WHITEPAPER

DeepTruth - AI

2024

CONTENTS

ABSTRACT

Due to the easy transfer of images, audio, and video sources through an internet connection, the amount of information online has exponentially grown in recent years. Nowadays, the physical identity of a person is more important than it ever was because it can be easily manipulated for purposes such as entertainment, spreading misinformation, manipulation, financial gain and fraud, unauthorized access, revenge pornography, and political misleading. The ease of finding information about the field of machine learning, modern and well-documented tools like Pytorch, Tensorflow, or Keras, easy access to open-source datasets and pre-trained models, rapid improvement of deep learning methods such as Generative Adversarial Networks, and inexpensive computational unit assets have made it possible for state-of-the-art deepfakes to be created and serve one of the purposes mentioned above. This report aims to present the results of an experiment conducted using the Vision Transformer (ViT) architecture, a cutting-edge model in the field of computer vision, to classify videos as either deepfake or authentic. By leveraging the powerful attention mechanisms of the Vision Transformer, this report seeks to demonstrate its potential in accurately identifying manipulated media, thereby contributing to the broader efforts in combating malicious activities facilitated by deepfake technology. Through rigorous testing and analysis, this report will provide insights into the performance, strengths, and limitations of the Vision Transformer in this critical application area.

1 INTRODUCTION

1.1 Context

According to Rana et al. (2022), the term "Deepfake" is derived from "Deep Learning (DL)" and "Fake", and it describes specific photo-realistic video or image contents created with DL's support. This word was named after an anonymous Reddit user in late 2017, who applied deep learning methods for replacing a person's face in pornographic videos using another person's face and created photo-realistic fake videos.

Threat actors are using disinformation campaigns and deepfake content to misinform the public about events, to influence politics and elections, to contribute to fraud, and to manipulate shareholders in a corporate context. Many organisations have now begun to see deepfakes as an even bigger potential risk than identity theft, according to Europol (2022). Even though deepfakes pose a significant risk solely by their inherently deceptive nature, they are not the greatest threat. However, the more immediate concern lies in how the concept of deepfakes can be exploited to cast doubt on authentic content. Exaggerated media coverage and speculation regarding the impact of deepfakes have diverted attention from actual instances where they have made an impact. Worth mentioning is the concept of "Crime as a Service" (CaaS), wherein criminals offer access to tools, technologies, and expertise to facilitate cyber and cyber-enabled crime. In Europol (2022) it is mentioned that CaaS is anticipated to progress alongside current technologies, potentially automating crimes like hacking, adversatial machine learning, and deepfakes. The increasing prevalence of disinformation and deepfakes is poised to profoundly affect public perception of authority and media information. As deepfakes become more widespread, trust in authorities and official information erodes. Experts express concerns that this could result in a scenario where citizens no longer share a common reality, causing societal confusion regarding reliable sources of information—a phenomenon sometimes termed as 'information apocalypse' or 'reality apathy.'

The Europol (2021) report shows that deepfake technology can facilitate varous criminal activities, including:

- harassing or humiliating individuals online;
- perpetrating extortion and fraud;
- facilitating document fraud;
- falsifying online identities and fooling 'know your customer' mechanisms;
- non-consensual pornography
- online child sexual exploitation;
- falsifying or manipulating evidence for criminal justice investigations;

- disrupting financial markets;
- distributing disinformation and manipulating public opinion;
- supporting the narratives of extremist or terrorist groups;
- stoking social unrest and political polarisation;

Recently, deepfake technology has advanced to remarkable heights, surpassing previous limitations. What initially started as a specialized and technically intricate pursuit has swiftly transformed into a pervasive trend. Nonetheless, with the evolution of technology, the distinction between creativity and deceit becomes increasingly blurred, presenting both astonishing possibilities and troubling implications for shaping reality. Home Security Heroes (2023) mentioned that there are 550% more deepfake videos online in 2023 than in 2019. Their research indicates that the digital realm will host a staggering 95820 deepfake videos, of which 98% are of pornographic nature. 99% of deepfake videos features women as the primary subjects, while only 1% of the content features men. South Korea is the country most targeted by deepfake forgery and 94% of those featured in deepfake content videos work in the entertainment industry.

Over the past few years, significant focus from scholars and specialists has been directed towards countering such dangers through advancements in Deepfake detection. This attention has spurred the development of numerous techniques aimed at identifying and mitigating the proliferation of Deepfake content.

1.1.1 Major Types

Understanding the types of deepfake content is crucial not only to analyze the technological process itself but also to decipher the evolution of this content, its purposes, and where it is heading. By comprehending the various types of deepfakes, we can gain a deeper perspective on their impact in different fields, as well as their implications in contemporary society. From harmless uses such as entertainment and artistic creation to dangerous ones like political manipulation or information fraud, the diversity of deepfake types demonstrates the breadth and complexity of this technological phenomenon.

