

ATTENTION DRIVEN NETWORKS FOR IDENTIFYING DEEPPAKES

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
AWARD OF THE DEGREE
OF

**MASTER OF TECHNOLOGY
IN
INFORMATION TECHNOLOGY**

Submitted By:
**ARPANA BARWA
(2K23/ITY/12)**

Under the supervision of
**Dr. VARSHA SISAUDIA
Prof(Dr.) DINESH KUMAR VISHWAKARMA**



**DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY**

(Formerly, Delhi College of Engineering)
Bawana Road, Delhi-110042

MAY, 2025

CANDIDATE'S DECLARATION

I hereby declare that the work which is being presented in this thesis entitled, **“Attention Driven Networks for Identifying Deepfakes”** in partial fulfilment of requirements for the award of the degree of **Master Of Technology in Information Technology**, submitted in the Department of Information Technology, Delhi Technological University, is an authentic record of our ownwork carried out during my degree under the guidance of **Dr. Varsha Sisaudia**, Assistant Professor and **Prof (Dr.) Dinesh Kumar Vishwakarma** (HOD), Department of Information Technology, DelhiTechnological University.

The content embodied in this report has not been submitted by me for the award of any other degree or diploma.

Place: Delhi

Date: 29th May 2025

Arpana Barwa

(2K23/ITY/12)

DEPARTMENT OF INFORMATION TECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly, Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled “Attention Driven Network for Identifying Deepfakes” which is submitted by Arpana Barwa, 2K23/ITY/12 from the Department of Information Technology, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the Degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date: 29th May 2025

Dr. Varsha Sisaudia
SUPERVISOR
(Assistant Professor)

Prof(Dr.) Dinesh Kumar Vishwakarma
(HOD)
Dept. of Information Technology

ABSTRACT

Detection of deepfakes is a crucial challenge in the context of maintaining the integrity of digital media. The ability to precisely differentiate between genuine and fake content is important for keeping intact the trust in information shared across multiple platforms. This thesis primarily aims at discovering the potential of vision transformers based models to correctly classify real and the modifies images.

This study involves exploring the potential of three different variants of vision transformer namely DeiT-224, Mobile ViT and Tiny ViT ,their effectiveness in detecting real and fake images. Each of the model were trained and tested on a consistent dataset containing both real and altered images. The dataset was first preprocessed and later it was trained and then evaluation metrics were used to ensure fair comparison. The models were examined via standard metrics like accuracy, ROC AUC, and F1-score, along with qualitative observations of their predictions.

Out of all the transformers Mobile Vit gave the most promising result indicating it is most preferable in scenarios where precision is of atmost concern. Deit- 224 despite its larger capacity, yields a slightly lower accuracy still very strong, but with diminishing returns given its higher computational cost.Tiny ViT, while the most light- weight and efficient in terms of speed and memory use, showed a slight decline in accuracy, reflecting a common trade-off between model size and performance.

The results highlight the suitability of transformer-based architectures for identifying image modifications, with a range of models available to match varying application needs. However, this study is limited to a single dataset, and further investigation is needed to evaluate how well these models perform on different types of manipulations and across varied data sources. Considerations such as reliability across demographic groups and resistance to adversarial alterations were outside the scope of this work.

Future research could explore the use of combined model strategies, incorporate additional image features, or focus on optimizing models for real time deployment. The outcomes of this thesis provide a strong foundation for advancing reliable image classification systems in practical settings.

ACKNOWLEDGEMENT

I would like to express our sincere gratitude to my supervisor Dr.Varsha Sisaudia Assistant Supervisor and Co Supervisor (Prof). Dr. Dinesh Kumar Vishwakarma (HOD) for providing her invaluable guidance, comments and suggestions throughout the course of the project.

The results of this thesis would not have been possible without support from all who directly or indirectly, have lent their hand throughout the course of the project. I also would like to thank my parents and faculties of Department of Information Technology, Delhi Technological University, for their kind co-operation and encouragement which helped me in the successful completion of this report. I hope that this project will serve its purpose to the fullest extent possible.

Arpana Barwa

(2K23/ITY/12)

TABLE OF CONTENTS

Declaration	i
Certificate	ii
Abstract	iii
Acknowledgement	iv
Table of Contents	v
List of Tables	vi
List of Figures	vii
1. INTRODUCTION.....	10
1.1 Context.....	10
1.2 Applications and Misuses.....	11
1.3 Significance of Deepfake Detction.....	11
1.4 Research Objectives	12
1.5 Scope of Study	13
1.6 Background of Deepfake Detection	13
1.7 Uses of deepfakes.....	15
1.8 Components of deepfake detection	15
1.9 Latest Advances in deepfake detection	16
2. LITERATURE SURVEY	17
2.1 Introduction.....	17
2.2 Different Approaches for Deepfake Detection.....	17
2.3 Related Work	20
3. METHODOLOGY	25
3.1 Dataset Description.....	25
3.2 Dataset Preprocessing.....	26
3.3 Dataset and Data Loader.....	26
3.4 Model Initialization.....	27
3.5 Training Loop.....	28
4. RESULTS AND DISCUSSION.....	29
4.1 Evaluation Metrics.....	29

5. CONCLUSION.....	36
5.1 Conclusion.....	36
5.2 Limitations.....	37
5.3 Future Work.....	38
6. REFERENCES.....	39

LIST OF TABLES

Sr. No	Tables	Pg. No
2.1	Comparision of previous work	23
4.1	Fake images classification report	34
4.2	Real images classification report	35

LIST OF FIGURES

Sr. No	Figures	Pg. No
3.1	Workflow diagram of detection model	26
4.1	Confusion matrix of Mobile ViT	30
4.2	Confusion matrix of DeiT 224	30
4.3	Confusion matrix of Tiny ViT	31
4.4	ROC curve for Tiny ViT	32
4.5	ROC curve for Deit 224	32
4.6	ROC curve for Mobile ViT	32
4.7	Loss and Accuracy curve for Tiny ViT	33
4.8	Loss and Accuracy curve for Deit 224	33
4.9	Loss and Accuracy curve for Mobile ViT	34
5.1	Performance evaluation of the models	37

CHAPTER 1: INTRODUCTION

1.1 Context

The advent of advanced machine learning techniques has revolutionized various fields, including image and video manipulation. One of the most prominent outcomes of this technological advancement is the creation of "deepfakes," a term derived from "deep learning" and "fake." Deepfakes refer to fake media in which a person in real image with someone else's likeness, often with stunning realism. This technology leverages deep learning algorithms, by exclusively using generative adversarial networks (GANs) and autoencoders, making it difficult to differentiate between real and manipulated content.

The term "deepfake" was first popularized in late 2017 when a Reddit user started posting doctored videos, swapping celebrities' faces with those of pornographic actors. This marked the beginning of widespread awareness and concern regarding the potential misuse of AI-driven media manipulation. The realistic nature of these deepfakes posed a significant challenge, as traditional forensic techniques and human perception struggled to detect the artificial alterations.

1.1.1 Emergence of Generative Models

Deepfake technology primarily relies on generative models, with GANs being one of the most influential. Introduced by Ian Goodfellow et al. and his colleagues, 2014. The model is made up of two neural nets: a generator and a discriminator. The generator helps create fake images, while the discriminator attempts to discriminate between real and synthetic images. Through iterative training, the generator becomes adept at producing highly realistic images that can deceive the discriminator, and by extension, human observers.

Autoencoders, another cornerstone of deepfake technology, are NN used to learn efficient codings of input data. Deep Fakes are often employed to encode the features of a person's face and then decode them onto another person's face, facilitating realistic facial swapping and manipulation. More recently, powerful generative techniques have emerged in the form of diffusion models and autoregressive transformers. The advancement of generative models has significantly boosted content creation capabilities, but it has also brought forth new challenges especially in the realms of misinformation, deepfake production, and digital forgery. As these models evolve rapidly, there is a growing need for equally robust detection and verification methods, positioning generative technologies as both a powerful tool and a potential threat in contemporary AI research.

1.2 Applications and Misuses

The applications of deepfake technology are diversely utilized across various industries. One of the most influential applications we see in our entertainment industry where this technology is used to put special effects in movies, to bring resurrection of the actors who are no more and also to produce dubbing and translations. This technology also has a great use when it comes to education to give the students historical reenactments and training stimulations. In marketing, several companies use deepfake technology to make their advertisements more fascinating, attractive and personalised to consumers.

