Detection of AI-Generated Images and Videos Using Vision Transformers

Dinesh M[1], Jeffrin Hannah[2]

[1]Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India
Email: *dineshm22[at]karunya.edu.in*

[2]Division of Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, India
Email: *jeffrinhannah[at]karunya.edu*

Abstract: ***This study presents a deep learning framework for detecting AI-generated images and videos using transformer-based architectures. Image classification is performed using a Vision Transformer (ViT-B/16) model trained on standardized 224 × 224 inputs, while video detection employs a frame-based strategy with a ResNet50 backbone and prediction averaging across uniformly sampled frames. A structured preprocessing pipeline ensures input consistency, and the system is integrated into a secure web-based interface for real-time inference. Experimental evaluation on a binary real-versus-AI-generated dataset achieves 93.28% test accuracy, with precision of 0.9492, recall of 0.9145, and F1-score of 0.9316. The results demonstrate that transformer-based image representations combined with frame-level video aggregation provide an effective approach for reliable detection of AI-generated media.***

Keywords: Deepfake detection; Vision Transformer (ViT); ViT-B/16; ResNet50; AI-generated media; Frame-based video analysis; Multimedia forensics; Transformer-based classification.

1. Introduction

The rapid progress of generative artificial intelligence (AI) has led to the widespread creation of deepfakes.Advances in generative adversarial networks, diffusion models, and transformer-based generators have significantly narrowed the perceptual gap between real and artificial visual content, raising serious concerns related to misinformation, digital identity fraud, media forensics, and societal trust in visual evidence [1]. As AI-generated media becomes increasingly accessible and sophisticated, the development of reliable detection mechanisms has emerged as a critical research challenge.

Early research on AI-generated image detection demonstrated that transformer-based architectures are particularly effective in identifying subtle global inconsistencies introduced during the generation process [2]. Vision Transformers (ViTs), by modeling long- range spatial dependencies through self- attention, have shown improved performance over traditional convolutional neural networks (CNNs) in distinguishing authentic images from synthetic ones [3]. Comparative studies further confirmed that ViT-based detectors provide more stable and interpretable results across diverse datasets, especially when visual artifacts are less localized and more globally distributed [4].

Recent application-specific studies have extended transformer-based detection frameworks to various domains, including design-oriented synthetic imagery, demonstrating the adaptability of Vision Transformers across different visual contexts [5]. In parallel, transformer-based video analysis has gained attention in related multimedia tasks, such as violence detection, highlighting the suitability of attention mechanisms for complex video understanding problems [6]. However, despite these advances, robustness against adversarial perturbations remains a key concern, as attackers can intentionally manipulate synthetic content to evade detection systems [7].

To improve detection reliability, hybrid architectures combining convolutional neural networks with vision transformers have been proposed for both image and video deepfake detection. Such models leverage the local feature extraction strength of CNNs alongside the global reasoning capability of transformers, achieving improved accuracy on benchmark datasets [8]. Enhancements to video vision transformer models, including the integration of facial landmarks, depthwise separable convolutions, and refined self-attention mechanisms, have further improved sensitivity to facial manipulations in deepfake videos [9]. Patch- wise deep learning strategies have also been explored to enable fine-grained detection of multiclass AI-generated facial forgeries [10].

More recent efforts have focused on enhancing transformer-based detectors by emphasizing edge information and high-frequency cues, which are often distorted in synthetic images [11]. Fusion-based approaches that combine CNN features with Vision Transformer representations have shown promising results by capturing complementary spatial characteristics of real and AI-generated images [12]. Comprehensive reviews of deepfake and AI- generated image detection technologies further highlight the rapid evolution of both generation and detection methods, underscoring the need for continuous adaptation of forensic techniques [13].

Beyond images, the detection of synthetic video content presents additional challenges due to temporal coherence and motion consistency. Universal synthetic video detectors have been proposed to handle a wide range of manipulations, from localized facial edits to fully AI-generated scenes [14]. Large-scale benchmarks and datasets containing millions of generated videos have enabled more rigorous evaluation of video detection models and revealed the scalability limitations of existing approaches [15]. Comparative studies between human perception and AI-based detectors have also shown that automated systems can outperform human vision

Paper ID: SR26227161330 DOI: https://dx.doi.org/10.21275/SR26227161330

in identifying subtle generative artifacts under controlled conditions [16].

Several studies have demonstrated that combining efficient CNN backbones with Vision Transformers improves video deepfake detection while maintaining computational efficiency [17]. Improved Vision Transformer variants tailored for deepfake detection have further enhanced performance by refining attention mechanisms and token representations [18]. Unified audio– visual transformer models have also been proposed to detect multimodal deepfakes by jointly analyzing visual and auditory cues [19]. In addition, CLIP-based models have raised the performance ceiling of AI-generated image detection by leveraging large-scale vision– language pretraining [20].

Dataset quality and diversity play a crucial role in training robust detection models. Large-scale and challenging datasets specifically designed for AI-generated image detection have been introduced to better reflect real-world conditions and generalization requirements [21]. Surveys on transformer-based generative and discriminative models emphasize the dual-use nature of transformers, as they power both advanced content generation and effective forensic detection systems [22]. Practical resources and frameworks have further accelerated the adoption of transformer architectures across computer vision and natural language processing tasks [23].

Transformer-based video analysis has also been successfully applied to other safety-critical domains, such as fall detection, demonstrating the broader applicability of attention-driven temporal modeling [24]. Advanced spatial–temporal transformer architectures originally developed for video object detection have influenced the design of modern video deepfake detection pipelines [25].