As mentioned in Masood et al. (2021), deepfake content can be categorized into the following types:

- Face Swapping - involves replacing the face of one person with that of another in a video or image, creating a deceptive portrayal where the actions of the source person are attributed to the target person. These deepfakes are often used to exploit the fame or standing of well-known individuals by placing them in situations they never experienced, potentially harming their reputation, such as in cases of non-consensual pornography.
- Lips Syncing - involves altering the movements of a person's lips to match a particular audio recording. This technology aims to create the illusion of someone speaking in

a manner dictated by the attacker, regardless of what the actual individual said or intended.

- Puppet Master - involves replicating the expressions of the target person, including eye movements, facial expressions, and head movements. These deepfakes aim to manipulate the source person's expressions, and potentially even their entire body movements in a video, to animate them according to the impersonator's intentions.
- Face synthesis and attribute manipulation - involves creating highly realistic facial images and altering facial attributes. These manipulations are often employed to propagate misinformation on social media platforms through the use of fake profiles.
- Audio deepfakes - concentrate on producing the voice of the target speaker through deep learning methods, allowing them to utter statements they haven't actually said. These fabricated voices can be created using either text-to-speech synthesis (TTS) or voice conversion (VC) techniques.

The deepfake technology encompasses various types, including the one mentioned. These are often utilized for deceptive purposes, such as spreading disinformation and damaging reputations. However, alongside the evolution of deepfake creation methods, there is a parallel effort to develop detection techniques to identify and mitigate the spread of misleading content. Further exploration into both the advancement of deepfake generation and detection methods is crucial in addressing the ethical and societal challenges posed by this technology.

1.1.2 Detection

Deepfake Detection refers to the process of identifying and distinguishing manipulated or synthetic media, known as deepfakes, from genuine or unaltered content. It involves the development and application of various techniques, algorithms, and models to scrutinize media content for signs of manipulation of artificial generation. These methods often analyze both visual and auditory aspects of the media to identify inconsistencies, artifacts, or anomalies that may indicate the presence of a deepfake. Common approaches to deepfake detection include examining spatial and temporal inconsistencies, analyzing facial and body movements, detecting artifacts left by the generation process, and assessing contextual clues.

The clarity and consistency of the classification in Naitali et al. (2023) regarding the clues of deepfake sources are commendable. This work explores multiple categories of clues and provides extensive explanations for each, along with models that identify deepfakes based on these clues. Below, only the most relevant detection methods will be provided, which perform well even in limited situations.

In the overview of face swap deepfake detection techniques and their limitations created by Masood et al. (2021), there is a clear difference of best evaluation performance between the methods that are using handcrafted features and those that use deep learning-based features. There is a widespread recognition that deep learning-based models currently demonstrate

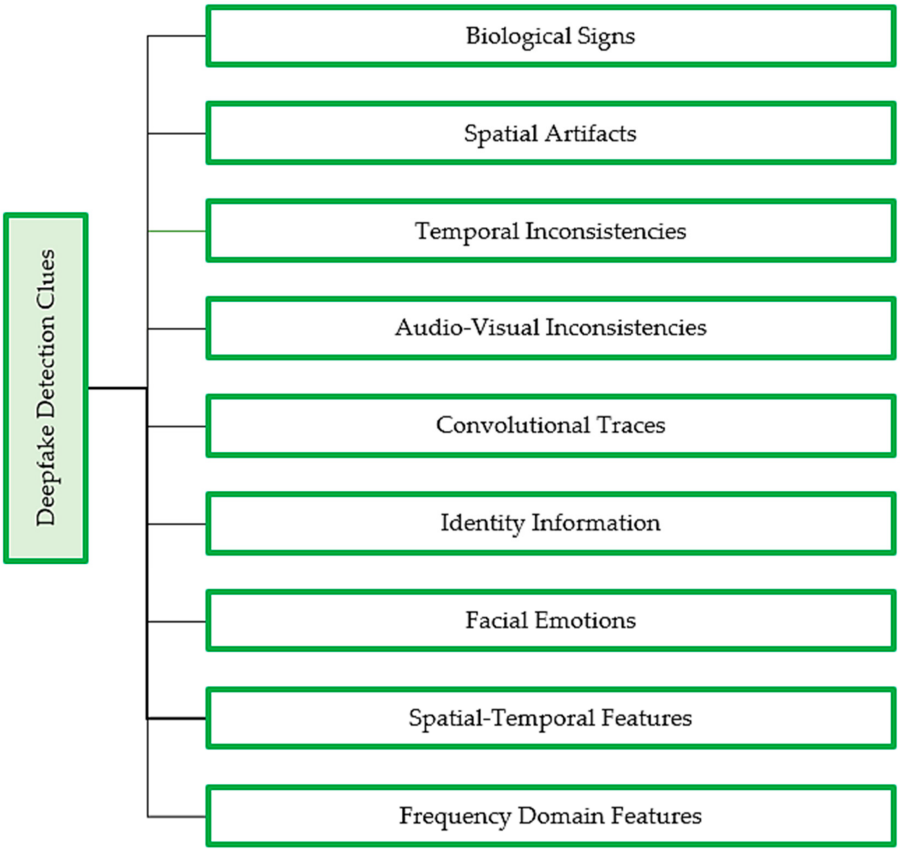

Figure 1: Clues of deepfake analysis Naitali et al. (2023)