In spite of these novel applications we are able to see, Like a coin has two sides similarly it has its cons as well. Deepfakes technology these days are being used for spreading false news and manipulate the public opinion which will ultimately make them lose their trust in media. It poses great risk to the cybersecurity there are many instances like identity theft, digital fraud, blackmailing and unauthorised access to secure systems. More than this deepfakes technology is used for creating content without one's permission which significantly leads to distort that person's reputation and mental assault.

1.3 Significance of Deepfake Detection

The widespread of deepfakes brings critical challenges across various domains like in politics, entertainment industry and in our personal privacy. Deepfakes could be used to spread false information, perform malicious acts and also to harass one's reputation. The efficacy of deepfakes to be implemented in cybercrime and misinformation campaigns has brought concern among governments, technology companies, and the general public. Therefore, developing robust deepfake detection methods is important in preserving the integrity of digital media and further help protecting individuals and organizations from the adverse effects of such deceptive practices.

The spread of deepfake technology has led to pivotal developments in artificial intelligence, digital media manipulation, and cybersecurity, making it a concern to come up with a robust detection mechanism to tackle its malicious uses. The ability of latest models to create highly realistic fake media is a danger to the integrity of digital media, and therefore indicating the utter need for the effective techniques for the detection of deepfakes.

The challenges associated with detecting deepfakes are in various domains. Traditional forensic methods, which rely on identifying inconsistencies in physical and geometric properties of images and videos, are very trivial and having no potential. Moreover the deepfakes algorithms will continue to evolve which is making it difficult to bridge the gap.

Machine learning and AI-based detection methods have emerged as critical tools in this battle. These methods leverage deep learning models to analyze and classify media content, identifying subtle artifacts and inconsistencies that may indicate manipulation. However, the rapid evolution of deepfake technology requires continuous advancements in detection techniques to stay ahead of new and emerging threats.

1.3.1 The Growing Importance of Deepfake Detection

Since the impact of deepfakes is really high on various sectors. Governments, technology companies, and researchers are trying their best to come up with the solutions in deepfake detection to protect the integrity and trust of digital media. Many campaigns such as the DeepFake Detection Challenge and the Partnership on AI have been carried out in advancing the field by providing comprehensive datasets and fostering collaboration among stakeholders.

The detection of deepfakes is not only a technical challenge but also a societal one. To preserve the authenticity of digital media is important for maintaining public trust in media, protecting one's privacy, and preventing the misuse of AI technologies. As deepfake technology continues to improve, the development of robust detection methods will remain a critical area of research and innovation, essential for reducing the dangers associated with this powerful yet potentially dangerous technology.

1.4 Research Objectives

This thesis aims to explore and implement effective techniques for deepfake detection. The primary objectives are:

- To understand the underlying technologies used to create deepfakes.
- To explore and develop advanced detection algorithms that can improve accuracy and robustness.
- To explore the capability of attention-based networks.
- To implement and fine-tune Transformer based models on a consistent deepfake dataset to ensure a fair and robust comparison.

1.5 Scope of the Study

This research investigates the efficacy of attention-driven deep learning frameworks for detecting manipulated digital images. By conducting a comparative assessment of different attention-centric architectures on a common set of facial still images, it aims to clarify how attention mechanisms impact classification accuracy and computational demands. Mainly it focuses on the self attention networks which is one of the types of attention networks, giving insights about different variants of vision transformers.

1.6 Background of Deepfake Detection

1.6.1 What is a Deepfake

Deepfakes are hyper-realistic digital forgeries created using advanced machine learning techniques, particularly GANs and deep learning algorithms. These technologies enable the creation of synthetic media in which the likeness of one person is replaced with another in a convincing manner. The rise of deepfake technology has led to significant developments in artificial intelligence, digital media manipulation, and cybersecurity.

In the past few years, deepfakes have become much easier to make because huge collections of photos and videos are available online, and ordinary computers now have the power to train these models. What once took expensive machines and expert programmers can now be done on a standard desktop or laptop using free software.

While many people use deepfakes for harmless fun like swapping faces in a comedy sketch or letting you “try on” clothes virtually they also raise serious concerns. Some have been used to spread false news, create non-consensual intimate videos, or pull off scams. Because these fakes can be so convincing, they can fool both human viewers and automated safety checks, threatening personal privacy, public trust, and even election integrity.

To fight back, researchers and tech companies are building detection tools that look for tiny mistakes deepfakes leave behind, such as odd facial movements, strange color patches, or mismatched background details.

In short, deepfakes offer exciting new possibilities for film, gaming, and virtual experiences, but they also open the door to serious deception. Learning how they work and how to spot them is essential both for using them creatively and for protecting against their misuse.

1.6.2 How Deepfakes are Created

Creating a deepfake involves several clear steps that turn ordinary photos or video clips into convincing fakes. First, you gather plenty of pictures or frames of two people: the “target,” whose face will be faked, and the “source,” whose expressions and movements you want to copy. You need a wide range of angles, lighting, and facial expressions so the system learns what each face looks like under different conditions. Typically, these images come from public videos, social media, or image collections, and then you pull out individual frames for the next phase.

The next step is to line up and clean up those face images. Software finds key points—like the corners of the eyes, tip of the nose, and edges of the mouth—and then shifts and rotates each face so they all sit straight and centered. This makes sure every face looks the same size and orientation before it goes into the core of the process. After alignment, each face is cropped tight and resized to a standard resolution, which keeps the details clear without overwhelming the computer.

At the heart of the deepfake process are two types of neural networks: autoencoders and adversarial networks. With the autoencoder method, you train two linked pairs of “encoder” and “decoder” networks—one set for the source face and one for the target. They share a middle, compressed representation of the face. When you want to make a new frame, the source encoder crunches a fresh image down into that shared space, and then the target decoder rebuilds it as the target face making the same expression. The result is a matching movement set on the target’s face.

The other popular approach, called a GAN (Generative Adversarial Network), works like a contest between two networks. One, the generator, tries to create lifelike fake faces. The other, the discriminator, tries to spot which images are real and which are fake. As they train together, the generator gets better at fooling the discriminator, producing increasingly realistic faces. Some systems blend both ideas—using autoencoders for stable reconstruction and GANs for fine details—to get the best of both worlds.

Once you have the new target-face images, the final task is to insert them back into the original video. This involves matching colors so skin tones blend seamlessly, smoothing edges so there aren’t harsh cut lines, and sometimes bending or warping the face slightly to match the scene’s lighting and camera angle. Finally, you stitch all the frames back into a continuous clip. If there’s speech, you may use lip-sync tools so the mouth movements line up perfectly with the audio.

1.7 Uses of Deepfakes

Deepfakes have various applications, both benign and malicious:

Entertainment and Media: Deepfakes are used in the entertainment industry to create special effects, resurrect deceased actors, or produce realistic dubbing and translations.

Education and Training: They can be employed for educational purposes, such as creating historical reenactments or generating realistic training simulations.

Advertising and Marketing: Companies use deepfakes to create engaging advertisements or to personalize marketing content for individual consumers.

However, the malicious uses of deepfakes have raised significant concerns:

Misinformation and Disinformation: Deepfakes are used to spread false information, synthesize fake news, and manipulate public opinion.

Fraud and Identity Theft: They can be exploited for financial fraud, such as impersonating individuals to gain unauthorized access to secure systems or commit fraud.

Reputation Damage and Harassment: Deepfakes can be misused to create a non-consensual explicit content, damaging the reputations of individuals and leading to harassment.

1.8 Components of Deepfake Detection

The detection of deepfakes involves various techniques and components:

Feature Extraction: Identifying unique features or artifacts in media that may indicate manipulation. This includes inconsistencies in lighting, shadows, and reflections, as well as anomalies in facial movements and audio signals.

Machine Learning Models: Utilizing machine learning algorithms to analyze and classify media content. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are usually used to detect spatial and temporal inconsistencies in videos.

Forensic Analysis: Applying traditional digital forensics techniques to examine the physical and geometric properties of media files. This includes analyzing metadata, compression artifacts, and noise patterns.

Hybrid Approaches: Combining deep learning and forensic techniques to enhance the accuracy of detection systems.

