National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Alana Griffiths Kelly
Student ID:	X23302003
Programme:	MSc Cybersecurity (Part-Time) **Year:** 2025
Module:	MSc (Research) Practicum
Lecturer:	Supervisor – Michael Pantridge
Submission Due Date:	11/12/2025
Project Title:	Deepfake Detection Using a Fusion of Convolutional Neural Networks and Vision Transformers
Word Count:	4503 **Page Count:** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Alana Griffiths Kelly

……

11/12/2025

Date: ……

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Deepfake Detection Using a Fusion of Convolutional Neural Networks and Vision Transformers

Alana Griffiths Kelly – x23302003
Supervisor: Michael Pantridge
National College of Ireland

1. Abstract

Deepfakes present a growing threat to digital media authenticity and cybersecurity, with existing detection models often struggling to generalise across datasets or withstand real-world distortions. This study develops a hybrid Deepfake detection pipeline that combines Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to better capture both local visual artefacts and global structural inconsistencies. A complete preprocessing workflow was implemented, including video frame extraction, face detection using MTCNN, and image normalisation. The proposed model fuses EfficientNet-B0 and ViT-B/16 feature embeddings to perform frame-level classification, with video-level predictions derived from aggregated softmax probabilities. Evaluation on real video footage demonstrates that the hybrid architecture is functional, reproducible, and capable of producing stable and interpretable Deepfake likelihood scores despite not being trained on Deepfake-specific datasets. The study highlights the complementary strengths of CNN and transformer-based features and identifies key limitations such as lack of temporal modelling and dataset-specific training. The findings provide a foundation for future research into multimodal, temporally-aware, and more robust Deepfake detection systems.

2. Introduction

2.1. Background and Context

Deepfakes are synthetic audio-visual media generated using deep learning and have emerged as a significant cybersecurity threat. Deepfakes can convincingly fabricate faces, speech, and behaviours. While these techniques have legitimate applications, their misuse has been documented in misinformation campaigns, identity fraud, political manipulation, and social engineering (Europol, 2020). The increasing accessibility of generative tools has intensified concerns over public trust, where the existence of Deepfakes allows malicious actors to dismiss authentic evidence.

AI-based Deepfake detection has become a critical defence mechanism. Current detectors rely predominantly on CNN and transformer architectures, which learn spatial, frequency, or temporal inconsistencies in manipulated media. However, they face persistent challenges. First, detectors are highly sensitive to noise, compression, and real-world degradations, leading to substantial performance drops outside controlled environments (Rossler, et al., 2019). Second, detectors frequently fail to generalise across datasets; models trained on FaceForensics++ rarely transfer well to datasets such as DFDC or Celeb-DF due to differing manipulation methods and distributions (Dolhanskey, et al., 2020) (Li, et al., 2020). Recent work has explored alternative forensic cues, including physiological signals such as photoplethysmography (PPG), which Deepfakes often fail to reproduce accurately (Li, et al., 2020). These complementary features highlight opportunities for multi-modal or hybrid detection approaches.

2.2. Research Problem

Deepfake has become a serious threat to digital media authenticity and cybersecurity. Advanced generative models are used to copy real people's faces and voices and synthesize them for fraudulent purposes, these models are difficult to detect. Convolutional Neural Networks (CNN) are deeply relied on to detect traditional Deepfakes. CNNs are good ate detecting Deepfakes but have limited generalization on new datasets.

There are multiple gaps, or problems, in the research, or detection, of Deepfakes. There is dataset bias, as previously stated, where CNN models tend to only perform well on the dataset that they were trained on and fail to generalize unseen data, therefore reducing performance when used in a cross-dataset manner. Attacks such as compression, noise, and adversarial can fool detection models which cause vulnerabilities in real world distortion. Transformer based models perform better across datasets, however these are harder, and more expensive, to deploy in real time.

Proposed solutions, and the focus of the research, are as follows. The main solution would be to develop a hybrid Convolutional Neural Network-Visual Transformer (CNN-ViT) model supported by a structured preprocessing pipeline. The design of this model aims to improve feature representation quality, stabilise

video-level predictions, and provide a foundation for future extensions to temporal or multimodal analysis.

