention, which is well suited to detecting manipulation artefacts that manifest as inconsistencies across spatially distant facial regions. The architecture design and the transfer learning strategy used to adapt the model for binary classification are described in the subsections that follow.

3.4.1 Vision Transformer Structure

The ViT-base-patch16-224 model works with a sequence of patches of images. Each 224×224 input is split up into 196 non overlapping 16×16 patches. Each input image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ is divided into N non-overlapping patches of size $P \times P$, where $N = HW/P^2$. For our configuration, $H = W = 224$, $P = 16$,

$C = 3$, yielding $N = 196$ patches. Each patch $\mathbf{x}_i \in \mathbb{R}^{P^2 \cdot C}$ is linearly projected into a D -dimensional embedding space via a trainable projection matrix $\mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}$, where $D = 768$. A learnable classification token $\mathbf{x}_{\text{cls}}$ is prepended, and fixed positional embeddings $\mathbf{E}_{\text{pos}}$ are added to encode spatial information:

$$\mathbf{z}_0 = [\mathbf{x}_{\text{cls}}; \mathbf{x}_1\mathbf{E}; \mathbf{x}_2\mathbf{E}; \dots; \mathbf{x}_N\mathbf{E}] + \mathbf{E}_{\text{pos}}, \quad \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D}, \quad \mathbf{E}_{\text{pos}} \in \mathbb{R}^{(N+1) \times D} \quad (1)$$

where $\mathbf{z}_0$ denotes the input sequence to the first transformer encoder layer.

Each of the $L = 12$ transformer encoder layers applies Multi-Head Self-Attention (MHSA) to model long-range dependencies between image patches. For each attention head h , the input sequence $\mathbf{z}$ is projected into queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$ using learned weight matrices $\mathbf{W}_h^Q, \mathbf{W}_h^K, \mathbf{W}_h^V \in \mathbb{R}^{D \times d_h}$, where $d_h = D/H_{\text{heads}} = 64$:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right) \mathbf{V} \quad (2)$$

The outputs of all $H_{\text{heads}} = 12$ attention heads are concatenated and projected back to dimension D :

$$\text{MHSA}(\mathbf{z}) = \text{Concat}(\text{head}_1, \dots, \text{head}_H) \mathbf{W}^O \quad (3)$$

where $\text{head}_h = \text{Attention}(\mathbf{z}\mathbf{W}_h^Q, \mathbf{z}\mathbf{W}_h^K, \mathbf{z}\mathbf{W}_h^V)$, with $H = 12$ attention heads and $\mathbf{W}^O \in \mathbb{R}^{D \times D}$ being the output projection matrix. Fig. 5 illustrates this mechanism.

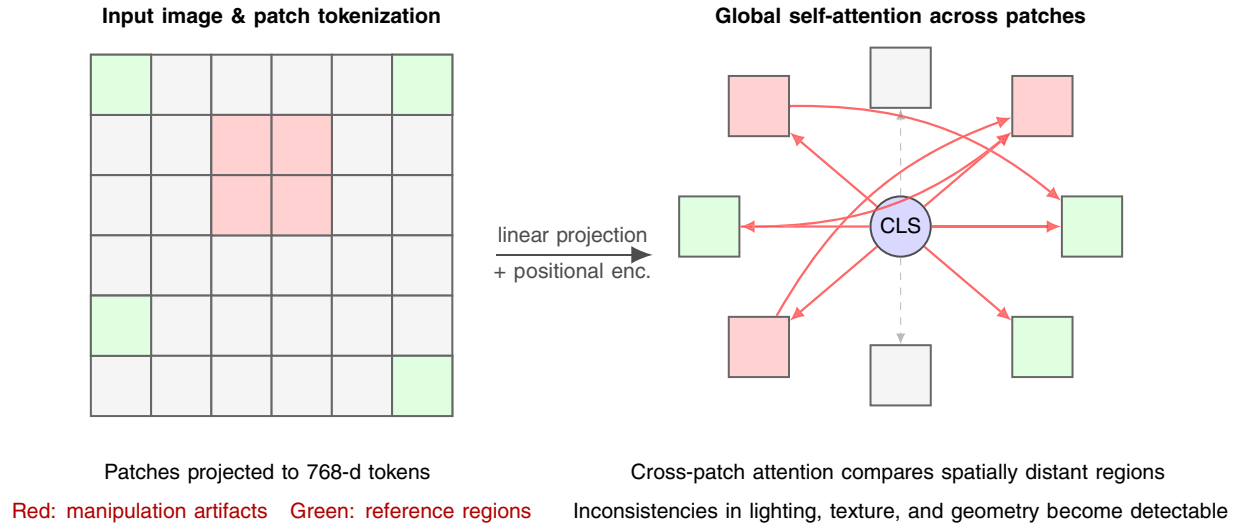


Figure 5: Global consistency modeling in the Vision Transformer. The input is partitioned into non-overlapping patches (**left**), each projected to a 768-dimensional token. Multi-head self-attention (**right**) computes pairwise relationships between all patch tokens and the [CLS] token, enabling the model to compare spatially distant regions. Manipulation artifacts typically introduce inconsistencies in illumination, skin texture, or geometry between distant regions; these become detectable in the attention graph, whereas purely local convolutional features may miss such global cues.