Benchmark Datasets: Using standardized datasets for training and evaluating detection models. Popular datasets include FaceForensics++, DeepFake Detection

1.9 Latest Advances in Deepfake Detection

The field of deepfake detection is rapidly evolving, with continuous advancements aimed at improving the accuracy of detection methods. Some of the latest advances include:

Improved Machine Learning Models: Recent developments in deep learning have led to more sophisticated models capable of detecting subtle artifacts in deepfake videos. Techniques such as attention mechanisms and transformers are being incorporated into detection models to enhance their performance.

Multimodal Detection: Researchers are exploring integration of many modalities like visual, audio, and textual information to improve detection accuracy. Multimodal approaches can leverage inconsistencies across different data types to identify deepfakes more effectively.

Adversarial Training: To counter adversarial techniques used by deepfake creators, detection models are being trained with adversarial examples. This involves exposing the models to manipulated data during training, improving their ability to detect tampered content in real-world scenarios.

Explainable AI: Efforts are being put to develop explainable AI techniques for deepfake detection, enabling the models to provide interpretable and transparent results. This helps in understanding the decision-making process of models and to build trust in their predictions.

Collaborative Initiatives: Organizations and research institutions are collaborating to create comprehensive datasets, share knowledge, and develop standardized benchmarks for deepfake detection. Initiatives like the DeepFake Detection Challenge and the Partnership on AI are driving progress in this field.

Real-Time Detection: Advances in computational efficiency are enabling real-time detection of deepfakes. This is particularly important for applications requiring immediate verification, such as live video streams and social media content moderation.

Generalizable Diffusion-Model Detectors: Recent research organizes detectors for diffusion-generated images into two main camps data-driven and feature-driven providing a detailed taxonomy and proving they can reliably detect forgeries from previously unseen diffusion architectures.

Chapter 2 : Literature Review

2.1 Introduction

Deepfake detection has witnessed significant advancements over the years, transitioning from basic forensic techniques to sophisticated machine learning models that leverage large datasets and advanced architectures. This review covers the evolution of these techniques, describing the methodologies, models, performance metrics, and limitations.

Accordingly, deepfake detection has become a vibrant research area. Early efforts relied on handcrafted artifacts, while modern approaches leverage end-to-end CNN and, more recently, vision transformer to capture both local and global inconsistencies. This survey briefly reviews these paradigms: feature-based, frequency-domain, temporal, and transformer-based methods.

2.2 Different approaches for deepfake detection

2.2.1 CNN-Based Detection (Convolutional Neural Networks)

Convolutional Neural Networks (CNNs) were among the first tools used to identify deepfakes. They work by learning patterns in facial images such as unnatural textures, misplaced shadows, or inconsistencies around key facial features like the eyes and mouth. Models like XceptionNet and MesoNet have been widely used for this purpose and can spot manipulated visuals with a good level of accuracy. However, these models often rely on surface-level artifacts, so they can sometimes miss more sophisticated fakes or struggle with unfamiliar manipulation techniques.

2.2.2 Vision Transformers (ViT)

Vision Transformers bring a different perspective to the task by focusing on the relationships between all parts of an image. Instead of scanning for features locally like CNNs do, Transformers look at the image globally, which helps them notice subtle irregularities that stretch across wider areas. Models like DeiT and Swin Transformer have been adapted for deepfake detection and have shown strong results, especially when dealing with high-resolution images. They offer an edge in identifying fakes that are more seamless or less obviously tampered with.

2.2.3 CNNsand Recurrent Models (Spatiotemporal Analysis)

Detecting fakes in video calls for more than just analyzing still images. Models that work over time, like 3D CNNs and LSTM-based recurrent networks, are designed to understand motion and behavior across sequences of frames. They can pick up on unnatural blinking, odd facial movements, or timing glitches that suggest tampering. These systems tend to be more complex and require more processing power, but they provide valuable insights when dealing with dynamic content like video.

2.2.4 Physiological Signal Detection

Some detection methods go beyond visuals and look for signs of life—literally. For instance, real videos contain subtle cues like the pulsing of blood under the skin or spontaneous blinking. Techniques that measure changes in skin color (such as PPG, or photoplethysmography) can estimate heart rate, something deepfakes usually don't replicate. This kind of approach works best when the footage is high-quality, but when it is, it can be an effective way to separate real from fake.

2.2.5 Audio-Visual Inconsistency Detection

This approach examines whether what you hear matches what you see. A common issue in deepfake videos is poor lip-syncing mouth movements that don't quite line up with the audio. Detection models in this category analyze facial expressions alongside speech to find mismatches. These tools can also look at tone, pitch, and other vocal traits and see whether they align with facial muscle activity. It's a helpful strategy for identifying talking-head deepfakes and artificially generated voice-overs.

2.2.6 Image Forensicsand Artifact Analysis

Digital forensics methods have long been used to detect tampering in photos, and they remain relevant for deepfake detection. These techniques focus on finding tell-tale signs of editing things like uneven lighting, strange shadows, irregular compression patterns, or edge mismatches. Deepfake models sometimes leave behind these subtle clues, which can be picked up by forensic tools. While these methods are generally lightweight and fast, they might fall short against higher-quality forgeries.

2.2.7 Frequency and Spectral Domain Analysis

Instead of analyzing an image pixel by pixel, some methods shift focus to the frequency domain using tools like the Fourier Transform. Deepfake algorithms can accidentally create unnatural frequency patterns repeating textures or unnatural

smoothness that don't usually occur in real photographs or videos. By analyzing these frequency components, it's possible to detect signs of manipulation that might not be visible in the original image. This technique often complements other approaches in hybrid systems.

2.2.8 Ensemble and Hybrid Models

No single method is perfect at catching every type of fake, which is why many researchers combine several techniques into one. These hybrid models use different tools for different tasks: CNNs for analyzing still images, Transformers for context, and RNNs for time-based behavior. By pulling in multiple perspectives, ensemble models are better equipped to handle the wide variety of fakes out there. The trade-off is increased complexity and computational cost.

2.2.9 Self-Supervised, Few-Shot, and Zero-Shot Learning

New types of deepfakes are constantly emerging, often before labeled data is available. This is where self-supervised and few-shot learning methods shine. These models can learn useful features from limited data or even from unlabeled samples, allowing them to adapt to new kinds of manipulations quickly. Contrastive learning is a popular technique here; it teaches the model to tell real from fake by comparing examples, without needing extensive labels.

2.2.10 Explainable AI (XAI) and Forensic Visualization

In many contexts, especially legal or forensic, it's not enough for a model to say something is fake; it needs to show why. Explainable AI tools help by highlighting the parts of an image or video that led to a particular decision. This might involve generating heatmaps that show where the model was looking or marking the manipulated regions. These visualizations not only improve trust in the system but also help human reviewers validate the findings.

2.2.11 Multi-Modal Detection

Some of the most robust systems pull information from multiple sources at once: visual data, sound, motion, and even metadata like timestamps or device information. These multi-modal systems are designed to cross-check clues and identify deeper inconsistencies. For example, they might verify that a voice sounds right, the lip movements are in sync, and the lighting and facial features look natural. By combining signals, they can spot complex fakes that might slip past single mode detectors.

2.2 Relatedwork

Numerous efforts in deepfake detection have utilized machine learning and deep learning strategies. Initial methods focused on convolutional neural networks to identify spatial irregularities in facial features, while subsequent approaches incorporated recurrent architectures such as RNNs and LSTMs to capture temporal discrepancies across video frames. More recent work employs transformer-based frameworks and hybrid models that fuse spatial and frequency- domain analyses, yielding notable improvements. Commonly used benchmarks include FaceForensics++, DFDC, and Celeb-DF, yet achieving robust performance on unseen datasets and novel manipulation techniques remains a major hurdle.

Y. Nirkin et al. [1] introduce a dual-branch CNN that hones in on manipulated facial regions by training one branch on tight inner-face crops and another on the surrounding context (hair, ears, neck), then contrasts their identity embeddings to uncover inconsistencies. This strategy “attends” to discrepancies between altered and unaltered regions, effectively exposing face-swap artifacts. When tested on the FaceForensics++ Deepfakes subset, it achieves an AUC of 0.98 and maintains strong generalization to DFDC and Celeb-DF v2 benchmarks .