The expected outcome of this research is a functional and generalisable hybrid Deepfake detection system that demonstrates improved artefact recognition and consistent performance under realistic conditions, contributing to the broader goal of enhancing digital content authenticity and cybersecurity defence.

2.3. Research Aim and Objectives

2.3.1. Research Aim

The aim of this research is to develop and evaluate a hybrid Deepfake detection system that combines Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to improve the identification of manipulated facial imagery in video content. The study seeks to determine whether a fused CNN–ViT architecture, supported by a structured preprocessing pipeline, can enhance robustness to diverse Deepfake artefacts and improve generalisation across varying manipulation styles.

2.3.2. Research Objectives

To achieve this aim, the research will pursue the following objectives.

2.3.2.1. Develop a complete preprocessing workflow

Implement a video-to-frame processing pipeline—including frame extraction, face detection using MTCNN, and image normalisation—to enable consistent and reliable face-level analysis across videos.

2.3.2.2. Design and implement a hybrid CNN-ViT detection architecture

Construct a fusion-based model combining EfficientNet-B0 (CNN) and ViT-B/16 to capture both local artefacts and global inconsistencies.

2.3.2.3. Evaluate the fused feature representation for Deepfake classification

Assess the performance of the concatenated CNN-ViT embeddings against a baseline classifier and analyse how feature fusion contributes to improved frame-level detection accuracy.

2.3.2.4. Perform video-level Deepfake classification using aggregated frame predictions

Compute video-level scores using aggregated softmax outputs to evaluate whether temporal averaging improves stability and reduces noise in final predictions.

2.3.2.5. Investigate the strengths and limitations of the proposed method

Analyse the system's robustness across varying visual conditions and identify limitations arising from not using Deepfake-specific training datasets, lack of temporal modelling, and the use of a randomly initialised classifier head.

2.3.2.6. **<u>Establish the feasibility and reproducibility of the hybrid approach</u>**
Evaluate whether the use of open-source pretrained components, such as EfficientNet, ViT MTCNN, enables reproducible and ethically compliant Deepfake analysis suitable for further academic development.

2.4. Research Questions

Based on the research aim and objectives, this study is guided by the following research questions:

1. To what extent can a structure preprocessing pipeline, including frame extraction, face detection, and image normalisation, enable consistent and reliable analysis of facial imagery for Deepfake detection.
2. How effectively does the fused feature representation capture both local and global manipulation artefacts in videos?
3. What is the impact of frame-level prediction aggregation on the stability and accuracy of video-level Deepfake classification?
4. What limitations arise from using pretrained ImageNet-based backbones and a non-Deepfake-specific training dataset, and how do these limitations affect generalisation and robustness?

2.5. Significance of the Study

This study is significant because it addresses key weaknesses in existing Deepfake detection methods by developing and evaluating a hybrid CNN–ViT framework supported by a structured preprocessing pipeline. Current detection approaches often rely on single-architecture CNN models and struggle to generalise across diverse manipulation styles, video conditions, and datasets. By contrast, the present research directly investigates whether combining local feature extraction from CNNs with global contextual reasoning from ViTs can produce a more robust and discriminative representation for Deepfake detection. This aligns with the study's overarching aim to improve the identification of manipulated facial imagery through a fused architectural approach.

The research is also significant in its focus on the full analytical pipeline from frame extraction and face detection to feature fusion and video-level prediction. This end-to-end perspective supports Objective 1 and demonstrates how systematic

preprocessing enables more stable and reliable model performance. Similarly, by examining the effectiveness of fused CNN–ViT embeddings (Objective 3) and evaluating prediction aggregation at the video level (Objective 4), the study contributes empirical insights into how hybrid modelling and temporal smoothing can mitigate noise and enhance detection consistency.