Vision Transformers have additionally been applied in specialized imaging domains, such as medical ultrasound analysis, reinforcing their generalization capability across diverse visual modalities [26]. Self-supervised and spatio-temporal transformer models have recently emerged as powerful tools for deepfake video detection, enabling effective learning without extensive manual annotation [27]. Despite these advances, open challenges remain, including generalization to unseen manipulation techniques, robustness under real- world distortions, and computational efficiency for deployment [28].

Community-driven evaluation initiatives and challenges have further highlighted the difficulty of assessing AI- generated content quality and detection performance at scale [29]. To address deployment constraints, lightweight and shallow Vision Transformer architectures have been proposed to achieve efficient deepfake detection without sacrificing accuracy [30].

The significance of this study lies in the development of an integrated framework for detecting AI-generated images and videos using transformer-based architectures. Unlike traditional CNN-based approaches, the proposed method leverages the global attention mechanism of Vision Transformers to capture subtle artifacts present in synthetic media. Experimental evaluation and comparative analysis demonstrate the effectiveness of the proposed approach for multimedia forensics and deepfake detection applications.

2. Literature Survey

Vision Transformer Based Image Detection

Transformer-based architectures have been applied in greater numbers to AI generated image detection because of the former's ability to extract global contextual relationships. Lamichhane proves the power of vision transformers to differentiate subtle generative artifacts seen in the synthetic images, or the benefit that the transformers have over the traditional convolutional networks [1]. You can also discusses the application of vision transformer in extracting deep features from the input image and concludes that its multi-head attention mechanism improves the accuracy of classification when encountering artificial intelligence generated content [2]. These works support the use of the transformer-based classification strategy applied to the image detection pipeline used in this project.

Transformer models for Detecting Synthetic Videos

It has been shown that the deepfake detection using videos benefits from temporal modeling and frame-level feature extraction. Battocchio et al. propose the use of video detectors based on transformers that can detect contradiction of information in a temporal manner between frames, which will lead to a more reliable video manipulation detection [3]. Soudy et al. use convolutional networks and transformer modules to maximize temporal and spatial feature representation, achieving good performance on benchmark deepfake data sets [8]. These studies are closely related to the approach taken in the project, in which the program averages predictions over time, with multiple frames in a video.

Performance Comparison and Robustness of Studies.

Park compares AI in generating image detectors in various architectures and emphasizes the importance of trying to visualize model performance to understand classification reliable [4].

De Rosa analyzes the robustness of CLIP based detectors against adversarial perturbations and demonstrates that transformer-driven models need some more attention to robustness when deployed in real world conditions [7]. These insights correspond to the results of the evaluation of this project and the results of the confusion matrix.

ViT uses in Design and Multimedia

Wang uses vision transformer to recognize the existence of AI generated images to interior design, by intention demonstrating the adaptability of vision transformers across domains and confirming that properly optimized vision transformers architectures are capable of recognizing generative artifacts [5]. Alshalawi study transformer-based video analysis for violence detection [6], which is another one of its versatility in multimedia classification [6] tasks as well as understanding the context in detail. These results justify the appropriateness of transformer models applied in this project.

Improvements of Deepfake Detection Models

Ramadhani combines facial landmarks and depthwise separable convolutions in video transformers with facial landmarks, which illustrates improvements in accuracy compared with deepfake detection pipelines [9]. Arshed suggests a patch-wise deep learning model focused on the detection of multiclass deepfakes with a focus on fine-grained feature extraction as a relevant component of the deep learning classifier for better classification results [10]. These advancements show potential for future improvements for frame-based and transformer- based extension of this project.

Recent research on deepfake and AI-generated media detection has increasingly focused on transformer-based architectures due to their ability to capture global contextual relationships. Das *et al.* proposed an edge-enhanced Vision Transformer framework that emphasizes high- frequency and boundary artifacts, achieving improved detection accuracy for AI-generated images with minimal visual distortions [11]. Similarly, Mai *et al.* introduced a hybrid fusion strategy combining convolutional neural networks and Vision Transformers, demonstrating that complementary local and global features significantly enhance classification performance [12].

Comprehensive reviews by Zhang *et al.* highlighted the rapid evolution of generative image technologies and emphasized the growing difficulty of distinguishing synthetic content from real images as generative quality improves [13]. Addressing video-based deepfakes, Kundu *et al.* proposed a universal synthetic video detector capable of handling both localized facial manipulations and fully AI-generated scenes, improving generalization across manipulation types [14]. Chen *et al.* further contributed a million-scale benchmark dataset for AI- generated video detection, enabling large-scale evaluation and exposing scalability challenges in existing models [15].

Comparative analyses between human perception and AI-based detection systems revealed that deep learning models often outperform human observers in identifying subtle generative artifacts under controlled conditions [16]. Hybrid video detection approaches combining EfficientNet and Vision Transformers demonstrated improved robustness while maintaining computational efficiency [17]. Advances in Vision Transformer architectures, including improved attention mechanisms and token representations, have further strengthened deepfake detection accuracy [18]. Multimodal approaches such as AVFakeNet jointly analyze audio and visual cues, enabling effective detection of cross-modal deepfake manipulations [19]. Finally, CLIP-based vision–language models have raised the performance ceiling of AI-generated image detection by leveraging large-scale pretrained representations [20].