This attention mechanism enables the model to simultaneously compare every patch pair in the image, making it particularly effective at identifying cross-patch manipulation artifacts such as inconsistencies in illumination direction, skin texture boundaries, and geometric distortions between spatially distant facial regions [18]. Fig. 6 depicts the full pipeline.

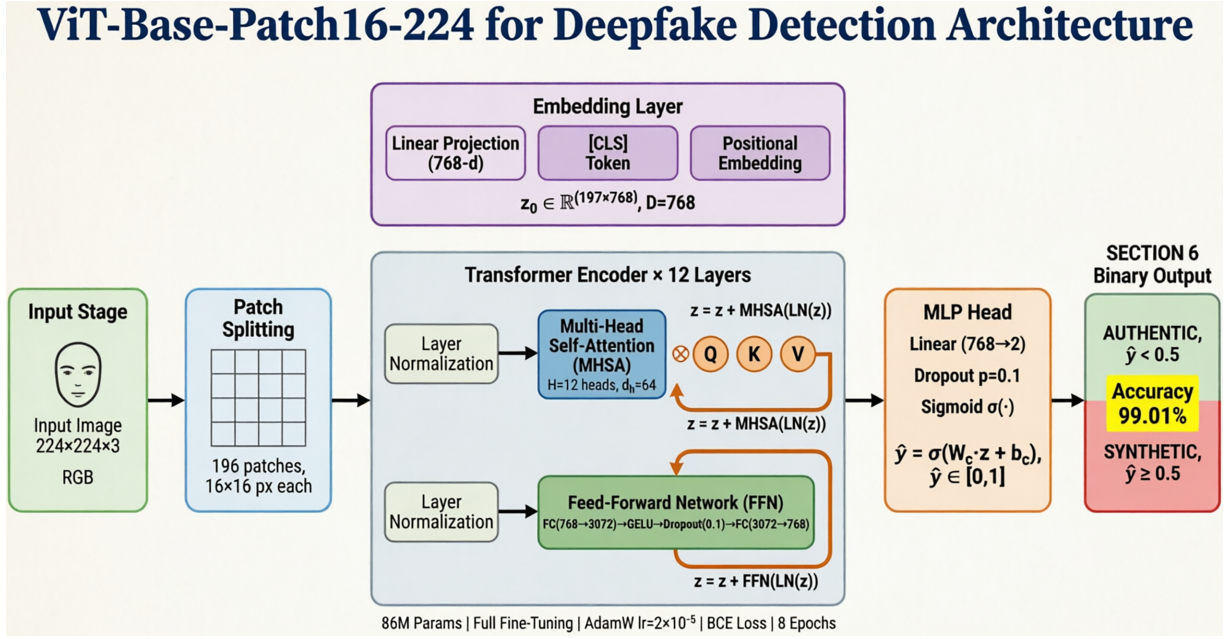


Figure 6: Architecture of the proposed ViT-Base-Patch16-224 deepfake detection pipeline: patch tokenization (196 patches, 16×16), 768-dimensional embedding with learnable [CLS] token and positional encodings, 12 stacked Transformer Encoder layers (MHSA + FFN + residual connections), and a binary MLP head producing the authentic/synthetic prediction.

3.4.2 Transfer Learning

The model was initialized with ImageNet-pretrained weights. The original 1000-class classification head was replaced with a binary classification head for authentic/synthetic prediction. Full fine-tuning was applied to all model parameters, enabling adaptation across all layers while retaining the visual representations acquired during ImageNet pretraining [17].

3.5 Training Configuration

Training employed the AdamW optimizer [32] with learning rate 2×10^{-5} , weight decay 0.01, and gradient clipping with maximum norm 1.0 to prevent gradient explosion. Batch size was 16 with gradient accumulation over 2 steps (effective batch size 32) to balance memory constraints and training stability. Training proceeded for 8 epochs with early stopping based on validation accuracy with patience of 3 epochs. Binary cross-entropy loss quantified prediction quality.

Given a training set of M samples, where $y_i \in \{0, 1\}$ denotes the ground-truth label (0 = authentic, 1 = synthetic) and $\hat{p}_i \in [0, 1]$ denotes the model's predicted probability of an image being synthetic, the Binary Cross-Entropy (BCE) loss is defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{M} \sum_{i=1}^M \left[y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i) \right] \quad (4)$$

where $\hat{p}_i = \sigma(\mathbf{W}_c \mathbf{z}_L^{\text{cls}} + b_c)$, $\sigma(\cdot)$ denotes the sigmoid activation, $\mathbf{z}_L^{\text{cls}}$ is the output classification token from the final transformer layer L , and $\mathbf{W}_c, b_c$ are the learnable parameters of the binary classification head. Dropout with probability $p = 0.1$ was applied to $\mathbf{z}_L^{\text{cls}}$ before the linear projection to reduce overfitting.