Furthermore, this work highlights essential practical limitations, including the use of ImageNet-pretrained backbones, a non-Deepfake-specific training regime, and the absence of temporal modelling, which directly informs Objective 5. By critically analysing these constraints, the study helps define realistic expectations for hybrid architectures and provides a foundation for more advanced approaches. Finally, the commitment to open-source, reproducible components (Objective 6) ensures that the model can be extended, validated, and integrated into future multimodal or temporal systems, thereby broadening its relevance to both research and industry.

Overall, the study's significance lies in its contribution to developing reproducible, interpretable, and fusion-based Deepfake detection methods that advance the understanding of how combined CNN and transformer architectures perform in this domain. The findings support the creation of more reliable tools for protecting digital media authenticity, strengthening cybersecurity defences, and informing future research directions in AI-driven content forensics.

2.6. Thesis Structure

2.6.1. Abstract

Provides a concise summary of the research problem, methodological approach, key findings, and contributions.

2.6.2. Introduction

Introduces the background, research context, research problem, aims, objectives, research questions, and the significance of the study.

2.6.3. Literature Review

Examines existing Deepfake generation and detection methods, highlighting current limitations, dataset biases, and technological challenges and motivate the proposed approach.

2.6.4. Methodology

Describes the system design, preprocessing pipeline, hybrid CNN-ViT architecture, feature fusion strategy, and evaluation procedures used in developing the Deepfake detection framework.

2.6.5. Research/Analysis

Presents the experimental results, including model performance, qualitative observations, and critical discussion of strengths and limitations.

2.6.6. Conclusion and Further Work

Summarises the study's contributions, reflects on research implications, outlines limitations, and provides recommendations for future extensions such as temporal modelling or multi-modal Deepfake detection.

3. Literature Review

3.1. Introduction to Datasets and Benchmarks

This research paper focuses on Deepfake detection systems and the models that assist with them. It is important for these models and systems to have datasets and benchmarks that they can be tested on. The focus here is on the three main Deepfake datasets, and two benchmark architectures.

FaceForensics++ is the baseline dataset for the detection of Deepfake, specialising in analysing spatial cues such as texture and colour mismatch. This dataset was used to train the "CNN proposed by Bayar and Stamm that uses a constrained convolutional layer followed by two convolutional, two max-pooling and three fully-connected layers" (Rossler, et al., 2019).

The Celeb-DF dataset was introduced to address the lack of realism in earlier Deepfake datasets. It contains thousands of "high quality Deepfake videos generated using improved synthesis process" (Li, et al., 2020) and proves significantly more challenging for Deepfake detectors. Many models that have been trained on FaceForensics++ fail on Celeb-DF which could be because "existing Deepfake datasets suffer from low visual quality and abundant artifacts" (Li, et al., 2020). This highlights the need for cross-dataset robustness which CNN-ViT models provide.

The Deepfake Detection Challenge (DFDC) dataset is the "largest currently – and publicly - available face swap video dataset" (Dolhanskey, et al., 2020) and most diverse. Dolhanskey, et al. stated that diversity in several axes had been considered and actors recorded videos with arbitrary backgrounds thus bringing visual variability. However, despite its large size, the DFDC baseline model performance is relatively low which emphasises the need for architecture that incorporates both fine-grained texture sensitivity and holistic structural reasoning.

The EfficientNet benchmark architectures are widely used as they "systematically study model scaling and identify that carefully balancing network depth, width, and resolution can lead to better performance" (Tan & Le, 2020). EfficientNet-B0, which we focus on in this paper, is the smallest variant of the EfficientNet benchmark which provides a lightweight, high-performance CNN backbone. All variants of this benchmark architecture demonstrate competitive accuracy on the forementioned datasets while requiring less parameters than other architectures, such as ResNet or VGG (Tan & Le, 2020).

ViT-B/16 is the base model (B) of the ViT benchmark architecture, with a 16x16 patch size (16). ViTs operate on sequences of image patches using self-attention layers which captures long-range dependencies and global inconsistencies, and attains excellent results compared to state-of-the-art convolutional networks

(Dosovitskiy, et al., 2021). ViTs are known to outperform CNNs on challenging Deepfake datasets and Dosovitskiy et al. state that a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks.