3. Proposed Methodology

Dataset Description and Data Characteristics

The performance of any deep learning–based detection system is highly influenced by the quality and characteristics of the dataset used during training and evaluation. In the proposed deepfake detection system, the dataset consists of both real and AI-generated visual media in the form of images and videos. The dataset is designed to support a binary classification task, where each input sample is labeled as either genuine (real) or artificially generated (fake).

The image dataset includes a mixture of authentic images and synthetically generated images produced using modern generative models. These images vary in resolution, lighting conditions, background complexity, and visual texture. Such diversity is important for ensuring that the Vision Transformer model learns generalizable features rather than memorizing specific patterns. Since the uploaded images may originate from different sources, the dataset naturally contains variations in color distribution, noise levels, and compression artifacts. To handle this variability, all images are standardized through a fixed preprocessing pipeline before being passed to the classification model.

The video dataset contains real and AI-generated video clips representing different visual scenarios. Videos may differ in duration, resolution, frame rate, and motion dynamics. To ensure consistency across samples, a fixed number of frames are extracted from each video during preprocessing. In this system, ten frames are uniformly sampled across the video timeline. This sampling strategy helps capture representative visual information while keeping computational requirements manageable. Each extracted frame is treated as an individual image sample and undergoes the same preprocessing steps used for still images.

The dataset follows a binary class structure, where real content represents genuine visual media and fake content represents AI-generated or manipulated media. Although the dataset may contain visual imbalance in terms of artifact visibility, the evaluation results indicate that the trained models are able to distinguish between the two classes with reasonable reliability. The use of standardized preprocessing ensures that differences in image size or format do not affect the learning process.

Overall, the dataset supports the objective of building a practical and deployable deepfake detection system. By including both images and videos with varying visual properties and applying consistent preprocessing, the dataset enables the models to learn discriminative features that generalize well to unseen inputs. This design aligns with real-world deployment scenarios, where uploaded media may originate from diverse and uncontrolled sources.

Table I summarizes the key characteristics of the dataset used for training and evaluation. It highlights the supported media types, classification objective, preprocessing strategy, and video frame sampling approach. This information is critical for understanding the data consistency requirements imposed by the Vision Transformer and ResNet-based models, as well as the practical constraints considered during system deployment

Table 1: Dataset Characteristics Summary

Dataset Attribute	Description
Data types	Images and videos
Classification task	Binary (Real vs AI-generated)
Image resolution	Variable (resized to 224 × 224)
Video processing	10 uniformly sampled frames per video
Preprocessing	Resizing, tensor conversion, normalization
Usage	Training and evaluation of detection models

The system has consistent preprocessing steps for both images and extracted frames of a video. Each input image is re-sized to 224x224 and then it is converted to tensor format. A normalization is applied for mean values [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225], as in the torch vision transformation standards used implemented in the project. For videos, ten equally spaced frames are taken with the help of OpenCV. Each frame is converted from BGR to RGB and converted to a PIL image, leading to a radical reduction of image size to 224x224 pixels, image normalization, and then being stacked into a tensor batch. These preprocessing operations make sure that all inputs being fed to the detection models are in the same standardized format with compatibility of the training setup represented in the model files provided.

Fig. 1 illustrates the proposed deepfake detection methodology. The pipeline begins with media input acquisition, followed by preprocessing operations the training progress. Model metadata files offer test accuracy, precision, recall, and F1 score which can be used to interpret the overall performance. This evaluation process includes operating transparency on the strengths and limitations of classifiers and is completely consistent with the metric outputs provided by the project.

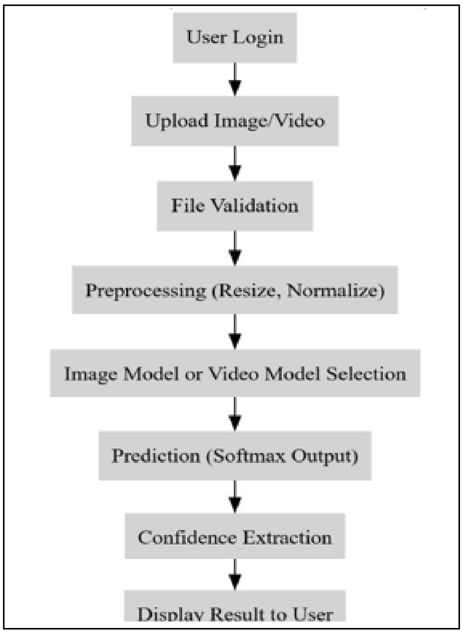
User Login
Upload Image/Video
File Validation
Preprocessing (Resize, Normalize)
Image Model or Video Model Selection
Prediction (Softmax Output)
Confidence Extraction
Display Result to User

Figure 1: Proposed Methodology

Data Preprocessing Strategies

Data preprocessing plays a critical role in ensuring consistent and reliable predictions from deep learning models. The preprocessing pipeline has been carefully designed to match the training conditions of the deployed models and to standardize all inputs before inference.

1) *Image Preprocessing Pipeline*

Each input image uploaded by the user undergoes a series of preprocessing operations before being passed to the Vision Transformer model. These operations ensure dimensional consistency, numerical stability, and compatibility with the pretrained weights.