the most outstanding performance in distinguishing between fabricated and authentic digital media. These models utilize advanced neural network architectures referred to as backbone networks, which have shown exceptional effectiveness in tasks related to computer vision. Notable examples of such architectures are VGG (Simonyan & Zisserman (2014)), Resnet (Targ et al. (2016)), EfficientNet(Tan & Le (2019)), Inception (Szegedy et al. (2016)), known for their excellence in the feature extraction stage. Deep learning-based detection models surpass traditional methods by integrating extra techniques to boost their effectiveness. Meta-learning is one of such techniques, empowering the model to learn from previous experiences and adapt its detection capabilities accordingly. Through these techniques, these model improve their ability to identify patterns and differentiate between authentic and altered content.

Moreover, in the training of deep learning-based detection models, data augmentation plays a pivotal role. This method involves enriching the training dataset with artificially generated or modified samples, thereby enhancing the model's ability to generalize and detect various forms of deepfake media. Data augmentation enables the model to learn from a wider range of examples and improves its robustness against different types of manipulations. Attention mechanisms have also proven to be valuable additions to deep learning-based detection models. By directing the model's focus toward relevant features and regions of the input data, attention mechanisms enhance the model's discriminative power and improve its overall accuracy (Naitali et al. (2023)). In summary, the integration of deep learning-based architectures, meta-learning,

data augmentation, and attention mechanisms represents a significant advancement in the field of deepfake detection.

The comparison between different detection methods presented in Masood et al. (2021) offers a accurate and consistent review of a wide variety of detection techniques. In the research Li & Lyu (2019), the detection mechanism proposed uses VGG16, ResNet50, ResNet101, ResNet152 as techniques, DLib facial landmarks as features, and the best evaluation performances of AUC(Google Developers (Year Accessed)) are 84.5 for VGG16, 97.4 for ResNet50, 95.4 for ResNet101, 93.8 for ResNet152. The dataset used is DeepFake-TIMIT(Idiap Research Institute (2024)) and the limitation identified is that it is not robust for multiple video compression. Another notable deepfake detection mechanism is the one proposed in Güera & Delp (2018). This approach uses CNN/RNN as technique, deep features and the accuracy obtained on a customized datased is 97.1%. One important limitation is that this model is applicable to short videos only.

Worth mentioning is another type of deepfake detector architecture, the vision transformer (Dosovitskiy et al. (2020)), long for ViT. It represents a robust alternative to convolutional neural networks (CNNs), showcasing remarkable performance advancements. These models have demonstrated superiority over state-of-the-art CNNs, offering nearly fourfold improvements in both computational efficiency and accuracy. The central piece of ViTs are transformers, a category of non-sequential deep learning models, which leverage self-attention mechanisms to assign varying degrees of importance to different parts of input data Han et al. (2021). A notable transformer variant is Video Transformer (Neimark et al. (2021)), optimized for efficiently processing large-scale video data, thereby maximizing computational resource usage and minimizing runtime. This capability enables full video processing during test time, making VTNs particularly well-suited for handling lengthy videos.