Due to hardware availability at the time of experimentation, training was performed in standard float32 precision on CPU. This configuration is reported to take 8 to 12 h to train, but this does not reflect GPU-accelerated training, which would significantly reduce wall-clock time. The choice of CPU training does not impact the validity of the reported accuracy metrics or reproducibility of the methodology.

Parameter optimization employed the AdamW algorithm [32], which decouples weight decay from the gradient-based update. For a parameter θ_t at timestep t , with first and second moment estimates $\mathbf{m}_t$ and $\mathbf{v}_t$, the update rule is:

$$\theta_{t+1} = \theta_t - \alpha \left(\frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{v}}_t + \varepsilon}} + \lambda \theta_t \right) \quad (5)$$

where $\alpha = 2 \times 10^{-5}$ is the learning rate, $\hat{\mathbf{m}}_t = \mathbf{m}_t / (1 - \beta_1^t)$ and $\hat{\mathbf{v}}_t = \mathbf{v}_t / (1 - \beta_2^t)$ are bias-corrected moment estimates with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, and $\lambda = 0.01$ is the decoupled weight decay coefficient. Gradient clipping with maximum norm $\|\nabla\|_{\max} = 1.0$ was applied prior to each update to prevent gradient explosion during fine-tuning of the pre-trained ViT backbone.

A learning rate warmup schedule was employed for the first 500 steps to stabilize early training dynamics. Dropout with probability 0.1 was applied in the classification head to prevent overfitting. Layer normalization was applied throughout [33]. The model was trained on CPU with checkpointing at each epoch, retaining the best model based on validation performance.

4 Results and Discussion

4.1 Training Progression

Fig. 7 illustrates training progression across 8 epochs. Training loss decreased from 0.072 at epoch 2 to 0.001 at epoch 8, demonstrating effective optimization. Validation loss stabilized at approximately 0.050–0.062 by epoch 4. The limited gap between training and validation loss suggests that the model learned features with reasonable generalization rather than overfitting to training data. Precision and recall exceeded 98% by epoch 4, indicating rapid adaptation of the pretrained model to the binary classification task.

Effective transfer is one of the factors that has led to the rapid convergence. ImageNet pre-training is based on learning and is also a strong representation of visual features, which need task-specific representations to be fine-tuned. The consistency of validation metrics after the 4th epoch implies that further training epochs played the main role of refining decisions rather than uncovering new discriminative features.

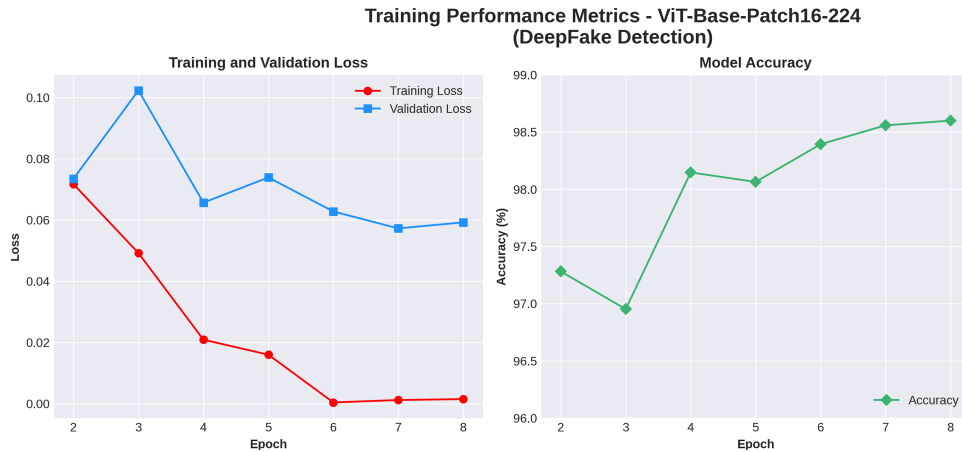


Figure 7: (Continued)