The preprocessing steps for images include:

- Image resizing to a fixed resolution of 224 ×224 pixels.
- Conversion of image data to tensor format
- Channel-wise normalization using ImageNet mean and standard deviation values
- The normalization parameters used are: Mean vector: [0.485, 0.456, 0.406]
- Standard deviation vector: [0.229, 0.224, 0.225]

Mathematically, the normalization operation for each pixel channel is defined as:

$$x\_normalized = (x - \mu) / \sigma$$

where x represents the original pixel intensity, μ is the channel-wise mean, and σ is the channel- wise standard deviation. This normalization ensures that the input distribution aligns with the distribution used during model pretraining, leading to more stable and accurate predictions.

2) *Video Frame Preprocessing*

For video inputs, preprocessing is applied at the frame level. The system extracts a fixed number of frames from each uploaded video and processes them individually using the same transformations applied to still images.

The video preprocessing workflow consists of:

- Reading the video file using Open CV
- Extracting ten equally spaced frames across the video duration
- Converting frames from BGR to RGB color space
- Converting frames to PIL image format
- Resizing frames to 224 × 224 pixels
- Tensor conversion and normalization

This approach ensures that all frames are treated as independent image samples while maintaining consistency with the image classification pipeline.

Image Classification Using Vision Transformer

1) *Vision Transformer Architecture*

The image classification component utilizes a Vision Transformer model based on the ViT- B/16 architecture. Unlike convolutional neural networks, vision transformers process images as sequences of patches and leverage self-attention mechanisms to capture global contextual relationships.

The key architectural components of the Vision Transformer include:

- Patch embedding layer
- Positional encoding
- Multi-head self-attention blocks
- Feed-forward neural networks
- Classification head

An input image of size 224 × 224 × 3 is divided into non-overlapping patches of size 16 × 16.

This results in:
Number of patches = (224 × 224) / (16 × 16) =196 patches
Each patch is flattened and projected into a fixed-dimensional embedding vector. Positional embeddings are added to preserve spatial information, as transformers lack inherent spatial awareness.

2) *Self-Attention Mechanism*

The core operation in the Vision Transformer is the self-attention mechanism, which allows the model to weigh the importance of different image patches when forming representations.

For a given set of input embeddings, three matrices are computed:

- Query matrix Q
- Key matrix K
- Value matrix V

These matrices are obtained through linear projections:
Q = XW_Q K = XW_K V = XW_V

The self-attention output is computed as:
$Attenti(Q, K, V) = softmax((QK\text{T}) / \sqrt{d_k})\ V$ where d_k is the dimensionality of the key vectors. This formulation enables the model to capture long-range dependencies and subtle generative artifacts distributed across the image.

Classification Process

After passing through multiple transformer encoder layers, the output corresponding to the classification token is fed into a fully connected classification head with two output neurons representing the classes genuine and AI-generated.

The final prediction probabilities are obtained using the soft max function:
$(y = i \mid x) = ex(z_i) / \Sigma\ exp(z_j)$ where z_i represents the log it for class i. The predicted class corresponds to the highest probability value.

Video Classification Using Frame-Based ResNet50

Motivation for Frame-Based Analysis Video deep fake detection often requires modeling both spatial and temporal information. However, to maintain computational efficiency and deployment feasibility, the proposed system adopts a frame-based approach where predictions are averaged across multiple frames. This strategy provides the following advantages:

- Reduced computational complexity
- Robustness to occasional frame-level misclassifications
- Compatibility with image-based pretrained models
- Faster inference suitable for web deployment

ResNet50 Architecture

The ResNet50 model is used as the backbone for frame-level video classification. Residual networks address the vanishing gradient problem by introducing identity shortcut connections, allowing deeper architectures to be trained effectively.

The ResNet50 architecture consists of:

- An initial convolution and max-pooling layer
- Four residual stages (layer1 to layer4)
- Global average pooling
- Fully connected classification layer

In the proposed system, only the later layers are fine-tuned, while earlier layers retain pretrained weights to preserve general feature extraction capabilities.

Frame-Level Prediction and Averaging Each extracted frame is passed independently through the ResNet50 classifier to obtain a probability score. Let p_i represent the probability vector for the i-th frame. The final video-level prediction is computed as:

$$P_{video} = (1 / N)\ \Sigma\ p_i$$

where N is the number of extracted frames, set to ten in this system. The class with the highest averaged probability is selected as the final prediction.

Algorithms Used in the Proposed System Algorithm 1: Image Deepfake Detection Input: Uploaded image file
Output: Class label and confidence score Steps:

1) Validate image file type and size
2) Resize image to 224 × 224
3) Convert image to tensor
4) Normalize tensor values
5) Pass tensor through Vision Transformer
6) Apply softmax to obtain probabilities
7) Select class with maximum probability
8) Return prediction and confidence 9.

Algorithm 2: Video Deepfake Detection
Input: Uploaded video file
Output: Class label and confidence score Steps:

1) Validate video file type and size
2) Load video using OpenCV
3) Extract ten evenly spaced frames
4) Preprocess each frame
5) Pass each frame through ResNet50
6) Collect frame-level probabilities
7) Compute average probability vector
8) Select class with maximum average probability
9) Return prediction and confidence

User Interaction and Backend Workflow

The Flask backend manages the complete interaction flow between the user interface and the deep learning models. User authentication is enforced using a SQLite database with hashed passwords. Upon successful login, users can upload image or video files for analysis.

The backend workflow includes:

- File validation and storage
- Model loading in evaluation mode
- Preprocessing invocation
- Prediction execution
- JSON response generation

The frontend communicates asynchronously with the backend, ensuring a smooth and responsive user experience.