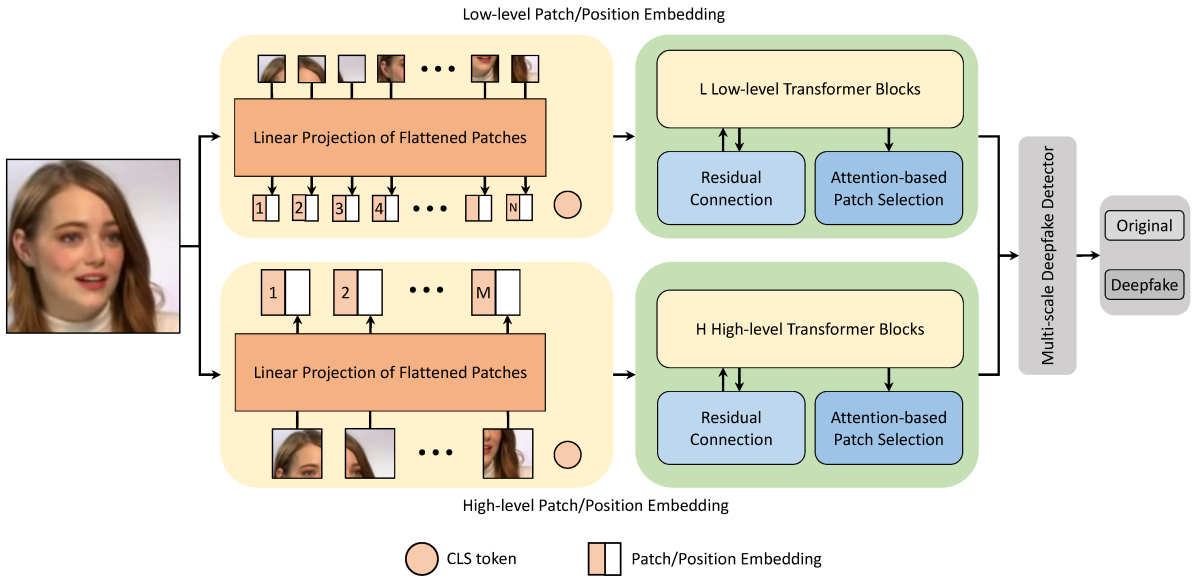

Figure 2: Vision Transformer-Based Deepfake Detection Khormali & Yuan (2022)

1.2 Report Structure

Up to this point, information has been provided regarding the field of study, classical and modern approaches to solving the problem, and well-known datasets that have contributed to the realm of deepfake content. The underlying reasons for the existence of this field, the various types of such content, along with its associated risks and future implications, have been outlined. In Chapter "3 METHOD", a Vision Transformer method, inspired by the article Dosovitskiy et al. (2020) will be presented, detailing its architecture, implementation, and the specific steps taken to apply it to the task of deepfake detection. This chapter will outline the preprocessing techniques used to prepare the video data, the training process, including the choice of loss functions and optimization algorithms, and the rationale behind selecting the Vision Transformer for this experiment. Additionally, the chapter will discuss the evaluation metrics used to assess the model's performance, the experimental setup, and the challenges encountered during the training and testing phases. The goal is to provide a comprehensive understanding of how the Vision Transformer method was utilized and the results it yielded in identifying deepfake content.

1.3 Related Work

The Vision Transformer (ViT) architecture, introduced by Dosovitskiy et al. (2020), represents a significant shift in the design of deep learning models for computer vision tasks, traditionally dominated by convolutional neural networks (CNNs). The ViT leverages the transformer architecture, originally proposed by Vaswani (2017) for natural language processing, to process image data, thereby challenging the long-standing dominance of CNNs in the field of computer vision. Prior to the introduction of the ViT, transformers had primarily been used in natural language processing (NLP). The transformer architecture's self-attention mechanism, which allows for capturing long-range dependencies in sequential data, was initially considered unsuitable for image processing due to the two-dimensional nature of images and the prohibitive computational cost associated with processing high-resolution images. However, the introduction of the ViT demonstrated that by splitting images into patches and treating these patches as tokens in a sequence, transformers could be effectively applied to vision tasks. This approach was shown to be highly effective, particularly in scenarios with large-scale data, where ViTs could outperform state-of-the-art CNNs. Early methods for detecting deepfakes primarily relied on CNNs to identify inconsistencies or artifacts in visual content. These methods focused on pixel-level anomalies, such as irregularities in eye blinking, lighting inconsistencies, and unnatural facial movements. Approaches like MesoNet (Afchar et al. (2018)) and XcenptionNet(Chollet (2017)) demonstrated some success in detecting manipulated images and videos, but they often struggled with high-quality deepfakes that effectively masked such artifacts.

The introduction of Vision Transformers has brought a new dimension to deepfake detection.

By treating image patches as tokens and leveraging self-attention mechanisms, ViTs can capture both local and global features in visual data more effectively than traditional CNNs. This capability is particularly advantageous in deepfake detection, where subtle global inconsistencies and fine-grained local anomalies must be identified. The application of ViTs to deepfake detection is a relatively recent development, but early results suggests that they offer improved performance, particularly in scenarios where large-scale datasets are available for training. To further enhance detection capabilities, hybrid models that combine the strengths of CNNs and ViTs have been proposed. These models typically use CNNs for initial feature extraction, followed by transformers to model long-range dependencies. The Swin Transformer, proposed in Liu et al. (2021) and its variants, which introduce hierarchical attention mechanisms, have shown promise in detecting deepfakes in more complex scenarios, such as those involving multiple people or highly realistic audio-visual manipulations. Moreover, extentions of ViTs, such as the Data-efficient Image Transformer (DEiT) and Swin Transformer, have been adapted for deepfake detection tasks. These models are designed to operate efficienstly with limited data, making them suitable for scanarios where annotated deepfake datasets are scarce.

2 METHOD

In this report, we utilize a Vision Transformer (ViT) architecture for deepfake detection. The ViT model is selected due to its ability to capture global context from input images, which is crucial for identifying subtle manipulations in deepfakes. The architecture operates on image patches rather than the entire image, allowing it to process high-resolution inputs efficiently.

2.1 Dataset

The Deepfake Detection Challenge (DFDC) (Dolhansky et al. (2020)) dataset stands at the forefront of research into multimedia manipulation, particularly focusing on facial manipulation techniques.