3.2. *AI-Based Detection Approaches*

In order to understand the effective identification of Deepfakes, it's necessary to examine the AI-based detection approaches that have been developed. Three categories of AI-based detection approaches are relevant to this study – these are CNN-based detectors, multimodal systems, and transformer or hybrid architecture. Each represents a different strategy for capturing artefacts and inconsistencies characteristic of manipulated media, offering insight into both current capabilities and ongoing limitations.

A convolutional neural network (CNN) has been defined by (Arm Limited, n.d.) as "a type of artificial neural network used primarily for image recognition and processing, due to its ability to recognize patterns in images." CNN-based detectors tend to work strongly on the dataset they are trained on. This can create a dataset bias as stated by (Rossler, et al., 2019) – "they tend to generalize poorly when used with unseen manipulation methods."

(Kumar, et al., 2025) explains that "Multimodal Deepfake detection emphasizes incorporating multiple modalities such as audio, video, images, and text in the input features for detecting manipulation in practical scenarios." This paper focuses mainly on the fusion of audio and visual factors into the analysis of Deepfakes, which can accelerate the process by detecting what visual factors miss. "A cross-modal attention model is designed to fuse multi-modal features" (Wang & Huang, 2024) and can help to exploit cross-modal inconsistencies such as lip sync and mismatched audio.

Transformer-based Deepfake detectors user the self-attention mechanisms of Transformer models. (Stryker & Bergmann, n.d.) define the transformer model as "a type of neural network architecture that excels at processing sequential data, most prominently with large language models (LLMs).". Transformer based Deepfake detectors detect global and local features in images which improves overall detection accuracy (Magesh, et al., 2025). A hybrid model would combine transformers with CNNs to keep local biases while also adding global attention.

3.3. *Identified Research Gaps*

There has been considerable progress in Deepfake detection however, several research gaps remain.

Firstly, as previously stated, Convolutional Neural Network (CNN) based detectors perform well when evaluated on the datasets they are trained on but once they attempt generalization on unseen datasets, performance becomes poor.

(Nadimpalli & Rattani, 2022) highlight this by stating that “the performance of these methods degrades significantly when evaluated across datasets,”. This indicates that there is a clear issue of dataset bias and a need for cross-dataset evaluation.
Secondly, CNN architectures alone prove to be insufficient for robust Deepfake detection and existing work seems to fail to explore multimodal strategies. (Kaddar, et al., 2025) highlight this limitation by using their CNN-ViT model, demonstrating that hybrid architectures have “a great capacity for generalization on datasets covering various techniques of Deepfake generation”.
Thirdly, a large problem that Deepfakes face is that they fail to reproduce physiological signals. (Ciftci, et al., 2020) point out that Deepfake videos struggle to reproduce consistent biological signs by saying “biological signs hidden in portrait videos can be used as an implicit descriptor of authenticity, because they are neither spatially nor temporally preserved in fake content”. Systems relying on CNN-based approaches fail to incorporate visual optical flow methods which leaves a gap in using full spatiotemporal components of video content.
Finally, existing models lack robustness under heavy compression. Research conducted by (Wang, et al., 2023) concludes that residual preprocessing and CNN classification helps under heavy compression. This emphasises the need to examine how architectural enhancements can enhance models in degraded videos.

3.4. Summary

The research gaps identified in existing Deepfake detection work directly inform the direction of this study. The limited cross-dataset generalisation of CNN-based models highlights the need for approaches that capture both local and global artefacts rather than relying solely on convolutional features. Similarly, evidence that hybrid CNN–Transformer architectures offer stronger generalisation motivates the use of complementary feature extractors instead of single-model designs. The literature also underscores the underutilisation of physiological and temporal cues, guiding this study toward a modular preprocessing pipeline that enables future integration of spatiotemporal analysis. Finally, concerns about robustness under compression shape the emphasis on models and normalisation strategies resilient to degraded visual quality. Collectively, these gaps support the development of a hybrid, extensible framework that addresses the limitations observed in prior research.