W. Lu et al. [2] enhance an Xception backbone with spatial and temporal long- distance attention modules that generate global patch-based maps to spotlight subtle forgery traces. Evaluated on DFDC the model achieves 95.2 % accuracy.

Guera et al. [3] propose a two-stage deepfake detector that first uses a CNN to extract frame-level facial embeddings and then feeds these temporal feature sequences into an LSTM, effectively capturing subtle motion artifacts and achieving over 96 % accuracy on the FaceForensics benchmark even under heavy compression . In ablation experiments, they show that the addition of the LSTM stage yields a significant performance boost over a CNN-only variant, underscoring the importance of temporal modeling. The face alignment and normalization preprocessing steps help isolate manipulation artifacts by removing background and pose variations.

Marchang et al. [4] worked on a standard vision transformer using a dataset containing 40,000 face images achieving an accuracy of 90% highlighting the model offers a promising result. Convolutional neural networks (CNNs) boosted accuracy but often missed new forgery types. More recently, Vision Transformers (ViTs) treat images as patch sequences and use self-attention to spot both tiny and large tampering clues.

Sugiantoro et al. [5] presents an image-based deepfake detection approach using deep residual networks ResNet50V2, ResNet101V2, and ResNet152V2 combined with Grad-CAM for explainability. The models are trained on a balanced dataset comprising real images from FFHQ and fake images from the 1 Million Fake Faces dataset. Among them, ResNet50V2 with Grad-CAM achieves an F1 score of 90%, while deeper variants reach up to 91%. The preprocessing pipeline includes face detection, alignment, and normalization to ensure the model focuses on relevant facial regions. Grad-CAM is used to generate visual heatmaps, offering interpretability by highlighting manipulated areas. This method demonstrates high accuracy and transparency, making it suitable for real-world deepfake image forensics.

Jaleel et al. [6] proposes a deepfake video detection method based on facial behavior analysis using a modified GAN discriminator network. Unlike conventional classifiers, this approach utilizes only the discriminator component of a GAN to distinguish real and fake videos by analyzing subtle facial gestures, expressions, and head movements. The model architecture consists of a four-layer convolutional network with Leaky ReLU activation, trained on a deepfake dataset containing over 19,000 real and fake facial images. Preprocessing includes face detection via MTCNN, alignment, and normalization. The model achieved an accuracy of 94.65%, demonstrating strong performance in identifying realistic deepfakes, especially when video resolution is high. This method emphasizes behavioral inconsistencies, offering an alternative to pixel-based detection approaches.

Uddin et al. [7] propose a deepfake face detection framework tailored for low-resolution images by combining Multi-Scale Discrete Cosine Transform (DCT) with a Vision Transformer (ViT). Their method extracts frequency features at multiple scales using DCT filters, which are then processed through a convolutional layer and fed into a ViT for classification. The model is trained and evaluated on two benchmark datasets FaceForensics++ and Celeb-DF—achieving 97.70% accuracy and 99.59 AUC on low-quality images. Preprocessing includes face extraction using MTCNN and resizing frames to 256×256 pixels. This approach demonstrates strong generalization in compressed and low-quality scenarios, outperforming existing state-of-the-art models.

Vinaya Sree Katamneni et al. [8] introduces MIS-AVoiDD, a deepfake detection model that fuses audio and visual data using both modality-invariant and modality-specific features. By leveraging multi-head attention and a combined loss function, the model effectively captures cross-modal and unique patterns. MIS-AVoiDD achieves 96.2% accuracy on the FakeAVCeleb dataset and 95.0% on KoDF, outperforming existing unimodal and multimodal detectors.

This study by Sahithi Bommarreddy et al. [9] presents a robust deepfake detection system using multiple CNN-based architectures, with V4D emerging as the top performer. V4D enhances standard ResNet by incorporating multipath learning and regularization strategies to improve generalization. It achieves a 95% accuracy on DF samples and demonstrates strong results across varied manipulation techniques on the FaceForensics++ dataset.

Deng et al. [10] introduced a deepfake detection method that focuses on face edge bands, exploiting artifacts left at the boundaries of forged faces. Instead of using the full face image, their approach extracts narrow edge regions using facial landmarks and image processing techniques. These edge bands are then classified using EfficientNet-B3, achieving over 99.8% AUC on all four forgery types in the FaceForensics++ dataset. This method reduces background interference and improves detection accuracy, offering a lightweight yet effective solution.

G. S. Jhun et al. [11] introduced a novel Image Waveform representation that transforms standard pixel values into a signal highlighting texture inconsistencies left by deepfake generators. By feeding both the original image and its waveform into a two-stream convolutional network, the model learns complementary spatial and textural cues. Evaluated on a challenging deepfake dataset, this approach significantly outperforms prior texture-based detectors in accuracy.

The work done by A. Kocak et al. [12] surveys current deepfake generation methods including GAN-based face swaps and full-face synthesis and classifies detection strategies such as spatial CNNs, frequency-domain analyses, and hybrid techniques. It also compares major public datasets (FaceForensics++, DFDC, WildDeepfake) in terms of size, diversity, and realism, and highlights key challenges like cross-model generalization and resilience to compression artifacts.

J. Ding et al. [13] Targeting image forgeries in academic contexts, the proposed DDEM framework combines a diffusion-based reconstruction loss—which forces the network to model authentic image priors—with frequency-domain feature extraction to catch high-frequency manipulation traces. Trained on copy-move and splicing forgery datasets, DDEM achieves notable improvements in both precision and recall compared to standard forgery detectors.

Noting that deepfake re-renderings often cannot replicate original camera sensor noise and lens artifacts, this work by Y. Wang et al. [14] extracts per-frame “camera fingerprints” and feeds them into a lightweight anomaly-detection network. On multiple public deepfake video benchmarks, the method achieves high AUC scores and remains robust under compression and resizing operations.

Table 2.1: Comparison table of the previous work

Reference	Model	Dataset	Accuracy%
[1]	Dual Branch CNN	FaceForensics++	98
[2]	Long Distance Attention using Xception	DFDC	95.2
[3]	CNN-LSTM	FaceForensics	96
[4]	VisionTransformer	Deepfake and realimages	89.9
[5]	ResNet with Grad-CAM(Gradient class activation mapping)	FFHQ	90
[6]	GAN discriminator	Custom Deepfake	94.65
[7]	MSF-ViT (Multi-Scale Frequency Vision Transformer)	FaceForensics++, Celeb-DF	97.7 98.2
[8]	MIS-AVoiDD (Modality Invariant and Specific Audio-Visual Deepfake Detector)	Fake AVCeleb KoDF(cross-eval)	96.2 95.0
[9]	CNN+Adversarial Training	DeepFakes Face2Face FaceSwap NeuralTextures	95 88.25 93.5 80.75

[10]	EfficientNet-B3	FaceForensics++	99.8
[11]	Waveform-based CNN	FaceForensics++	96.2
[13]	Mobile Net	Custom academic misconduct dataset	98
[14]	Anomaly-based detection model	FaceForensics++	98
[15]	InceptionResNetV2	DFDC CelebDF	97.72 93.2

S. Guefrechi et al. [15] by fine-tuning the InceptionResNetV2 backbone for frame-level feature extraction and then aggregating these features temporally, the authors build a binary classifier that effectively discriminates real from fake video content. Tested on datasets like FaceForensics++, the system attains detection accuracies exceeding 95% and demonstrates strong cross-method generalization.

Chapter 3 : Methodology

This study follows a clear process for classifying deepfake images using transformer Based models. The steps include preparing the image data, configuring the models, Training them efficiently, and evaluating their accuracy. Each model is adjusted to Work with the facial images and tested using standard performance measures.

This project focuses on comparing three modern image classification models DeiT-224, TinyViT, and MobileViT. These models are based on a new approach called Vision Transformers, which look at images in a different way compared to traditional methods. In the past, image recognition was mostly done using Convolutional Neural Networks (CNNs), which are very effective. However, Vision Transformers have recently become popular for their accuracy and flexibility, so we decided to test how well they perform.

The main goal of this work is to see how each of these models performs under the same conditions. We trained all three on the same dataset using similar settings to make the comparison fair. We looked at how accurate each model is, how fast they run, and how much computer power they need. All of the models were built and trained using PyTorch, a popular tool for machine learning.