The dataset started from 48190 total videos that average 68.8s each - a total of 38.4 days of footage. 3426 subjects were involved in total with an average of 14.4 videos each, with most videos shot in 1080p. The authenticity of the 48000 videos in the DFDC dataset was manipulated using various techniques, primarily focus on on face-swapping and audio-swapping methods. One of the methods used is the convolutional autoencoder model, especially chosen to reflect the technology behind existing deepfake videos on the internet. Another methods that represent state-of-the-art generation technology, are the GAN-based Methods. Neural Talking Heads (NTH) (Zakharov et al. (2019)) generates realist talking heads with few and one-shot learning settings, incorporating meta-learning and fine-tuning stages. FSGAN (Nirkin et al. (2019)) which employs GANs for face swapping and reenactment, considering pose and expression variations. DFDC dataset includes audio swapping methods alongside face-swapping techniques. Specifically, the TTS Skins (Polyak et al. (2020)) voice conversion method was used to manipulate the audio in some of the video clips. While the dataset primarily focuses on facial manipulation techniques, the inclusion of audio swapping adds another dimension to the manipulation types explored within the dataset.

To calculate the total number of videos in the training, validation and test sets, the sum of the number of videos in each set would add up to 133154 videos (119,154 for training, 4000 for validation and 10000 for testing). The DFDC dataset represents a significant contribution to the field of multimedia manipulation research, particularly in the context of deepfake detection and analysis. By compiling over 133,000 videos across training, validation, and test sets, the dataset provides a rich and diverse collection of manipulated and authentic multimedia content for researchers and practitioners to explore.

2.2 Image Extraction

The image classification task for deepfake detection is critical because it serves as the foundational step in identifying manipulated content within a video. Each frame extracted from the video is analyzed individually, and the accuracy of the model in classifying these frames as real or fake determines the reliability of the overall detection process. The success of deepfake detection hinges on the model's ability to consistently and accurately classify each frame, as even a small error rate could lead to false conclusions about the authenticity of the video. Therefore, optimizing the image classification task is crucial to ensure robust and reliable detection of deepfakes.

To enhance the accuracy of deepfake detection, the process begins by extracting faces from individual video frames using a face detection model like MTCNN (Multi-task Cascaded Convolutional Networks). MTCNN is particularly well-suited for this task as it efficiently detects faces and key facial landmarks, ensuring precise cropping of the facial region from each frame. Once the faces are isolated, these extracted facial regions are passed through a deepfake detection model, which focuses on identifying subtle manipulations and artifacts that are often present in deepfakes. By concentrating exclusively on the facial area, the detection model can more effectively analyze and identify potential signs of tampering, leading to more reliable and accurate deepfake detection. This approach also minimizes the influence of irrelevant background information, enabling the model to zero in on the most critical elements of the video—the faces—where deepfake manipulations are most likely to occur.

In deepfake detection using frames, a unique challenge arises when individual frames from a manipulated video appear to be unaltered. This situation can occur if the deepfake generation model fails to consistently apply manipulations across all frames, resulting in a video where some frames are technically part of a fake video but show no visible signs of alteration. These unaltered frames can mislead detection models, as they do not exhibit the typical artifacts or inconsistencies usually associated with deepfakes. This issue underscores the importance of considering the context of the entire video rather than solely relying on the analysis of isolated frames.

A future approach would be to train the deepfake detection model to incorporate not only the original frames extracted from videos but also additional manipulated frames where specific alterations, such as removing or distorting parts of the face, have been made. Methods for augmenting faces extracted from frames were addressed in a Das et al. (2021) written by the winner of the competition from which the dataset originates. One of the key benefits of using face cutout techniques in deepfake detection is that it improves the model's ability to focus on relevant regions of the face, particularly the manipulated areas. By selectively occluding certain facial regions, based on facial landmarks, the model avoids learning redundant or irrelevant features, which can reduce overfitting. These artificially modified frames introduce a wider variety of anomalies, enabling the model to learn and recognize a broader spectrum of potential manipulations. By exposing the model to these diverse examples during training, it

becomes more robust and less reliant on detecting only specific types of artifacts associated with traditional deepfakes. This approach enhances the model's ability to generalize to unseen manipulations, making it more effective at identifying a wide range of deepfakes, including those that use more sophisticated or subtle techniques. As a result, the model becomes more adaptable and reliable in real-world scenarios, where the nature of manipulations can vary significantly.

2.3 Vision Transformer Architecture

The Vision Transformer (ViT) architecture introduces a novel approach to image classification by leveraging the transformer architecture, which has been highly successful in natural language processing (NLP) tasks. The architecture adapts the transformer model for image recognition by treating images as sequences of fixed-size patches, analogous to tokens in NLP, and processing them with standard transformer layers.

In this architecture, an image is first divided into non-overlapping patches, typically 16x16 or 32x32 pixels in size, from a larger image, such as 224x224 pixels. Each patch is then flattened into a one-dimensional vector and linearly embedded into a vector of dimension D, which matches the transformer's input dimension. These embeddings are supplemented with position embeddings to retain information about the relative positions of patches within the image since the transformer architecture does not inherently capture spatial relationships.