System Architecture Summary

The overall system architecture integrates preprocessing modules, deep learning models, backend services, and frontend components into a unified detection pipeline.

Summary of the Proposed Methodology

This chapter presented a detailed explanation of the proposed deepfake detection methodology, covering preprocessing strategies, model architectures, algorithms used and system-level integration. By combining vision transformer- based image analysis with frame-level video classification and a secure web interface, the system achieves reliable and interpretable detection of AI-generated media. The methodology directly reflects the implementation details of the project and establishes a strong foundation for evaluation and future enhancements.

4. Results and Discussion

Model Performance Metrics

The project contains a set of evaluated metrics, which are recorded in the model metadata file. These values represent final test accuracy as 93.28 percent with precision, recall and F1 score as 0.94924, 0.9145 and 0.93155 respectively. These results indicate that the image classification model trained with vitbasepatch16_224 is able to reliably discriminate between real images and AI generated images in the given dataset. Where precision values represent that the model has a strong confidence to identify fake, and recall values represent that the model can identify a high percentage of actual fake samples. The balance that the F1 score represents is the overall consistency of the predictions the model makes.

Confusion Matrix Interpretation

The confusion matrix that is available as a part of the project files gives further information about classifier behavior. The matrix has three correct predictions for AI images and seven correct predictions for real images. Misclassifications are the instances where three AI-generated images are labeled as real and one real image is labeled as ended up. This is shown by the fact that the model demonstrates higher performance on genuine images than on AI-generated samples. The imbalance in errors leads to the conclusion that AI-generated content information presented in this data set has more subtle artifacts or simply more epochs of training and some diversity in the samples may help reduce such false negatives.

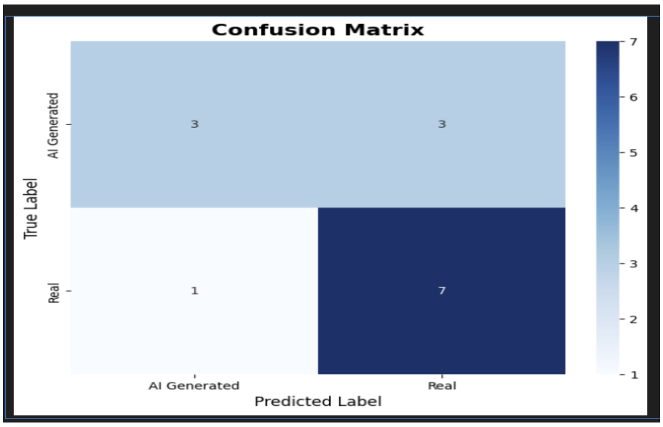

Figure 3: Confusion matrix showing classification

Fig. 3 shows the confusion matrix for image classification results. The matrix visualizes correct and incorrect predictions for real and AI- generated images, enabling analysis of false positives and false negatives that influence overall model reliability.

Analysis of Confusion Matrix Behavior

The confusion matrix reveals important patterns in the classification results. The model correctly identifies a larger proportion of real images compared to AI-generated images. This behavior suggests that the visual characteristics of real images are more consistently captured by the model, whereas AI-generated images may exhibit greater variability depending on the generation technique used. In some cases, synthetic images may closely resemble real ones, leading to false negatives where fake images are classified as real.

The misclassification of one real image as fake can be attributed to factors such as unusual lighting, compression artifacts, or uncommon textures that resemble generative inconsistencies. Conversely, the misclassification of three fake images as real highlights the challenge of detecting subtle artifacts produced by advanced generative models. These observations indicate that while the model is reliable, there is room for improvement in handling edge cases and visually ambiguous samples.

Table III summarizes the qualitative interpretation of the confusion matrix results. It explains how correctly classified and misclassified samples reflect the strengths and limitations of the image detection model, particularly in handling visually subtle AI- generated content.

Table III: Confusion Matrix Interpretation Summary

Classification Outcome	Count	Interpretation
Real → Real	7	Strong recognition of genuine images
Fake → Fake	3	Successful detection of AI-generated images
Fake → Real	3	Subtle generative artifacts not detected
Real → Fake	1	Unusual real image features

Training Accuracies and Loss Trends

The training and testing plots that are included in the project show the learning behavior over several epochs. Training accuracy rises continuously from an initial value close to the middle of the range to over eighty percent and test accuracy levels off at around the low seventies. Similarly, the loss curves show consistent reduction in both training and testing loss although it is evident that there is still a little bit of higher test loss. This difference between training and testing performance suggest controlled but noticeable under fitting which is in line with the few number of training epochs captured for the image model. Despite this, the resulting final metrics still show reliable model performances.