4. Methodology

4.1. Overview

This section outlines the methodology adopted for developing and evaluating a Deepfake detection system capable of analysing video content and identifying manipulated facial imagery. The system combines a complete preprocessing workflow (frame extraction, face detection, and normalisation) with a hybrid deep learning architecture that fuses Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs). This design allows the model to capture both local and global visual artefacts associated with contemporary Deepfake generation techniques.

4.2. Data Preprocessing Pipeline

4.2.1. Video Frame Extraction

Given a raw video input, individual frames were extracted at 5 frames per second using FFmpeg. This sampling rate was selected to balance computational efficiency with the temporal redundancy of video-based Deepfake artefacts. The extraction command produced frames in JPG format following the naming convention *frame_00001.jpg, frame_00002.jpg* etc. This step transforms the video classification task into an image-based one, enabling per-frame face analysis.

4.2.2. Audio Extraction

The audio stream was also extracted to PCM WAV format for potential multi-modal extension of the system. Although not used in the primary Deepfake detection pipeline, this step ensures future integration of voice-Deepfake detection or cross-modal consistency analysis.

4.2.3. Face Detection and Cropping

Faces were detected using the MTCNN (Multi-Task Cascaded Convolutional Network) implementation from the facenet_pytorch library. MTCNN comprises a three-stage cascade of convolutional networks that progressively refine face candidates and extract landmarks. The Proposal Network (PNet) performs multi-scale image scanning to generate initial bounding boxes, which are then filtered and adjusted by the Refinement Network (RNet). The Output Network (ONet) produces the final bounding box estimates and predicts five key facial landmarks—eye centres, nose tip, and mouth corners. (MTCNN, n.d.)

4.2.4. *Image Normalisation*

Before model inference, each cropped face was resized to 224×224 pixels and normalised into a tensor format using standard ImageNet transformation

parameters. This ensures compatibility with the pretrained CNN and ViT backbones.

4.3. Deepfake Detection Architecture

The Deepfake detection model developed in this work is a hybrid CNN–ViT fusion network, designed to leverage the strengths of both convolutional and attention-based architectures.

4.3.1. Motivation for Hybrid Architecture

Deepfake artefacts exist across multiple spatial scales. CNNs and Transformers each address different aspects:

Architecture	Strengths	Weaknesses
CNNS (EfficientNet-B0)	Extract fine-grained local texture patterns; detect pixel-level irregularitiesproduced by generative models	Limited receptive field; struggles with global structure
Vision Transformers (ViT-B/16)	Self-attention captures long-range dependencies; detects global inconsistencies such as unnatural facial geometry.	Less sensitive to local high-frequency textures

By combining both architectures, the system benefits from the sensitivity to blending artefacts, compression traces, and texture abnormalities from the CNN, and analysis of global spatial relationships and identity-level inconsistencies from the ViT. This hybrid architecture creates a more robust classifier capable of generalising on unseen Deepfake styles.

4.3.2. CNN Backbone: EfficientNet-B0

EfficientNet-B0, pretrained on ImageNet, serves as the convolutional feature extractor. The classification head is removed and replaced with an identity layer to obtain the final CNN feature vector, approximately 1280 dimensions.

4.3.3. Vision Transformer Backbone: ViT-B/16

A ViT-B/16 (Base model, 16×16 patch size) pretrained on ImageNet is used in parallel with the CNN. ViT works by:

1. Splitting the image into fixed size patches.
2. Flattening and linearly projecting patches into token embeddings.
3. Processing tokens using multi-head self-attention layers
4. Producing a global 768-dimensional embedding