The sequence of patch embeddings, along with their corresponding position embeddings, is fed into a standard transformer encoder. The transformer encoder consists of multiple layers of multi-head self-attention mechanisms and feedforward neural networks. The self-attention mechanism enables the model to weigh the importance of different patches relative to each other, capturing both local and global dependencies across the image. This is achieved by computing attention scores between all patches in the sequence and using these scores to generate a weighted sum of the input patches.

Following the self-attention mechanism, each token is passed through a feedforward network, typically consisting of two linear layers with a non-linear activation function in between, enhancing the model's capacity to learn complex representations of the input data. Both the multi-head self-attention and feedforward network components are followed by layer normalization and residual connections, which help stabilize training and improve the model's performance.

A classification token, which is a learned vector prepended to the sequence of patch embeddings before the transformer encoder, is used to represent the entire image after the final transformer encoder layer. This classification token is passed through a final fully connected layer, producing the logits corresponding to the predefined classes in the image recognition task.

When fine-tuning ViT using a pre-trained model like torchvision.models.ViT_B_16_Weights.DEFAULT, the process builds on the learned knowledge from a model that was previously trained on a large dataset such as ImageNet. This pre-trained model includes weight configurations that have already learned general image features, which can be adapted to specific tasks or datasets through fine-tuning. It involves replacing the final classification layer with one suited to the new task, while the remaining layers of the network are initialized with the pre-trained weights. During training, the model adjusts these weights to better fit the new dataset, effectively adapting the pre-trained model to the specific characteristics of the new images.