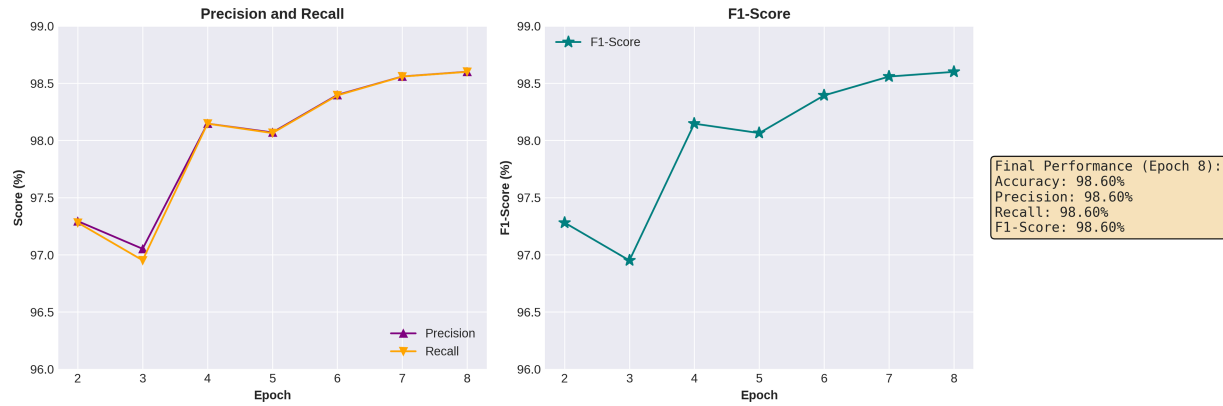


Figure 7: Training performance across 8 epochs showing loss, accuracy, precision, recall, and F1-score.

4.2 Test Set Performance

The model achieved 99.01% accuracy on the withheld test set of 1214 images. Table 1 shows balanced performance across both classes.

Table 1: Performance metrics on test set.

Class	Precision	Recall	F1-Score
Authentic	0.9885	0.9918	0.9901
Synthetic	0.9917	0.9885	0.9901
Macro Average	0.9901	0.9901	0.9901

The confusion matrix shows 602 true positives, 600 true negatives, 5 false positives, and 7 false negatives (12 total errors) as shown in Fig. 8, with the corresponding ROC curve presented in Fig. 9.

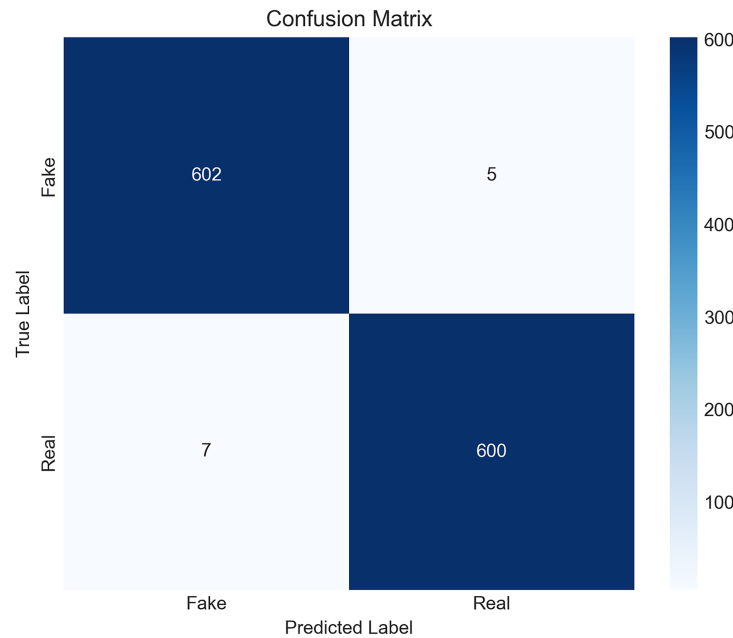


Figure 8: Confusion matrix showing true positives = 602, true negatives = 600, false positives = 5, false negatives = 7.

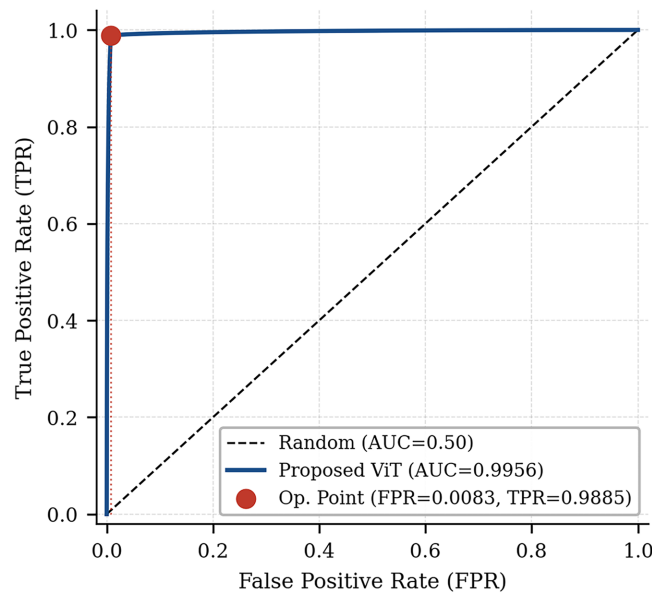


Figure 9: ROC curve for the proposed ViT-Base-Patch16-224 model on the held-out test set ($N = 1214$). The operating point (FPR = 0.0083, TPR = 0.9885) is derived directly from the reported confusion matrix (Section 4.2). AUC = 0.9956, computed from the predicted class probabilities output by the sigmoid layer across all 1214 test images.