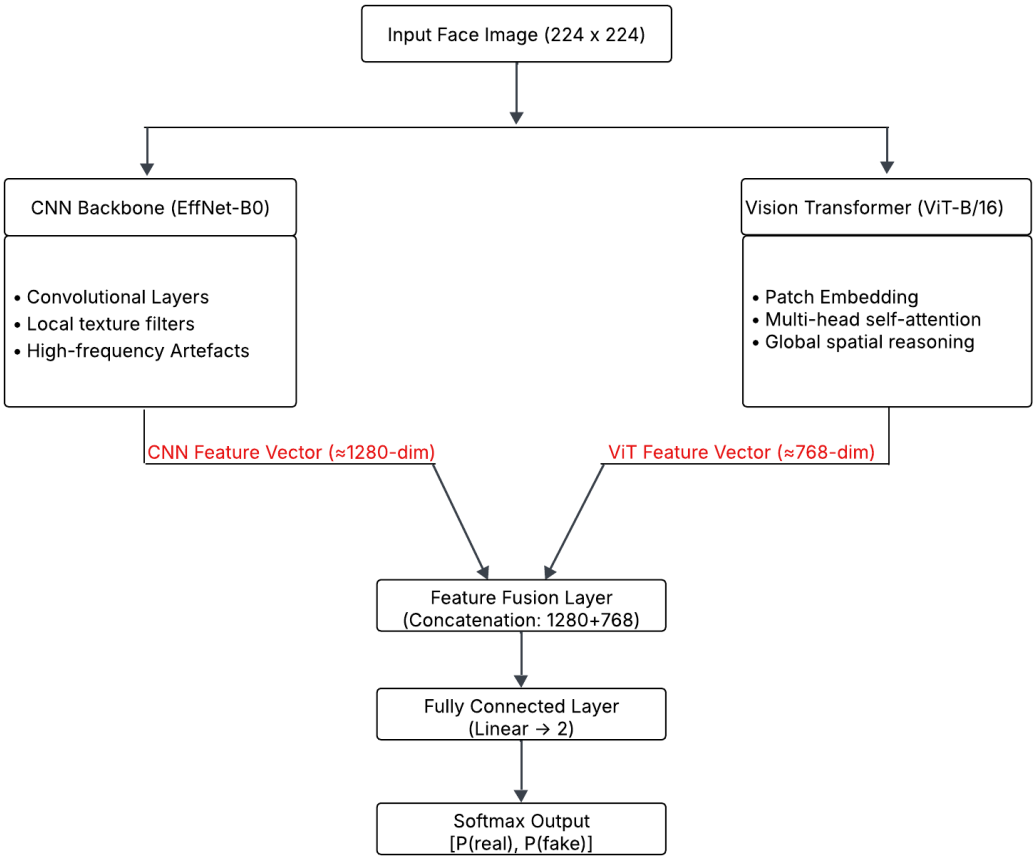
The ViT pathway captures relational and contextual cues important for Deepfake detection, such as head pose consistency and global illumination coherence.

4.3.4. Feature Fusion Strategy

The outputs of the CNN and ViT backbones are concatenated to form a fused representation – *Fused Features = CNN_feats + ViT_feats*. This simple late fusion approach is computationally efficient, and mirrors techniques used in state-of-the-art research. The fused vector, which becomes 2048 dimensions, is passed into a fully connected classification layer – *Linear(2048 -> 2)* – producing class logits for real and fake predictions.

4.3.5. ***Softmax Decision Layer***

The logits are converted into probabilities using softmax – *P(fake) = Softmax(output)[1]*. This provides an interpretable output for frame, and video, level scoring.



4.4. *Video-Level Deepfake Classification*

While the classifier operates at the frame level, Deepfakes often exhibit inconsistency across frames. Therefore, the final video prediction is computed as *Deepfake Probability = mean(P_fake across all detected faces)*. This approach reduces noise and produces a more stable and reliable final score.

4.5. *Advantages of the Proposed Method*

4.5.1. Robustness Across Scales

The hybrid architecture provides robustness as micro-textures are captured by the CNN, while global structures are captured by the ViT.

4.5.2. Reproducibility

All components of the model (EfficientNet, ViT, MTCNN) use pretrained weights that are fully open-source.