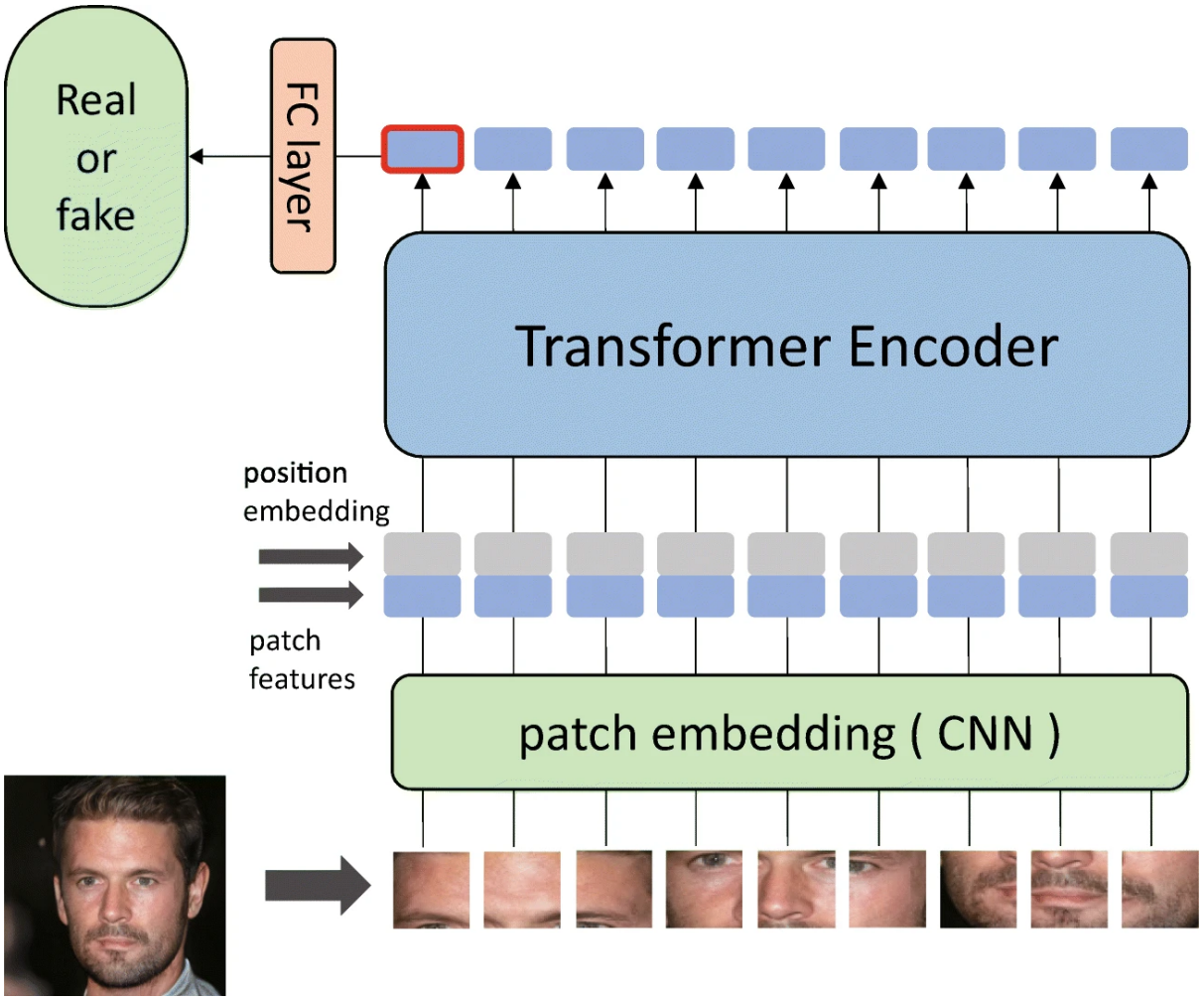

Figure 3: The model overview

3 RESULTS

In this section, the outcomes of applying our proposed deepfake detection algorithm to 10% of the dataset comprising authentic and synthetic media is evaluated. The evaluation reflects various performance metrics, providing insights into the efficacy and efficiency of the approach in identifying manipulated content.

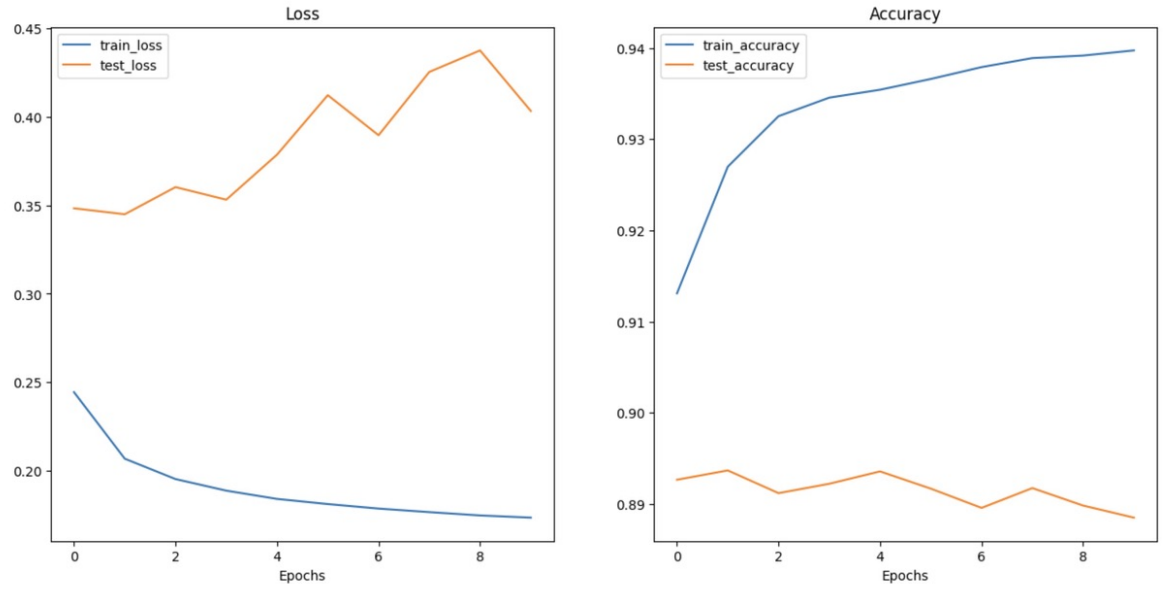

Figure 4: Accuracy and Loss Evolution during fine-tuning

The fine-tuning was done using 10% of the dataset, which was split into test and train. The training part has 6410 real images that come from 641 videos and 55594 fake images that come from 5559 videos. The testing part has 1819 real images that come from approximately 182 videos and 15042 fake images that come from 1504 videos. The model obtained an accuracy of 93,97%, which is better compared to the version that use Convolutional Neural Networks for the same task and for the same dataset.

Table 1: Performance of the method using 10% of the dataset

	precision	recall	f1-score	support
0	0.82	0.67	0.74	1819
1	0.95	0.98	0.97	15042
macro avg	0.89	0.83	0.86	16861

Overall, the ViT shows impressive results, with an overall accuracy of 93.97%, which outperforms previous models using Convolutional Neural Networks (CNNs) on the same task and

dataset. The model shows a strong ability to detect fake images, as demonstrated by its high precision (0.95) and recall (0.98) for class 1 (fake images), leading to an excellent F1-score of 0.97.

However, the model struggles slightly with identifying real images. While the precision for real images is 0.82, indicating that most images identified as real are indeed real, the recall is lower at 0.67, meaning that around one-third of real images are not correctly identified by the model. This results in a moderately lower F1-score of 0.74 for real images.

The macro average F1-score of 0.86 reflects the performance imbalance between the two classes, with the model excelling at detecting fake images but showing room for improvement in accurately identifying real images. Despite this, the model's high overall performance, particularly for fake images, makes it highly effective for tasks where detecting fake content is the priority.