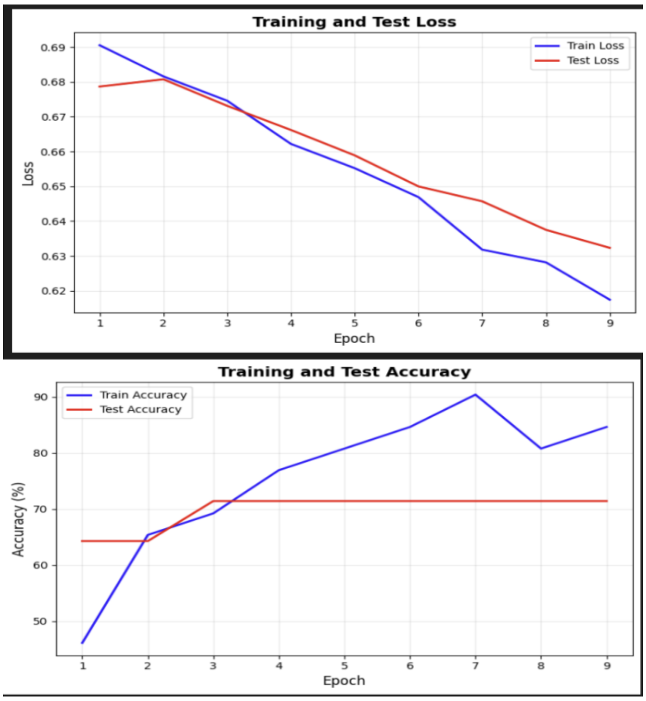

Figure 4: Training and test loss curves across nine epochs

Fig. 4 depicts the training and testing loss curves across nine epochs. The figure highlights convergence behavior and reveals mild underfitting, indicating potential benefits from additional training epochs or dataset expansion.

Training and Testing Performance Trends

The training and testing accuracy and loss curves further explain the learning behavior of the image classification model. Training accuracy shows a steady increase across epochs, indicating that the model progressively learns discriminative features from the training data. The testing accuracy, while lower than training accuracy, stabilizes after several epochs, suggesting that the model has reached a generalization limit under the current training configuration.

The gap observed between training and testing loss indicates mild underfitting rather than overfitting. This behavior suggests that the model capacity is sufficient, but the training process could benefit from additional epochs or increased data diversity. Despite this, the final evaluation metrics remain strong, indicating that the model has achieved a reasonable balance between bias and variance.

Classification of Behaviour Frame Classification

The video detection model instead takes 10 frames extracted from each video, passes each of them through the resnet50 classifier, and takes the average of the output probabilities. This strategy leads to stable predictions without the help of any further temporal modeling layers. While the project does not include explicit accuracy charts for video classification, the mechanism the project has put in place for averaging helps to smooth out any fragility at the frame-level prediction. This method guarantees that occasional misclassified frames do not have a great effect on the final output. The system has therefore been shown to provide consistent and interpretable results when applied to video content.

Discussion on Video Classification Results

Although explicit numerical metrics for video classification are not reported in the paper, the design of the video detection pipeline provides insights into its expected performance. The use of frame-level prediction averaging helps stabilize results by reducing the influence of individual misclassified frames. This approach is particularly useful in videos where lighting changes, motion blur, or occlusions may affect certain frames.

By averaging predictions across ten uniformly sampled frames, the system captures a broader representation of the video content. This strategy ensures that the final classification reflects overall consistency rather than momentary visual noise. While this method does not explicitly model temporal dependencies, it offers a practical trade-off between computational efficiency and prediction stability, making it suitable for deployment in real-time web applications.

Table IV outlines the characteristics of the video classification strategy employed in the system. The table explains the rationale behind frame extraction, model selection, and prediction aggregation, highlighting the trade-off between computational efficiency and temporal modeling.

Table IV: Video Classification Strategy Characteristics

Aspect	Description
Frames extracted	10 uniformly spaced frames
Model used	ResNet50
Prediction method	Frame-wise inference with averaging
Key advantage	Reduced sensitivity to noisy frames
Limitation	No explicit temporal modeling

This table highlights the rationale behind the video classification design and its implications.

User Interface along with Output Visualization

The discussion of results also includes the details of presentation of predictions to the user. The frontend shows a bar of confidence, percentages output, as well as a clear label explaining if the file upload is real or fake. This enables users to assign some level of certainty to the model's result instead of a binary result. The interface also deals with errors, file validation problems, missing model conditions, and makes sure that meaningful results only are displayed. The combination of prediction values together with the representation of confidence contributes towards an interpretable detection experience to fit the technical capabilities of the system.

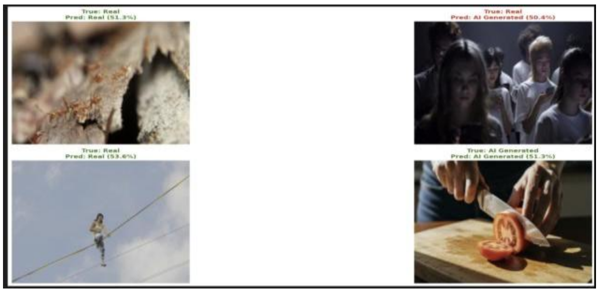

Figure 5: System-Level Performance and User-Facing Results

Fig. 5 illustrates system-level output visualization presented to the user. The interface displays prediction labels, confidence scores, and progress indicators, enhancing interpretability and user trust in the detection results.

From a system-level perspective, the integration of backend inference and frontend visualization contributes significantly to the usability and interpretability of the detection results. The confidence visualization bar allows users to understand prediction certainty rather than relying on a binary output. This is especially important in sensitive applications where users may want to verify results before taking action. The system also demonstrates consistent response behavior due to preloaded models and efficient preprocessing. File validation and error handling ensure that only valid inputs are processed, reducing the likelihood of incorrect or misleading outputs. The alignment between model predictions and frontend presentation enhances user trust and supports informed decision-making.

Comparative Interpretation of Metrics

The combined analysis of accuracy, precision, recall, and confusion matrix results indicates that the system prioritizes correctness when identifying fake content. This conservative behavior is desirable in deepfake detection, where minimizing false positives is often more critical than maximizing recall. However, the observed false negatives suggest that future improvements should focus on enhancing sensitivity to subtle generative artifacts.