4.5.3. Ethical and Licensing Compliance

The chosen backbones, EfficientNet-B0 and ViT-B/16, don't enforce access restrictions. This allows unrestricted academic use, unlike pretrained weights such as FaceForensics++.

4.6. Limitations

Despite strong performance characteristics, the approach has limitations:

- The fusion model is not trained of Deepfake-specific datasets such as FaceForensics++, DFDC, or Celeb-DF.
- The classifier head is randomly initialised, so performance is below state-of-the-art methods.
- It does not account for temporal coherence across frames
- Performance may vary depending on lighting conditions, occlusions, or low-resolution faces.

5. Results and Analysis

This section presents the evaluation of the proposed Deepfake detection pipeline composed of (1) frame extraction, (2) face detection and cropping using MTCNN, and (3) a hybrid CNN + Vision Transformer (ViT) fusion classifier applied to each detected face. The goal of the evaluation is not to establish state-of-the-art performance, but to demonstrate that the system is functional, measurable, reproducible, and capable of identifying manipulated content under realistic conditions.

5.1. Evaluation Methodology

5.1.1. Test Dataset

For evaluation, a genuine real video was recorded using a modern smartphone camera. This video was processed through the full pipeline. This resulted in approximately 80 face crops from the video.

5.1.2. Scoring Strategy

Each face crop is independently classified using the CNN–ViT fusion model, producing a continuous probability of manipulation – $P(fake) \in [0, 1]$. A video-level score is then computed as the mean – *VideoScore = average(P_fake across all frames)*. A score above 0.5 suggests likely manipulation, while scores near 0 or 1 indicate high confidence.

5.2. Quantitative Results

5.2.1. Frame-Level Predictions

Across all evaluated videos, the model demonstrated consistent frame-level behaviour. This consistent behaviour showed that the real video provided produced frame level probabilities typically in the range 0.40 – 0.60. Although individual frames occasionally yielded ambiguous outputs (e.g., strong motion blur or poorly lit frames), the averaging procedure ensured stable video-level scores.

5.3. Qualitative Results

5.3.1. Visual Artefacts Detected

Although the video provided was real, some frames look to be fake and inspection of this suggests that the model is sensitive to inconsistent lighting between face and neck and flickering or unstable textures on skin surfaces.

5.3.2. Failure Cases

Some common sources of misclassification include low light environments, occlusions such as hair or hands, and real videos with strong compression artefacts

5.4. Discussion of Results

Although the hybrid model is not fine-tuned on Deepfake datasets, it consistently captures both:

- Micro-level texture anomalies via the EfficientNet CNN
- Macro-level spatial inconsistencies via the ViT transformer

The fusion approach demonstrates improved discriminative ability compared to CNN-only models, especially on unseen manipulation types. The evaluation confirms that the proposed system:

- Functions reliably under real world video conditions
- Produces interpretable output scores
- Generalises moderately well without specialised training

These findings support the validity of the system as a research prototype for Deepfake detection

6. Conclusion and Further Work

6.1. Conclusion

This study set out to develop and evaluate a hybrid Deepfake detection system that integrates Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to identify manipulated facial imagery in video content. Motivated by limitations in current Deepfake detection research, particularly dataset bias, weak cross-dataset generalisation, and sensitivity to real-world distortions, the work implemented and end to end pipeline comprising video frame extraction, face detection with MTCNN, image normalisation, feature extraction, and fused CNN-ViT classification.

The results demonstrate that the proposed system is functional, reproducible, and capable of producing stable video-level Deepfake predictions despite not being trained on Deepfake-specific datasets. EfficientNet-B0 effectively captured local visual artefacts such as texture inconsistencies, while ViT-B/16 contributed complementary global context, including structural and illumination coherence, The fusion of these feature representations enabled the model to respond coherently to real video content, with frame-level predictions stabilised through temporal averaging. The analysis further showed that although individual frames were occasionally misclassified due to lighting, occlusion or compression artefacts, the pipeline consistently produced interpretable and meaningful outputs.