4 DISCUSSION

4.1 Performance Comparison

When comparing the performance of Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) in classifying deepfake content, several differences emerge. CNNs, which have traditionally excelled in image-related tasks due to their localized feature extraction through convolutional filters, are highly effective in detecting subtle artifacts and spatial inconsistencies that often characterize deepfake images. They are particularly strong in capturing local texture and pixel-level patterns. However, ViTs, which treat images as sequences of patches and use self-attention mechanisms to model both local and global dependencies, offer a different advantage. Their ability to capture long-range relationships between patches allows ViTs to better understand global context and higher-level features, making them potentially more effective in detecting the more sophisticated, globally-consistent manipulations found in advanced deepfake content. In practice, ViTs may outperform CNNs in tasks that require understanding global coherence in images, while CNNs may still excel in tasks focused on fine-grained texture analysis, leading to different trade-offs depending on the nature of the deepfake content being classified.

Another key difference in the performance comparison between Vision Transformers (ViTs) and Convolutional Neural Networks (CNNs) for classifying deepfake content lies in the training processes. In the CNN-based approach, a ResNet50 architecture was employed, which relies on pre-trained weights from models trained on large datasets like ImageNet. ResNet50, with its deep convolutional layers, is highly optimized for capturing hierarchical features, but its training is typically constrained to relatively smaller datasets like ImageNet (1K). On the other hand, the ViT architecture was trained using the larger ImageNet-21k (often referred to as ImageNet_16k), which contains millions more labeled images, providing a much richer and diverse feature set for transfer learning. This more extensive pre-training enables the ViT to capture broader and more complex visual patterns, which can be crucial in deepfake detection, where subtle manipulations may require deeper feature understanding. As a result, the ViT trained on ImageNet-21k is often better at generalizing to novel or unseen deepfake examples compared to ResNet50 trained on a smaller dataset, though it may require more computational resources and time to fine-tune effectively.

4.2 Limitations

One key issue encountered during the classification of deepfake content using the Vision Transformer (ViT) was related to the quality of the frames extracted from the videos. Specifically, during the frame extraction process, some frames that were labeled as part of a fake video actually resembled real, unmanipulated frames. This occurred because certain deepfake techniques did not apply their alterations consistently across all frames. In cases where the subject's face made rapid movements or exhibited complex expressions, the deepfake algorithm struggled to convincingly apply the fake overlay, resulting in certain frames that were essentially untouched by the manipulation. These real-looking frames, despite being part of a fake video, caused confusion during the ViT's training and testing processes, as they did not exhibit the typical artifacts or inconsistencies seen in most deepfakes. This limitation could negatively impact the model's ability to accurately classify deepfakes, as it may have led to misclassification or reduced overall performance, particularly when dealing with more advanced deepfake techniques that struggle with per-frame consistency.

Another significant limitation was the low percentage of real videos in the dataset. The dataset used for training and testing the ViT, was highly imbalanced, with a much smaller proportion of real videos compared to fake ones. This imbalance posed challenges for the model's ability to distinguish between genuine and manipulated content accurately. With fewer real examples to learn from, the model may have become biased toward identifying videos as fake, potentially leading to higher false positive rates.

5 CONCLUSIONS AND FUTURE WORK

This report explored the detection of deepfake content with a focus on the Visual Transformer (ViT) architecture, providing a detailed analysis of its strengths and limitations compared to traditional methods like Convolutional Neural Networks (CNNs). The ViT architecture demonstrated significant potential in detecting deepfakes due to its ability to model long-range dependencies and capture global context across image patches. This allowed the model to identify more sophisticated and globally consistent manipulations often found in deepfake content. However, several challenges were identified, including inconsistent frame manipulation in videos and dataset imbalances, particularly the low percentage of real videos. These limitations affected the ViT's performance, occasionally leading to misclassification and reduced robustness.

While the ViT shows promise in improving deepfake detection accuracy, particularly when trained on large datasets like ImageNet-21k, addressing the issues of frame inconsistency and dataset imbalance will be essential to enhance its effectiveness. Future work should focus on improving the dataset diversity and refining pre-processing techniques, such as better frame extraction methods, to mitigate the impact of unaltered frames in fake videos. Additionally, hybrid approaches that combine the strengths of ViT and CNN architectures could further boost detection capabilities in complex deepfake scenarios.

In addition to refining pre-processing techniques and addressing dataset imbalances, future work should also explore integrating audio content into the deepfake detection process. Many deepfake videos include manipulated audio along with visual alterations, and detecting inconsistencies between the visual and audio streams could provide a more comprehensive and accurate classification system. The introduction of audio analysis, in conjunction with the ViT model, could help in identifying mismatches between lip movements and speech, further improving the model's detection performance.

Moreover, another critical avenue for future work is utilizing the full dataset, as only 10% of the available data was used in this study. Expanding the dataset usage could improve the model's generalization and learning capacity, allowing it to better capture a wider range of deepfake variations and real content. Training on a more diverse and complete dataset would also help mitigate the bias introduced by the current dataset's imbalance and enhance the robustness of the ViT model in real-world deepfake detection scenarios.