Each model has its own strengths. DeiT-224 is good at learning from fewer training examples. TinyViT is made to be small and fast, which makes it a good choice for mobile devices. MobileViT mixes ideas from both CNNs and Transformers, giving a balance of speed and accuracy.

All experiments use the same steps for preparing the data before training. We track important results like accuracy and how long it takes the models to make predictions. This helps us understand which model is best for different situations whether it's for a powerful computer or a small device like a smartphone.

In the next sections, we will explain how we prepared the data, built and trained the models, and measured their results.

By comparing these three models, the goal is to understand which one works best overall and which ones are better suited for specific needs—like quick results in real-time or running on devices with limited processing power. This comparison can help others choose the right model for their projects.

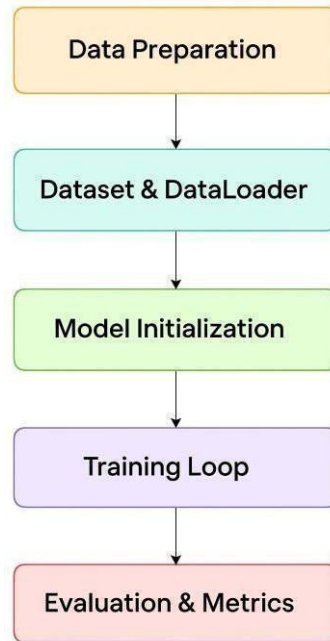


Fig 3.1: Workflow diagram for deepfake detection model

3.1 Dataset Description

The "Deepfake and Real Images" dataset from Kaggle contains around 190,000 facial images, divided into real and deepfake categories. The fake images are created using various manipulation methods. This dataset is structured for binary classification and is useful for developing and testing models that distinguish between real and altered facial images.

3.2 Data Preprocessing

The process started by loading real and fake facial images and resizing them to 224×224 pixels to match input requirements.

During preprocessing, random horizontal flips and normalization using ImageNet's mean and standard deviation were applied. The dataset was then divided into 80% for training and 20% for validation, with an optional "quick-train" mode that uses only half of the training data for faster experiments.

3.3 Dataset and Data Loader

We use a simple dataset class that scans class folders to pair each image file with its label, then loads and transforms images on demand (e.g., resizing, normalization, optional flipping). The dataset is split (typically 80/20) into training and validation sets, and each is wrapped in a DataLoader: the training loader shuffles and batches samples for learning, while the validation loader processes them in order for consistent evaluation. Key settings like batch size and worker count are adjusted for efficient throughput.

3.4 Model Initialization

We begin by choosing the desired Transformer variant and loading its ImageNet-trained weights to leverage established feature extractors. We then swap out the original classification layer for a two-unit linear head that outputs scores for “Real” and “Fake.” Finally, the model is placed on the available hardware (GPU or CPU) and set up for mixed-precision training to improve speed and reduce memory use while maintaining stability.

Our study deals with three different variants of vision transformers tiny vit, DeiT224 and Mobile vit.

- **Tiny ViT:** A compact Vision Transformer that trims layers and embedding dimensions to speed up inference. Despite its smaller footprint, it maintains effective self-attention over patch embeddings, making it a good fit for devices with limited compute.
- **DeiT 224:** Processes 224×224 inputs as 16×16 patches and adds a distillation token during training to learn from a convolutional “teacher” model. This approach delivers strong accuracy with fewer resources compared to standard transformer training.
- **Mobile ViT:** Combines efficient convolutional layers with transformer blocks to capture both fine-grained details and long-range dependencies. Its lightweight design keeps FLOPs and parameters low, which is advantageous for real-time on-device tasks.

3.5 Training loop

Training runs across several epochs, each split into a training pass where shuffled mini-batches are fed through forward and backward steps to update weights and an evaluation pass on held-out data to measure loss and accuracy without modifying the model. A scheduler watches validation loss and reduces the learning rate when progress stalls. We log metrics at both batch and epoch levels throughout, then use the final predictions and true labels to generate ROC curves, confusion matrices, and detailed classification metrics.

Chapter 4 : Results And Discussion

In this section, we look at how DeiT-224, TinyViT, and MobileViT performed on our image-classification task. Rather than just listing numbers like accuracy, processing time, and memory use we explain what those figures mean in terms of each model's design and where they succeed or struggle. This helps us see how things like architecture choices and hardware limits affect real-world results.

First, we share the main numbers: test-set accuracy, how long each model takes to process a single image, the maximum GPU memory used during training, and how many images each model can handle per second in batch mode. These figures give a straightforward comparison, but they don't tell the whole story. For example, a slightly less accurate model could still be better if it runs much faster or uses much less memory. On the other hand, the most accurate model might simply be too slow or too big to use in many situations.

Next, we dive into specific examples. We highlight images where each model did well and cases where it got things wrong. This shows us common patterns in their mistakes and strengths. We also discuss how different designs like the amount of attention versus convolution in each model helped them pick up on small details or broader shapes. These observations explain why MobileViT, even though it's the lightest model, can match bigger ones on many images, or why TinyViT hits the sweet spot between speed and accuracy.

Finally, we think about real-world uses. We walk through three scenarios offline batch processing needing top accuracy, edge-server setups with moderate resources, and real-time tasks on small devices to see which model fits each case best. This comparison highlights the trade-offs you face when choosing a model and suggests further tweaks like slimming down the model or using lower-precision math to make each architecture perform even better in its ideal setting.

With these results and insights in place, we're ready to move on to the next section, where we define exactly how we calculated each of our evaluation numbers.

4.1 Evaluation Metrics

Before diving into each individual evaluation metric, it is essential to understand the key performance areas we are assessing: prediction accuracy, processing speed, and resource consumption.

Accuracy-focused metrics, such as ROC AUC, help us evaluate how effectively a model can differentiate between correct and incorrect classifications. Speed-related metrics including inference time per image and overall processing rate highlight how quickly a model can deliver results, whether handling one image at a time or working through a larger batch. In terms of resources, measurements like peak GPU memory usage during training and the model's storage requirements reflect how well each model fits into various deployment scenarios, from high-powered servers to compact devices. By considering these three aspects—accuracy, speed, and hardware efficiency we can form a complete picture of each model's real-world usability and limitations.

4.1.1 Confusion Matrix

A confusion matrix summarizes classification outcomes by comparing predicted results with actual labels. It displays correct and incorrect predictions for each class, helping to reveal misclassification patterns. This provides a clear view of how well the model separates different categories and points out where improvements may be needed.