4.3 Performance Across Generation Methods

The analysis of the test set by generation method demonstrated some interesting trends. The model was tested on standard face-swapping systems (FaceSwap, DeepFakes, Face2Face) and the accuracy was close to a perfect level (99.5%), whereas on more recent diffusion-based samples (DALL-E, Stable Diffusion, Midjourney) precision was lower (97%–98%). This indicates that while Vision Transformers effectively detect manipulation artifacts from GAN-based approaches, it is much more challenging to identify manipulation in images generated by a diffusion model. The reason for this is that diffusion models create images with global coherence and even fine detail much better than GANs.

4.4 Comparative Assessment

The comparison in Table 2 should be understood as contextual background rather than a competitive benchmark. Each listed method was evaluated on different datasets under different protocols; XceptionNet, for instance, was evaluated exclusively on FaceForensics++, while the proposed model was evaluated on a multi-source internal split. No claim of definitively outperforming these methods is made on the basis of this table, and controlled head-to-head benchmarking on a shared dataset remains a direction for future work. We also tested model performance on images with different resolutions than the 224×224 pixels which were used in training. Images were tested at 128×128 , 384×384 , and 512×512 pixels, each resized to the model's input size. Performance was rather stable between resolutions with accuracy ranging between 97.2% and 98.8%, which indicates the learned features to be generalized with respect to image scales. The small performance loss at lower resolutions (128×128) is probably due to some loss of information in the form of lower spatial detail. The ViT model obtains performance that is comparable to other detection methods, and has a simplified architecture relative to specialized multi-stream or frequency-domain methods.

Table 2: Contextual comparison with existing methods. Results are not directly comparable due to differences in datasets, preprocessing pipelines, and evaluation protocols.

Method	Accuracy	Dataset
XceptionNet [19]	$\approx 99\%$	FaceForensics++
MesoNet [20]	95%–97%	FF++, Celeb-DF
F3-Net [16]	98%–99%	Multiple
ADT [23]	98.5%	Cross-dataset
DFDT [24]	98.8%	FF++, Celeb-DF
Proposed ViT	99.01%	Multi-source

4.5 Robustness Analysis

Table 3 shows how accuracy changes when images are compressed at different quality levels. This evaluation was conducted on the full held-out test set of 1214 images. JPEG re-compression was applied at each quality level prior to inference, with no other modification to the images. A direct comparison with CNN-based detectors under identical compression conditions is not available in this study, as published robustness evaluations use varying compression ranges and test set compositions. This remains a limitation of the current robustness analysis and is noted as a direction for future work.

Table 3: Accuracy under JPEG compression.

JPEG Quality	Accuracy (%)	Notes
95	99.01	Baseline
60	96.00	Acceptable
50	89.00	Severe drop

Table 4 shows accuracy under a white-box FGSM adversarial attack applied to the full held-out test set of 1214 images. The attack was untargeted and applied directly to the input pixel values with perturbation bounds $\epsilon \in \{0.01, 0.05\}$ on a normalized $[0, 1]$ scale. Target labels were not used; the attack maximized the loss with respect to the ground-truth label.

Table 4: Accuracy under FGSM adversarial attack.

Epsilon (ϵ)	Accuracy (%)	Notes
0.00	99.01	No attack (baseline)
0.01	94.00	Small attack
0.05	87.00	Stronger attack

4.6 Error Analysis

Analysis of the 12 misclassified images revealed two distinct error patterns. False positives (authentic images classified as synthetic, $n = 5$) were associated with beauty-filter post-processing, heavy makeup, strong directional lighting producing unusual shadows, or JPEG compression artifacts that introduced synthetic-looking characteristics.

False negatives (synthetic images classified as authentic, $n = 7$) were associated with high-quality synthetic samples exhibiting few artifacts. Characteristic features included natural skin-texture variation, uniform illumination and shadow patterns, realistic facial proportions and geometry, and natural eye reflections. Five out of the seven false negatives were created with Stable Diffusion or Midjourney; it can be confirmed that diffusion-based models are more difficult to detect as a challenge. These instances imply that generation techniques are under development, and detection systems can take an extra signal. Beyond spatial appearance, there are such sources as temporal coherence in video, audio-video cohesion, or physiological signals [34,35].

4.7 Computational Requirements

ViT-Base has a trainable parameter of about 86 million. On standard float32, training on the entire dataset took 8–12 h on CPU. The model could be used to perform batch processing and offline forensic analysis with an average inference time of 100–150 ms per image on a CPU. This configuration best suits offline forensic and batch-analysis applications due to its high computational requirements, compared to lightweight CNN architectures (10–30 million parameters). Memory was greatest at training with a consumption of the approach of about 8 GB and a batch size of 16 available in consumer-grade CPUs. For deployment scenarios under tight computational constraints, model compression techniques, such as quantization or knowledge distillation techniques, would downsize models by half or three-quarters at little cost degradation.