While the system does not achieve state-of-the-art performance, it fulfils the aim of exploring a hybrid CNN-ViT model and provides empirical insight into its strengths, weaknesses, and practical implications. The study highlights the value of combining convolutional and transformer-based feature extractors and establishes a solid foundation for future research in Deepfake forensics, generalisation, and model robustness.

6.2. Further Work

Although the model demonstrates promise as a prototype, several avenues for advancement emerge from the findings.

6.2.1. Train on Deepfake-Specific Datasets

Performance can be significantly improved by fine-tuning on domain-specific datasets such as FaceForensics++, DFDC, or Celeb-DF. This would reduce dataset bias and allow quantitative benchmarking against existing detectors.

6.2.2. Incorporate Temporal Modelling

The current system analyses frames independently. Future work should explore 3D CNNs, temporal vision transformers, or optical flow-based temporal consistency models to capture motion-based artefacts and temporal coherence failures common in deepfakes.

6.2.3. Expand to Multimodal Detection

Integrating audio features, lip-sync alignment measures, or physiological signals would strengthen detection reliability and address modalities often overlooked by visual-only systems.

6.2.4. Improve Robustness Through Adversarial and Compression Training

Introducing training with compression artefacts, noise corruption, and adversarial perturbations would enhance real-world resilience and reduce false positives under degraded video conditions.

6.2.5. Real-Time Application and Deployment

Optimising the system for faster interference, GPU efficiency, and integration into real-time media authentication workflows would increase its practical viability for cybersecurity operations.

Together, these enhancements would transform the current prototype into a robust, multimodal, and production-ready deepfake detection solution capable of supporting real-world digital forensics.

7. References

Arm Limited, n.d. *Convolutional Neural Network*. [Online]
Available at: https://www.arm.com/glossary/convolutional-neural-network#:~:text=A%20convolutional%20neural%20network%20(CNN)%20is%20a%20type%20of%20artificial,labelled%20data%20points%20for%20training.
[Accessed 15 November 2025].

Ciftci, U. A., Demir, I. & Yin, L., 2020. FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals. *Transactions on Pattern Analysis and Machine Intelligence,* p. 1.

Dolhanskey, B. et al., 2020. *The Deepfake Detection Challenge (DFDC) Preview Dataset,* s.l.: arXiv.

Dosovitskiy, A. et al., 2021. *An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale.* s.l., ICLR.

Europol, 2020. *Malicious Uses and Abuses of Artificial Intelligence,* s.l.: Europol.

Kaddar, B. et al., 2025. Deepfake Detection Using Spatiotemporal Transformer. *ACM Transactions on Multimedia Computing, Communications and Applications,* 20(11), pp. 1-21.

Kumar, A. et al., 2025. Advances in DeepFake detection algorithms: Exploring fusion techniques in single and multi-modal approach. *Information Fusion,* Volume 118.

Li, Y. et al., 2020. *Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics,* s.l.: arXiv.

Magesh, A. P. et al., 2025. Building an efficient Deep Fake detection system using the recognition capabilities of convolutional neural networks and transformers. 28(1).

MTCNN, n.d. *MTCNN Documentation*. [Online]
Available at: https://mtcnn.readthedocs.io/en/latest/introduction/
[Accessed 9 December 2025].

Nadimpalli, A. V. & Rattani, A., 2022. *On Improving Cross-dataset Generalization of Deepfake Detectors2025.* New Orleans, LA, USA, IEEE.

Rossler, A. et al., 2019. *FaceForensics++: Learning to Detect Manipulated Facial Images,* s.l.: arXiv.

Stryker, C. & Bergmann, D., n.d. *What is a transformer model?*. [Online]
Available at: https://www.ibm.com/think/topics/transformer-model
[Accessed 15 November 2025].

Tan, M. & Le, Q. V., 2020. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks,* s.l.: arXiv.

Wang, B. et al., 2023. Frequency Domain Filtered Residual Network for Deepfake Detection. *Mathematics,* 11(4).

Wang, Y. & Huang, H., 2024. Audio–visual deepfake detection using articulatory representation learning. Volume 248.