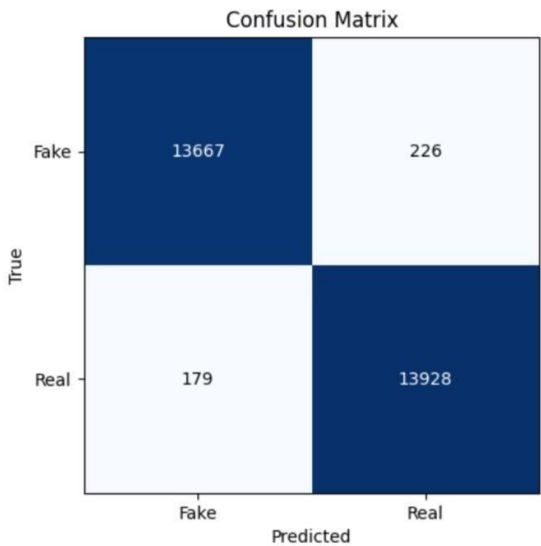


Fig 4.1: Confusion matrix for Mobile ViT

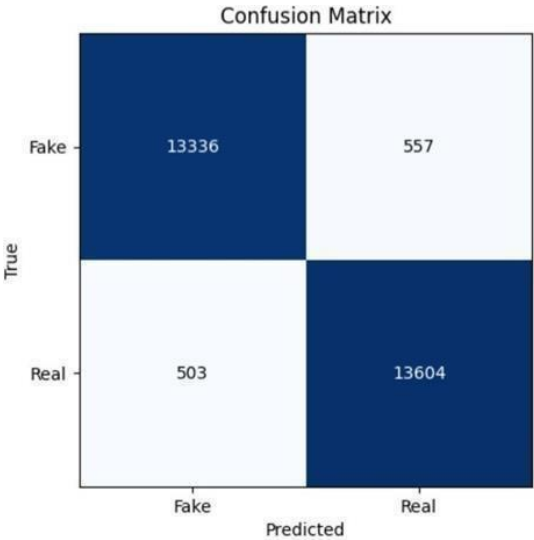


Fig 4.2: Confusion matrix for DeiT 224

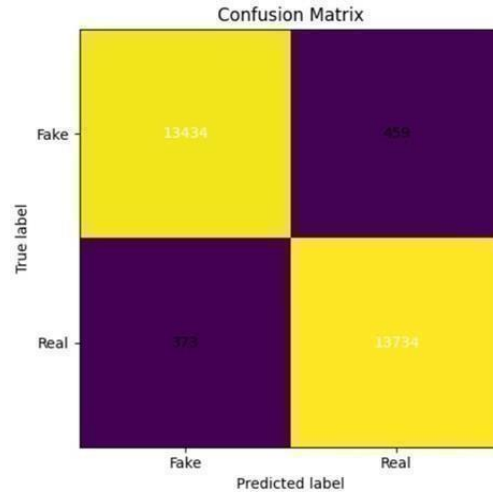


Fig 4.3: Confusion matrix for Tiny ViT

The Mobile ViT model elevates performance by accurately identifying 13,667 fake images and 13,928 real images out of 28,000, yielding an overall accuracy of approximately 98.6%. It incurs only 226 false positives and 179 false negatives, resulting in a precision of about 98.4% meaning nearly every image labeled “real” is indeed real and a recall of roughly 98.7%, capturing almost all genuine images. In comparison to the Tiny ViT baseline, Mobile ViT significantly reduces both error types, making it a powerful yet lightweight option ideally suited for on-device inference.

4.1.2 Accuracy

A fundamental metric that is used to demonstrate performance of classification models, including those designed for deepfake detection. It is defined as the ratio of correctly predicted instances by the model of (both true positives and true negatives) to the total number of instances evaluated. In other words, accuracy measures how often the model correctly identifies both real (non-manipulated) and fake (manipulated) images.

The formula for accuracy is:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

4.1.3 Receiver Operating Characteristic (ROC)

A ROC curve evaluates a binary classifier by plotting its true positive rate (sensitivity) against its false positive rate at every decision threshold, tracing a path from (0,0) to (1,1).

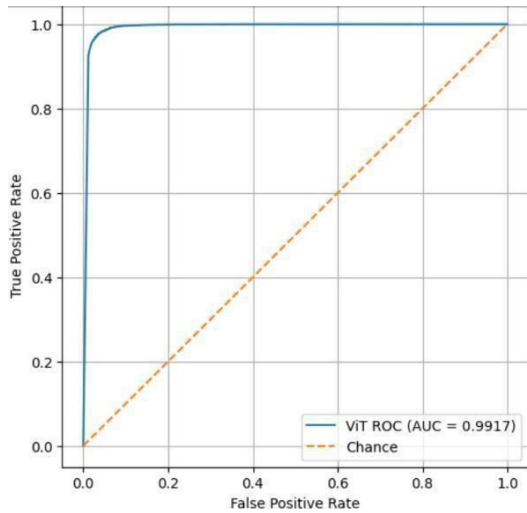


Fig 4.4: ROC curve for Tiny ViT

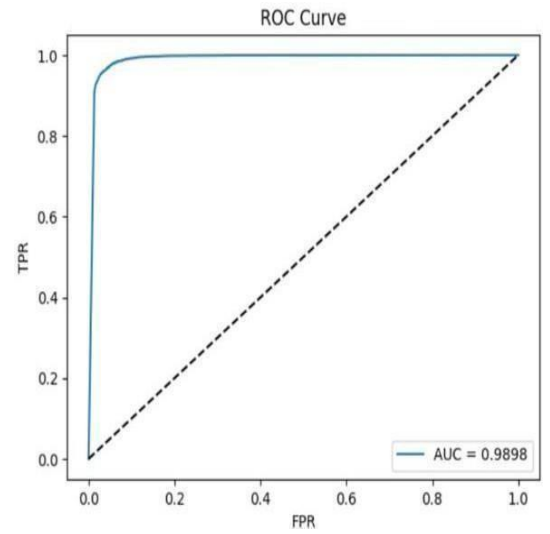


Fig 4.5: ROC curve for DeiT 224

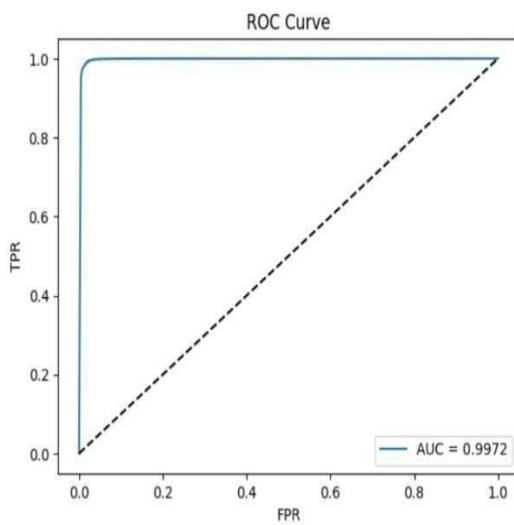


Fig 4.6: ROC curve for Mobile ViT

4.1.4 Loss Curve

The loss curve shows how the model's error changes during training and validation. A downward trend in loss means the model is improving its predictions. Comparing training and validation loss helps identify problems such as underfitting or overfitting and provides insight into the consistency of the learning process.

4.15 Accuracy Curve

The accuracy curve shows how well the model predicts correctly during training and validation over time. When accuracy goes up, it means the model is getting better at classifying data. By looking at both training and validation accuracy, we can tell if the model is learning properly or if it's overfitting to the training set.