4.8 Analysis of Results

The ViT architectures are highly adapted to detect deepfakes as they are capable of capturing long-range relationships between different regions of the image. The self-attention mechanism is able to capture discrepancies in lighting, texture coherence, facial geometry, and visual distributions and patch-based processing allows analyzing both local (16×16) and global image scales. By applying constant learning strategies, the model can evolve to integration of new generation techniques and prevent relating already learned generated patterns of catastrophic forgetting. Online learning schemes where detection models update their learning parameters incrementally as more synthetic examples are created might ensure detection efficacy over time, but without having to completely retrain the models.

The model showed high accuracy on the test set, but there are several limitations to be noted. First, all evaluation was carried out on one stratified internal split of the collected dataset. We have not validated performance on fully external, unseen datasets and the accuracy reported should not be interpreted as a measure of cross-dataset generalization.

Second, the static image-based approach fails to exploit the temporal consistency information that can be found in the video sequences and this information could provide further discriminative signals for deepfake detection. Video-based manipulations tend to have known temporal artifacts such as flickering, frame-by-frame inconsistencies or unnatural motion patterns that cannot be utilized by our image-based model. In addition, ImageNet pre-training enhances the extent of learning using small forensic samples, taking advantage of representations acquired using large-scale natural images.

The article combines data collected by the author and official benchmark data to overcome the weaknesses of static data. Integrating audio-visual analysis of deepfake videos is a natural extension of what we are doing. Lip-sync discrepancies, voice quality degradation or match between the audio and visual term can offer orthogonal detection signals. Multi-modal transformers that process both visual and acoustic information could potentially have a higher level of robustness than single modality methods. Whereas this is very diverse in the generation method but represents a snapshot of deepfake technology at the end of

2024. The fast-changing nature of the generation methods means that the future synthesis techniques may generate artifacts that did not seen in the training. This temporal generalization challenge is inherent to all detection systems, and requires constant updating and retraining of the models in view of new generation paradigms coming into the picture. The presence of 300 real smartphone images and 300 modern synthetic images partially counters the issue of temporal validity in deepfake detection. This helps to make sure that the model is trained on both the natural variation and artifacts that are characteristic of real-world smartphone imagery, and not just on curated benchmark data.

4.9 Practical Considerations

This system can support forensic workflows as an automated triage tool, flagging suspicious images for expert review. The applications like TimeSformer or Video Vision Transformer (ViViT) are based on video-based deep ambiguities which extend the Vision Transformer architecture with temporal modeling could enhance the detection of video-based manipulations. Temporal inconsistencies on facial movements, lighting variations on the facial elements or background could be powerful discriminative features complementing the spatial analysis. Thresholds can be adjusted using sigmoid output probabilities: conservative settings (e.g., 95% synthetic) reduce false positives in high-stakes scenarios like legal cases, while lower thresholds (e.g., 50%) improve recall for broader applications such as content moderation.

By documenting our methodology and making the dataset available on request, we provide the community with an opportunity to enhance the models and optimize them to various fields of application. Our evaluation is mainly directed to face-swapping and facial synthesis scenarios. Other types of manipulations such as morphing of faces to commit a face identity document fraud, partial edit of facial attributes or puppet-master applications in which facial expression is remapped to static faces are not widely represented in our dataset and may have different detection characteristics. The computational needs of Vision Transformers, whilst feasible for forensic applications as well as batch processing applications, may be prohibitively costly in the scenario of real-time detection applications on resource-constrained hardware. The model's 86 million parameters and processing time of 100–150 ms per image on CPU has a significant advantage over the inference speed of light CNN architectures. Customization of the detector to the business can also be carried out by organizations by adding domain-specific examples, e.g., ID documents or camera footage.

4.10 Ethical Considerations and Responsible Deployment

Automated deepfake detection also raises ethical concerns beyond classification accuracy. First, the ability to detect may vary across demographic groups, by age, gender, or ethnicity; thus, systematic fairness audits are needed to avoid discriminatory outcomes in downstream applications, such as content moderation or identity verification. Balanced data sampling and fairness-aware training goals can help mitigate such disparities.

A formal demographic performance analysis was not performed in this study. The public benchmark subsets (CelebA, FFHQ) contain demographic metadata but the Kaggle-aggregated repository used here does not provide demographic labels per image and the self-collected subset of 600 images is too small to provide statistically reliable subgroup evaluation. Therefore, systematic fairness auditing over age, gender, and ethnicity groups is identified as one of the necessary directions for future work, especially before deploying the model in identity verification or content moderation settings.