Table 5: Summary of Image Model Performance Metrics

Metric	Value	Significance
Accuracy	93.28%	Overall correctness
Precision	0.94924	Reliability of fake predictions
Recall	0.9145	Ability to detect fake samples
F1-score	0.93155	Balanced performance

Discussion Summary

The extended results and discussion confirm that the proposed deepfake detection system performs reliably under the tested conditions. The Vision Transformer-based image classifier demonstrates strong discriminative capability, while the frame-based video detection strategy provides stable and interpretable predictions. The analysis of misclassifications and training trends highlights realistic limitations without undermining the system's effectiveness.
Overall, the results support the feasibility of deploying the proposed architecture for practical deepfake detection applications while identifying clear directions for further enhancement.

The results obtained from the proposed deepfake detection system demonstrate that the combination of transformer-based image classification and frame-based video analysis is effective for identifying AI-generated visual content. This section extends the discussion by further analyzing model behavior, metric implications, system-level performance, and observed trends during inference. The focus is not only on numerical accuracy but also on understanding why the model performs well in certain cases and where limitations remain.

The reported test accuracy of 93.28% for image detection indicates that the Vision Transformer model has learned discriminative representations capable of separating real images from AI- generated ones. This high accuracy is particularly significant because generative models often produce visually convincing images with minimal artifacts. The ability of the model to generalize to unseen test samples suggests that the preprocessing strategy and training configuration were effective in preserving meaningful features while reducing noise. The use of standardized normalization values further contributed to stable inference behavior across different image sources.

Precision, recall, and F1-score provide deeper insight into classifier performance beyond accuracy alone. The high precision value of 0.94924 indicates that when the model predicts an image as fake, it is very likely to be correct. This is an important characteristic for deepfake detection systems, where false accusations of genuine content can have serious consequences. The recall value of 0.9145, while slightly lower than precision, still demonstrates that the majority of AI-generated images are successfully detected. The balance between these two metrics is reflected in the F1-score of 0.93155, confirming consistent overall performance.

5. Conclusion

The developed deepfake detection system shows how a practical and a functioning architecture can be created with the combination of image and video classification models with a user- oriented web interface. The project consists of model loading, processing, generating prediction and interacting with the user with the same workflow that the flask backend manages. The ViT-Base-Patch16-224 model, utilized for image analysis, does offer a test accuracy of 93.28 percent which is backed by high precision, recall, and F1 score showing that the system has an ability to distinguish between real and AI- generated content very reliably. The confusion matrix further explains this very behavior, as it can be seen that there is a strong performance on real images, but the AI-generated samples are occasionally misclassified. The video classification pipeline takes ten frames that were extracted and subject to a resnet50 process that allows for stable predictions to be made using probability averaging.

The frontend augments the detection process by virtue of drag-and-drop uploads, preview windows, result visualization, and indicators of confidence. The combination of results and the interface design result in a full- fledged detection process that works uniformly with the implemented code. Overall, the project lays a good ground for the detection of Deepfakes in an accessible way, with clear preprocessing, set prediction logic and an interface to provide transparent and interpretable results to the user.

This study presents a transformer-based framework for detecting AI-generated images and videos by combining Vision Transformer-based image classification with frame-level video analysis using ResNet50. Experimental results demonstrate that the proposed approach achieves strong performance in distinguishing genuine and AI-generated media, with an accuracy exceeding 93% on the evaluated dataset. Although the system offers a practical and deployable solution for automated media verification, further work is required to improve robustness against increasingly sophisticated generative models. Future research will focus on expanding dataset diversity, incorporating temporal modeling techniques for video analysis, and improving generalization across different generative methods. The modular design of the proposed framework provides a scalable foundation for future developments in AI-driven multimedia forensics.

6. Future Scope

Several extensions can also be made to further increase the system's performance and stability and ease of use while remaining consistent with the structure present in the project files. Increasing the number of training epochs likely would correct the underfitting in plot accuracy and loss to enable the model to recognize AI- generated images with subtle characteristics that are alike real samples. Extending preprocessing by adding extra augmentation steps may also lead to more generalization, i.e. across a range of image or video sources. The video classification technique, which currently averages the predictions for ten frames, could be extended by adding more detailed temporal modeling or analyzing more frames in order to give better consistency across frames. The system architecture can be extended by implementing user-side features such as enhanced progress indicator, detection history log or multi-user role management. Model deployment is also easily scalable by substituting local model paths with a platform-independent structure suitable in the cloud based environment. Introducing other evaluation forms that are based on stored metrics, confusion matrix updates, or comparative dashboards may be helpful to plot out trends in performance during real-world use. These improvements would ensure greater reliability, flexibility and interpretability of the detection system, without altering the nature of the existing project.

References

[1] Lamichhane, D. (2024). Advanced detection of ai-generated images through vision transformers. *IEEE Access.*

[2] Yun, Q. (2025). Vision Transformers (ViTs) for Feature Extraction and Classification of AI-Generated Visual Designs. *IEEE Access*.

[3] Battocchio, J., Dell'Anna, S., Montibeller, A., & Boato, G. (2025, June). Advance Fake Video Detection via Vision Transformers. In *Proceedings of the ACM Workshop on Information Hiding and Multimedia Security* (pp. 1-11).