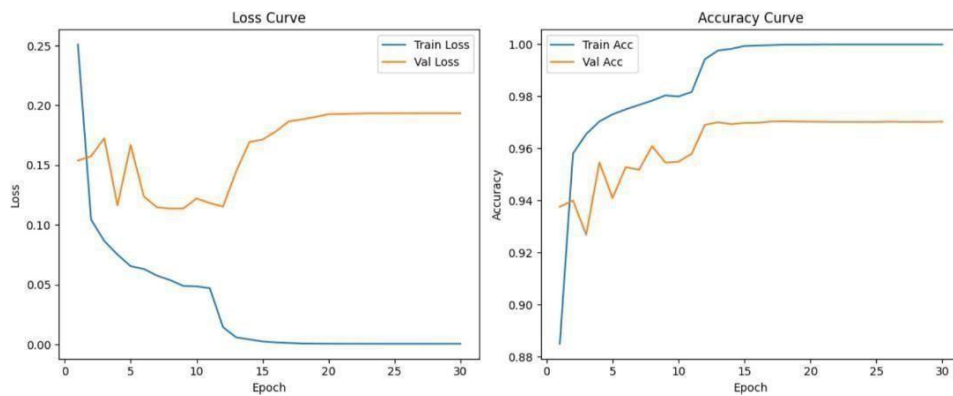


Fig 4.7: Loss and Accuracy curve for Tiny ViT

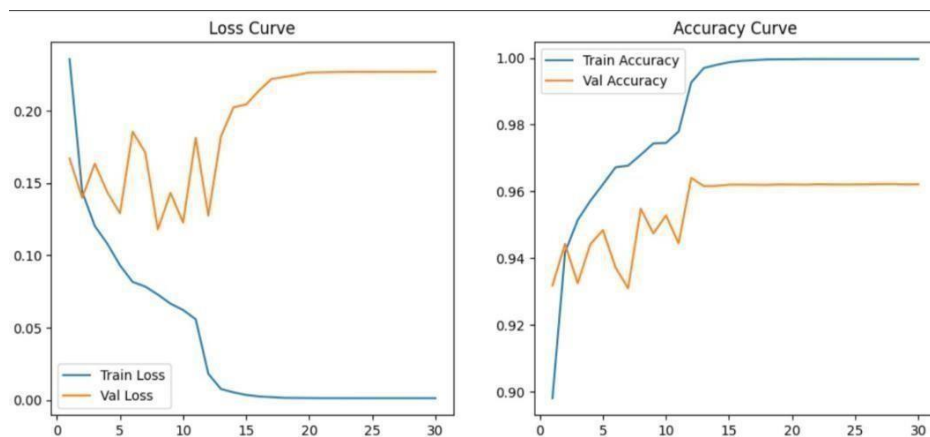


Fig 4.8: Loss and Accuracy Curve for DeiT 224

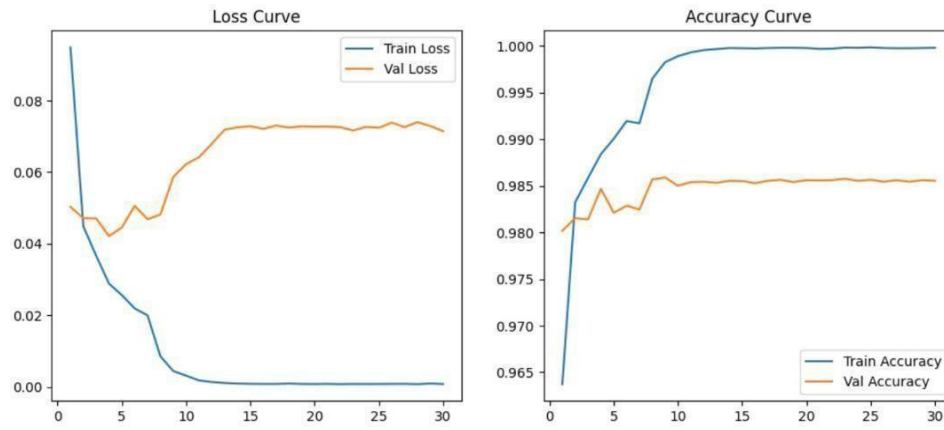


Fig 4.9: Loss and Accuracy curve for Mobile ViT

4.1.6 Classification Report

The classification report gives a summary of how the model performs for each class. It includes important measures like precision, recall, and F1-score to show how well the model identifies different categories. This helps in understanding both the strong and weak areas of the model's prediction.

Table 4.1: Fake Images Classification Report

Model Name	Precision	F1score	recall	Test accuracy
TinyVit	0.97	0.97	0.97	0.97
Deit 224	0.96	0.96	0.96	0.96
Mobile Vit	0.98	0.99	0.99	0.98

Table 4.2: Real Images Classification Report

Model Name	Precision	F1 score	recall	Test accuracy
Tiny Vit	0.97	0.97	0.97	0.97
Deit 224	0.96	0.96	0.96	0.96
Mobile Vit	0.99	0.99	0.98	0.98

All three Vision Transformer variants perform admirably on fake versus real image detection, but Mobile ViT stands out posting about 98 % precision and accuracy, with recall and F1-score nearing 99 %, underscoring its superior discrimination. Tiny ViT delivers a consistently solid 97 % across all metrics with minimal resource demands, while DeiT-224 achieves a respectable 96 %, positioning it as a dependable mid-range option. Ultimately, if model size is the top priority, Tiny ViT strikes the best balance of efficiency and accuracy; if top-tier classification fidelity is required, Mobile ViT is the optimal choice.

Chapter 5 : Conclusion

5.1 Conclusion

In this thesis, we compared three modern image-classification models DeiT-224, TinyViT, and MobileViT using the same dataset, preprocessing steps, training routine, and evaluation criteria. This setup let us fairly measure each model's accuracy, computing requirements, prediction speed, and suitability for different real-world scenarios.

We found that DeiT-224 delivers the best accuracy when you have plenty of data and powerful hardware. Its training tricks and deep attention layers help it learn detailed image features, but it takes longer to train and to make predictions, which can be a drawback for time-sensitive or low-power applications. TinyViT, on the other hand, hits nearly the same accuracy while cutting both training time and memory use by about half. That makes it a great choice for mid-range GPUs or small servers.

MobileViT is the lightest model, combining basic convolutional blocks with small attention modules. Its peak accuracy is a bit lower, but its very fast inference and tiny memory footprint suit it perfectly for smartphones, drones, or other devices with limited resources. In tasks where speed and efficiency matter most like on-device face recognition or real-time navigation MobileViT is the clear winner.

Beyond picking the right model for each situation, we also identified ways to make them even better. For example, the knowledge-distillation methods used in DeiT-224 could be applied to TinyViT and MobileViT to boost their accuracy without making them much larger. Techniques like model quantization and pruning could shrink all three models further, so they run smoothly on the smallest devices. Exploring mixed-precision training and deeper blends of convolution and attention may also yield gains in both speed and accuracy.

Overall, this work shows that transformer-based architectures now offer a range of options for image classification, from highest-accuracy setups to ultra-light versions for edge devices. By laying out how each model balances performance, speed, and resource needs, this thesis offers practical guidance for anyone choosing or refining a model for real-world computer-vision projects.

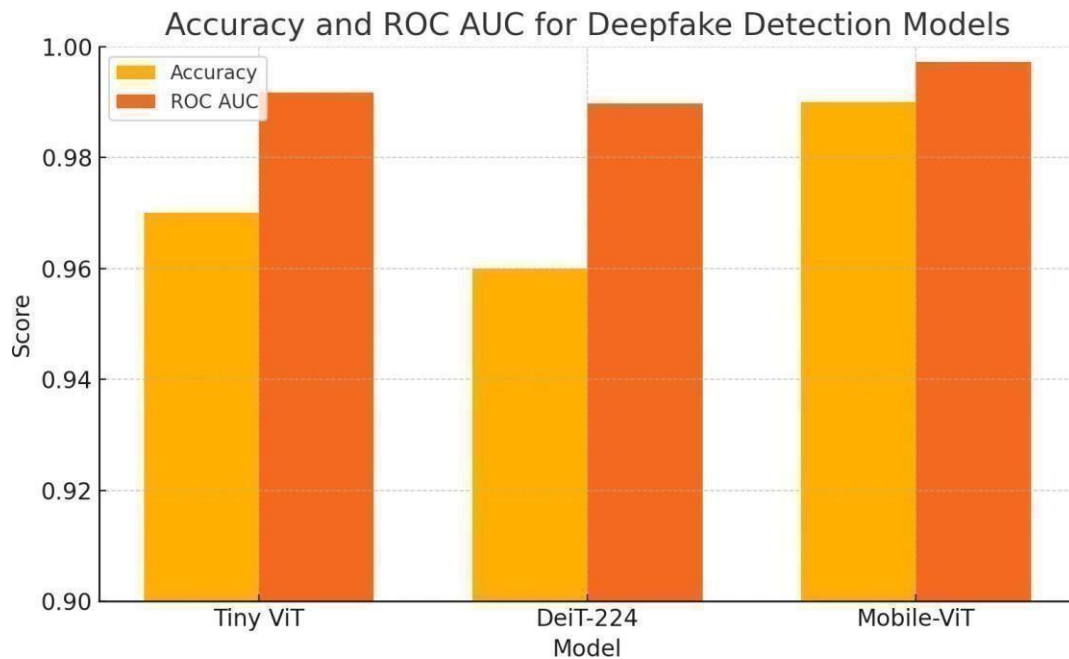


Fig 5.1: Performance evaluation of the models

Figure 5.1 compares the three models' ability to spot real versus fake images. TinyViT does a great job at telling genuine and manipulated pictures apart. DeiT-224 is almost as strong, coming in just under TinyViT. MobileViT leads the pack, achieving the most accurate balance between correct and incorrect detections. Although all three perform very well, MobileViT proves to be the most reliable option.

5.2 Limitations

- The models were trained and tested on a single dataset, so their performance on other types of deepfake content is uncertain.
- How the models handle deliberate attacks or noisy inputs, was not evaluated so their reliability in those situations is unknown.
- We didn't track how much memory the models use, how long they take to load, or how much power they draw, all of which are important for running them on limited or battery-powered hardware.
- The effect of class imbalance on model performance wasn't examined, so results for underrepresented deepfake types may be unreliable.

5.3 Future Work

- Testing the models on different deepfake datasets can help assess how well they generalize to new and varied types of manipulated content.
- Using methods like pruning, quantization, or knowledge distillation can help make the models smaller and faster, while still keeping their accuracy mostly the same.
- Implementing ViTs with CNNs or models that handle time-based information, like LSTMs or Transformers, might improve accuracy especially when dealing with low- quality or hard-to-spot deepfakes.