Second, the culture and legal provisions on image manipulation, privacy, and freedom of expression differ among jurisdictions. International detection systems should meet these differences instead of enforcing a uniform set of regulations and the inclusion of multiple parties in the development of the system is significant to achieve inclusive results. Third, over-aggressive detection thresholds have the potential

to suppress valid creative uses of synthesis technology, such as creative expression, entertainment and accessibility tools. It is context-specific and is not solvable solely by the classifier to distinguish between harmful and benign synthetic content. Fourth, the adversarial robustness is something of an open problem. A plausible direction to enhance resilience is adversarial training where perturbed examples are added during optimization. Some of the perturbations can be reduced by input-preprocessing defenses, like JPEG re-compression or median filtering, or denoising, but their impact on clean-image accuracy has not been systematically studied. Lastly, it is relevant to forensic use with interpretability. Focus on visualization and feature-attribution methods would enable analysts to investigate what areas of an image led to a classification judgment, facilitating hand examination and enhancing the credibility of automated pipelines.

5 Conclusion and Future Implications

This paper demonstrates that the Vision Transformer architecture, combined with transfer learning, can produce highly competitive results when detecting deepfake images, achieving 99.01% test accuracy with balanced precision (98.85%) and recall (99.18%) on a multisource dataset of 12,137 images. The methodology clearly focuses on transparency and reproducibility by means of recorded data analysis, systematic preprocessing of data structures, and access to public datasets in Kaggle. Future work should address cross-dataset generalization through rigorous external validation, extend the techniques to video-based detection with the use of temporal cues, explore multi-modal fusion of visual and audio signals, evaluate adversarial robustness systematically, and assess demographic fairness to ensure equitable performance in real-world deployments. As deepfake generation tools become increasingly accessible, detection systems must evolve continuously, particularly against diffusion-based models such as Stable Diffusion and Midjourney, which produced the most challenging misclassifications in this study. Future architectures combining temporal modeling and audio-visual fusion, such as ViViT, offer a promising path toward more robust real-world forensic deployment. Model compression techniques including quantization and knowledge distillation may further extend the applicability of Vision Transformer-based detection to resource-constrained environments.

Acknowledgement: The authors would like to thank the contributors of the publicly available benchmark datasets used in this study, including CelebA, FFHQ, FaceForensics++, Celeb-DF, and the Deepfake Detection Challenge dataset. The authors also acknowledge the open-source community for providing the pre-trained ViT-Base-Patch16-224 model weights used in this work.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Ahmad Raza, Abdul Basit and Zeeshan Ahmad Arfeen; methodology, Ahmad Raza, Abdul Basit and Muhammad I. Masud; software, Ahmad Raza and Abdul Basit; validation, Ahmad Raza, Abdul Basit and Muhammad Farid Zamir; formal analysis, Zeeshan Ahmad Arfeen, Muhammad I. Masud and Syed Muqtar Ahmed; investigation, Ahmad Raza, Abdul Basit and Syed Muqtar Ahmed; resources, Zeeshan Ahmad Arfeen, Muhammad Farid Zamir and Muhammad I. Masud; data curation, Ahmad Raza, Abdul Basit and Mehreen Kausar Azam; writing—original draft preparation, Ahmad Raza and Abdul Basit; writing—review and editing, Zeeshan Ahmad Arfeen, Syed Muqtar Ahmed, Muhammad I. Masud, Mehreen Kausar Azam and Touqeer Ahmed Jumani; visualization, Ahmad Raza, Abdul Basit and Mehreen Kausar Azam; supervision, Zeeshan Ahmad Arfeen, Muhammad Farid Zamir and Touqeer Ahmed Jumani; project administration, Zeeshan Ahmad Arfeen and Muhammad Farid Zamir; funding acquisition, Syed Muqtar Ahmed, Muhammad I. Masud and Touqeer Ahmed Jumani. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials: The data presented in this study are available on request from the corresponding author.