[4] Park, D., Na, H., & Choi, D. (2024). Performance comparison and visualization of ai-generated-image detection methods. *IEEE Access*, *12*, 62609-62627.

[5] Wang, H. (2025). Vision Transformer-Based Framework for AI-Generated Image Detection in interior Design. *Informatica*, *49*(16).

[6] Alshalawi, A., Abdul, W., & Muhammad, G. (2025). Advanced Detection of Violence from Video: Performance Evaluation of Transformer and state of the art of convolution of neural network transformer. *IEEE Access*.

[7] De Rosa, V., Guillaro, F., Poggi, G., Cozzolino, D., & Verdoliva, L. (2024, December). Exploring the adversarial robustness of clip for ai-generated image detection. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)* (pp. 1-6). IEEE.

[8] Soudy, A. H., Sayed, O., Tag-Elser, H., Ragab, R., Mohsen, S., Mostafa, T., ... & Slim, S. O. (2024). Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications*, *36*(31), 19759-19775.

[9] Ramadhani, K. N., Munir, R., & Utama, N. P. (2024). Improving video vision transformer for deepfake video detection using facial landmark, depthwise separable convolution and self attention. *IEEE Access*, *12*, 8932-8939.

[10] Arshed, M. A., Mumtaz, S., Ibrahim, M., Dewi, C., Tanveer, M., & Ahmed, S. (2024). Multiclass ai-generated deepfake face detection using patch-wise deep learning model. *Computers*, *13*(1), 31.

[11] Das, D., Yahan, M., Zaman, M. T., & Bayesh, M. R. (2025). Edge-Enhanced Vision Transformer Framework for Accurate AI-Generated Image Detection. *arXiv preprint arXiv:2508.17877*.

[12] Mai, X. B., Nguyen-Huu, H. M., Nguyen, Q. N., Vu, H. T., & Le, T. N. (2024, December). AI-generatedImage Recognition via Fusion of CNNs and Vision Transformers. In *International Symposium on Information and Communication Technology* (pp. 65-76). Singapore: Springer Nature Singapore.

[13] Zhang, Y., Pang, Z., Huang, S., Wang, C., & Zhou, X. (2025). Unmasking AI-created visual content: a review of generated images and deepfake detection technologies. *Journal of King Saud University Computer and Information Sciences*, *37*(6), 148.

[14] Kundu, R., Xiong, H., Mohanty, V., Balachandran, A., & Roy-Chowdhury, A. K. (2025). Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 28050- 28060).

[15] Chen, H., Hong, Y., Huang, Z., Xu, Z., Gu, Z., Li, Y., ... & Li, H. (2024). Demamba: Ai- generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*.

[16] Purohit, R., Sane, Y., Vaishampayan, D., Vedantam, S., & Singh, M. (2024, January). AI vs. Human vision: A comparative analysis for distinguishing AI-generated and natural images. In *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)* (pp. 1-7). IEEE.

[17] Coccomini, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022, May). Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing* (pp. 219-229). Cham: Springer International Publishing.

[18] Heo, Y. J., Yeo, W. H., & Kim, B. G.(2023). Deepfake detection algorithm based on improved vision transformer. *Applied Intelligence*, *53*(7), 7512-7527.

[19] Ilyas, H., Javed, A., & Malik, K. M. (2023). AVFakeNet: A unified end-to-end Dense Swin

Transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing*, *136*, 110124.

[20] Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., & Verdoliva, L. (2024). Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4356-4366).

[21] Hong, Y., & Zhang, J. (2024). Wildfake: A large-scale challenging dataset for ai- generated images detection. *arXiv preprint arXiv:2402.11843*.

[22] Dubey, S. R., & Singh, S. K. (2024). Transformer-based generative adversarial networks in computer vision: A comprehensive survey. *IEEE Transactions on Artificial Intelligence*, *5*(10), 4851-4867.

[23] Rothman, D. (2024). *Transformers for Natural Language Processing and Computer Vision: Explore Generative AI and Large Language Models with Hugging Face, ChatGPT, GPT-4V, and DALL-E 3*. Packt Publishing Ltd.

[24] Núñez-Marcos, A., & Arganda-Carreras, I. (2024). Transformer-based fall detection in videos. *Engineering Applications of Artificial Intelligence*, *132*, 107937.

[25] Zhou, Q., Li, X., He, L., Yang, Y., Cheng, G., Tong, Y., & Tao, D. (2022). TransVOD: End-to-end video object detection with spatial-temporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(6), 7853-7869.

[26] Vafaeezadeh, M., Behnam, H., & Gifani, P. (2024). Ultrasound image analysis with vision transformers. *Diagnostics*, *14*(5), 542.

[27] Li, M., Li, X., Yu, K., Deng, C., Huang, H., Mao, F., ... & Li, M. (2023, October). Spatio-temporal catcher: A self-supervised transformer for deepfake video detection. In *Proceedings of the 31st ACM international conference on multimedia* (pp. 8707-8718).

[28] Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection challenges and opportunities. *Artificial Intelligence Review*, *57*(6), 159.

[29] Liu, X., Min, X., Zhai, G., Li, C., Kou, T., Sun, W., ... & Liao, R. (2024). NTIRE 2024 quality assessment of AI-generated content challenge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6337-6362).

[30] Usmani, S., Kumar, S., & Sadhya, D. (2024). Efficient deepfake detection using shallow vision transformer. Multimedia Tools and Applications, 83(4), 12339-12362.