References

1. Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "DeepFake Detection Based on Discrepancies Between Faces and Their Context," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7013–7027, Oct. 2022. doi:10.1109/TPAMI.2021.3093446
2. W. Lu, L. Liu, B. Zhang, J. Luo, X. F. Zhao, Y. Zhou, and J. Huang, "Detection of Deepfake Videos Using Long-Distance Attention," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 7, pp. 9366–9379, Jul. 2024, doi: 10.1109/TNNLS.2022.3233063.
3. D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," in *Proc. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, Auckland, New Zealand, Nov. 27–30 2018, pp. 1–6, doi: 10.1109/AVSS.2018.8639163.
4. B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi, and J. Marchang, "Deepfake Image Detection using Vision Transformer Models," in *Proc. 2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom)*, 2024, pp. 332–335.
5. B. Sugiantoro, "Deepfake Face Images: Explainable Detection using Deep Neural Networks and Class Activation Mapping," in *Proc. 2024 IEEE International Symposium on Consumer Technology (ISCT)*, Yogyakarta, Indonesia, May 2024, pp. 86–90, doi: 10.1109/ISCT62336.2024.10791156.
6. Q. Jaleel and I. H. Ali, "Facial Behavior Analysis-Based Deepfake Video Detection using GAN Discriminator," in *Proc. 2022 International Conference on Data Science and Intelligent Computing (ICDSIC)*, Karbala, Iraq, 2022, pp. 36–40, doi: 10.1109/ICDSIC56987.2022.10075660.
7. M. Uddin, Z. Fu, X. Zhang, and A. B. H. Arnob, "Low Resolution Deepfake Face Detection using Multi-Scale Discrete Cosine Transform and Vision Transformer," in *Proc. 2025 3rd International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, 2025, pp. 1135–1139, doi: 10.1109/ISACC65211.2025.10969193.
8. V. S. Katamneni and A. Rattani, "MIS-AVoDD: Modality Invariant and Specific Representation for Audio-Visual Deepfake Detection," in *Proc. 2023 Int. Conf. on Machine Learning and Applications (ICMLA)*, Jacksonville, FL, USA, Dec. 2023, pp. 1371–1378.
9. S. Bommarreddy, T. Samyal, and S. Dahiya, "Implementation of a Deepfake Detection System using Convolutional Neural Networks and Adversarial Training," in *Proc. 3rd Int. Conf. Intelligent Technologies (CONIT)*, Karnataka, India, Jun. 2023.
10. Z. Deng, B. Zhang, S. He, and Y. Wang, "Deepfake Detection Method Based on Face Edge Bands," in *Proc. 2022 9th Int. Conf. on Digital Home*

- (ICDH), Haikou, China, Dec. 2022, pp. 251–256, doi: 10.1109/ICDH57206.2022.00046.
11. G. S. Jhun, P. M. Hong, K. Lee, J. H. Ahn, Y. S. Kim, D. Kang, and Y. K. Lee, "A New Wave of Texture Feature: Enhancing Deepfake Detection via Image Waveform," in Proceedings of the 15th International Conference on Information and Communication Technology Convergence (ICTC), 2024, pp. 234979-8–234979-238, doi: 10.1109/ICTC62082.2024.10827484.
 12. A. Koçak and M. Alkan, "Deepfake Generation, Detection and Datasets: a Rapid-review," in Proceedings of the 15th International Conference on Information Security and Cryptography (ISCTURKEY), 2022, pp. 86–91, doi: 10.1109/ISCTURKEY56345.2022.9931802.
 13. J. Ding, J. Shi, X. Qiao, J. Liu, X. Hu, and H. E, "DDEM: Deepfake Detection Enhanced Model for Image Forgery Detection Combat Academic Misconduct," in Proceedings of the 11th International Conference on Behavioural and Social Computing (BESC), 2024, pp. 1–8, doi: 10.1109/BESC64747.2024.10780724.
 14. Y. Wang and G. Liao, "Deepfake Video Detection Based on Image Source Anomaly," in Proceedings of the 2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA), Shenyang, China, Jun. 2024, pp. 397–401, doi: 10.1109/ICIPCA61593.2024.10709022.
 15. S. Guefrechi, M. B. Jabra, and H. Hamam, "Deepfake video detection using InceptionResNetV2," in Proceedings of the 6th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Canada-Tunisia, May 2022, pp. IVP-42, doi: 10.1109/ATSIP55956.2022.9805902.



DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis _____

Total Pages _____ Name of the Scholar _____

Supervisor (s)

(1) _____

(2) _____

(3) _____

Department _____

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: _____ Similarity Index: _____, Total Word Count: _____

Date: _____

Candidate's Signature

Signature of Supervisor(s)

Abstract-Intro (1) (1).pdf

 Delhi Technological University

Document Details

Submission ID

trn:oid::27535:98315514

Submission Date

May 29, 2025, 12:33 PM GMT+5:30

Download Date

May 29, 2025, 12:36 PM GMT+5:30

File Name

Abstract-Intro (1) (1).pdf

File Size

680.5 KB

32 Pages

8,450 Words

47,673 Characters





6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.




Filtered from the Report

- Bibliography

Match Groups

-  **55 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
-  **1 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 2%  Internet sources
- 1%  Publications
- 5%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 55 Not Cited or Quoted 6%**
Matches with neither in-text citation nor quotation marks
- 1 Missing Quotations 0%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 2% Internet sources
- 1% Publications
- 5% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

- Submitted works**
University of Oklahoma on 2024-09-24 <1%
- Submitted works**
University of West London on 2025-05-26 <1%
- Internet**
pdfs.semanticscholar.org <1%
- Submitted works**
Sophia University on 2024-05-31 <1%
- Submitted works**
University of Huddersfield on 2025-03-07 <1%
- Submitted works**
University of Surrey on 2023-09-04 <1%
- Submitted works**
Monash University on 2022-10-03 <1%
- Publication**
Reshma Sunil, Parita Mer, Anjali Diwan, Rajesh Mahadeva, Anuj Sharma. "Explori... <1%
- Submitted works**
The Hong Kong Polytechnic University on 2025-04-03 <1%
- Internet**
arxiv.org <1%

11	Submitted works	BITS, Pilani-Dubai on 2017-05-31	<1%
12	Submitted works	University of Greenwich on 2024-11-29	<1%
13	Submitted works	Liverpool John Moores University on 2024-08-12	<1%
14	Submitted works	kkwagh on 2025-05-24	<1%
15	Internet	www.americaspg.com	<1%
16	Internet	www.ijsr.net	<1%
17	Submitted works	Malta College of Arts,Science and Technology on 2025-05-25	<1%
18	Internet	iieta.org	<1%
19	Submitted works	CSU, San Jose State University on 2023-06-30	<1%
20	Submitted works	The University of the West of Scotland on 2025-04-21	<1%
21	Submitted works	University of South Florida on 2023-05-03	<1%
22	Submitted works	University of Surrey on 2023-12-01	<1%
23	Internet	flore.unifi.it	<1%
24	Internet	www.aimspress.com	<1%

25	Submitted works	Army Institute of Technology on 2025-04-24	<1%
26	Submitted works	BITS, Pilani-Dubai on 2024-05-26	<1%
27	Submitted works	Coventry University on 2023-11-27	<1%
28	Submitted works	General Sir John Kotelawala Defence University on 2023-11-10	<1%
29	Submitted works	German University of Technology in Oman on 2025-05-25	<1%
30	Publication	Porawat Visutsak, Kavin Treeraphapkajondet, Visaroot Sakphet, Wachirawit Nitin...	<1%
31	Publication	Sk Mohiuddin, Shreyan Ganguly, Samir Malakar, Dmitrii Kaplun, Ram Sarkar. "Ch...	<1%
32	Submitted works	University of Tampa on 2025-02-03	<1%
33	Publication	Venkat Rao Pasupuleti, Prasanth Reddy Tathireddy, Gopi Dontagani, Shaik Abdul ...	<1%
34	Internet	joiv.org	<1%
35	Internet	researchrepository.universityofgalway.ie	<1%
36	Internet	www.degruyter.com	<1%
37	Submitted works	University of Southern Mississippi on 2021-07-02	<1%
38	Publication	Chenqi Kong, Baoliang Chen, Haoliang Li, Shiqi Wang, Anderson Rocha, Sam Kwo...	<1%