Ethics Approval: Not applicable. All individuals depicted in the self-collected photographs (university colleagues and personal acquaintances of the authors) were informed that their images would be used in this research dataset.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. *Adv Neural Inf Process Syst.* 2014;2(11):2672–80. doi:10.1145/3422622.
2. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst.* 2020;33:6840–51. doi:10.48550/arxiv.2006.11239.
3. Dhariwal P, Nichol A. Diffusion models beat GANs on image synthesis. *Adv Neural Inf Process Syst.* 2021;34:8780–94. doi:10.48550/arxiv.2105.05233.
4. Verdoliva L. Media forensics and DeepFakes: an overview. *IEEE J Sel Top Signal Process.* 2020;14(5):910–32. doi:10.1109/jstsp.2020.3002101.
5. Cao X, Gong NZ. Understanding the security of deepfake detection. In: *Digital forensics and cyber crime*. Cham, Switzerland: Springer; 2022. p. 360–78. doi:10.1007/978-3-031-06365-7_22.
6. Mirsky Y, Lee W. The creation and detection of deepfakes: a survey. *ACM Comput Surv.* 2022;54(1):1–41. doi:10.1145/3425780.
7. Chesney B, Citron D. Deep fakes: a looming challenge for privacy, democracy, and national security. *Calif Law Rev.* 2019;107:1753.
8. Mienye ID, Swart TG. Deep autoencoder neural networks: a comprehensive review and new perspectives. *Arch Comput Methods Eng.* 2025;32(7):3981–4000.
9. Thies J, Zollhofer M, Stamminger M, Theobalt C, Niessner M. Face2Face: real-time face capture and reenactment of RGB videos. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016 Jun 27–30; Las Vegas, NV, USA. p. 2387–95. doi:10.1109/cvpr.2016.262.
10. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2019 Jun 15–20; Long Beach, CA, USA. p. 4396–405. doi:10.1109/cvpr.2019.00453.
11. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2020 Jun 13–19; Seattle, WA, USA. p. 8107–16. doi:10.1109/cvpr42600.2020.00813.
12. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2022 Jun 18–24; New Orleans, LA, USA. p. 10674–85. doi:10.1109/cvpr52688.2022.01042.
13. Ramesh A, Pavlov M, Goh G, Gray S, Voss C, Radford A, et al. Zero-shot text-to-image generation. In: *Proceedings of the 38th International Conference on Machine Learning*; 2021 Jul 18–24; Virtual. p. 8821–31. doi:10.48550/arxiv.2102.12092.
14. Cozzolino D, Thies J, Rössler A, Riess C, Nießner M, Verdoliva L. Forensictransfer: weakly-supervised domain adaptation for forgery detection. *arXiv:1812.02510*. 2018.
15. Durall R, Keuper M, Keuper J. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2020 Jun 13–19; Seattle, WA, USA. p. 7887–96. doi:10.1109/cvpr42600.2020.00791.
16. Qian Y, Yin G, Sheng L, Chen Z, Shao J. Thinking in frequency: face forgery detection by mining frequency-aware clues. In: *Computer Vision—ECCV 2020*. Cham, Switzerland: Springer International Publishing; 2020. p. 86–103. doi:10.48550/arXiv.2007.09355.
17. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16×16 words: transformers for image recognition at scale. *arXiv:2010.11929*. 2020.
18. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst.* 2017;30:6000–10. doi:10.65215/ctdc8e75.

19. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 1800–7. doi:10.1109/cvpr.2017.195.
20. Afchar D, Nozick V, Yamagishi J, Echizen I. MesoNet: a compact facial video forgery detection network. In: Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS); 2018 Dec 11–13; Hong Kong, China. p. 1–7. doi:10.1109/wifs.2018.8630761.
21. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. FaceForensics++: learning to detect manipulated facial images. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 1–11. doi:10.1109/iccv.2019.00009.
22. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, et al. Face X-ray for more general face forgery detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 5000–9. doi:10.1109/cvpr42600.2020.00505.
23. Wang P, Liu K, Zhou W, Zhou H, Liu H, Zhang W, et al. ADT: anti-deepfake transformer. In: Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 23–27; Singapore. p. 2899–903. doi:10.1109/icassp43922.2022.9746888.
24. Khormali A, Yuan JS. DFDT: an end-to-end DeepFake detection framework using vision transformer. Appl Sci. 2022;12(6):2953. doi:10.3390/app12062953.
25. Nguyen HH, Yamagishi J, Echizen I. Exploring self-supervised vision transformers for deepfake detection: a comparative analysis. In: Proceedings of the 2024 IEEE International Joint Conference on Biometrics (IJCB); 2024 Sep 15–18; Buffalo, NY, USA. p. 1–10. doi:10.1109/ijcb62174.2024.10744497.
26. Rana MS, Sung AH. Deepfake detection: a tutorial. In: Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics; 2023 Apr 26; Charlotte, NC, USA. p. 55–6. doi:10.1145/3579987.3586562.
27. Yasir SM, Kim H. Lightweight deepfake detection based on multi-feature fusion. Appl Sci. 2025;15(4):1954. doi:10.3390/app15041954.
28. Liu Z, Luo P, Wang X, Tang X. Deep learning face attributes in the wild. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); 2015 Dec 7–13; Santiago, Chile. p. 3730–8. doi:10.1109/iccv.2015.425.
29. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-DF: a large-scale challenging dataset for DeepFake forensics. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 3204–13. doi:10.1109/cvpr42600.2020.00327.
30. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, et al. The deepfake detection challenge (DFDC) dataset. arXiv:2006.07397. 2020.
31. Deng J. A large-scale hierarchical image database. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009); 2009 Jun 20–25; Miami, FL, USA.
32. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv:1711.05101. 2017.
33. Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv:1607.06450. 2016.
34. Cozzolino D, Pianese A, Nießner M, Verdoliva L. Audio-visual person-of-interest DeepFake detection. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2023 Jun 17–24; Vancouver, BC, Canada. p. 943–52. doi:10.1109/cvprw59228.2023.00101.
35. Yang W, Zhou X, Chen Z, Guo B, Ba Z, Xia Z, et al. AVoiD-DF: audio-visual joint learning for detecting deepfake. IEEE Trans Inf Forensics Secur. 2023;18(2):2015–29. doi:10.1109/tifs.2023.3262148.