National College of Ireland

An Optimised Deepfake Detection System Using Vision Transformers and FFT

MSc Research Project
Msc Data Analytics

Nobin Sabu
Student ID: x23313081

School of Computing
National College of Ireland

Supervisor: Eric Gyamf

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Nobin Sabu
Student ID:	x23313081
Programme:	Msc Data Analytics
Year:	2025
Module:	MSc Research Project
Supervisor:	Eric Gyamf
Submission Due Date:	11/08/2025
Project Title:	An Optimised Deepfake Detection System Using Vision Transformers and FFT
Word Count:	3684
Page Count:	20

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:	Nobin Sabu
Date:	13th September 2025

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):	

An Optimised Deepfake Detection System Using Vision Transformers and FFT

Nobin Sabu
x23313081

Abstract

The proliferation of deepfake media—AI-generated synthetic videos that closely resemble authentic content—poses a significant challenge to digital trust, public safety, and content verification.Conventional deepfake detection systems, especially the ones that employ convolutional neural networks (CNNs), tend to fail to perform due to real-world factors, including video compression and non-obvious ways of manipulating them. This study is undertaken to curb such limitations, as it follows in the proposal of a HybridSwinCNN, which is a dual-stream design with advantages of merging a spatial and frequency domain to perform better in terms of accuracy and robustness. The network comprises Swin Transformer to capture the semantic spatial features on RGB pictures and EfficientNet CNN to capture the high-frequency outliers in the log-magnitude spectrum Fast Fourier Transform (FFT).

The study uses the FaceForensics++ data which has 23,988 video frames extracted, preprocessed, trained, and validated using 20 epochs of model. This is evidenced in the fact that the HybridSwinCNN suggested attained the best validation performance with an accuracy of 85.86 percent and a final train accuracy of 99.47 percent, both indicative of good feature learning and generalization capability. The findings indicate the usefulness of the joint use of spatial and frequency features in detecting deep fake and specifically when the material under scrutiny is compressions and adversarially corrupt content. The solution proposed in the work is scalable and interpretable that can be applied in the situation of digital forensics, media verification and online content systems. Future directions will involve the best practise on generalization to unseen forgeries, explainability and making the model suitable to real-time applications.

1 Introduction

Deepfakes—AI-generated synthetic images and videos—have rapidly emerged as a significant threat to the authenticity of digital media. These manipulations are increasingly indistinguishable from genuine content, enabling malicious actors to spread misinformation, commit fraud, or impersonate public figures. With the rise of social media platforms and generative AI tools, detecting deepfakes has become a pressing issue for digital security, content verification, and public trust Amin et al. (2024). Traditional deepfake detection systems, especially those based solely on convolutional neural networks (CNNs), analyze pixel-level information from images but often fail under real-world conditions

such as compression, noise, or resolution loss. As the sophistication of deepfake generation techniques grows, so does the need for more resilient and intelligent detection approaches.

Most existing detection models are limited in that they focus only on spatial domain features—what the eye can see—without considering the subtle patterns hidden in the frequency spectrum. These frequency cues can reveal manipulation artifacts not obvious in the RGB domain. This shortfall is particularly critical when videos are compressed or tampered with post-generation. To address these weaknesses, machine learning methods that incorporate both spatial and frequency information are essential Liao and Chen (2024). Such models can analyze more comprehensive signal characteristics, improving accuracy and robustness, especially in low-quality or adversarial settings.

Even though deepfake detection has seen significant development, existing approaches still leave considerable room for enhancement. Many studies rely on either deep CNNs or Transformer-based architectures, with minimal attention paid to fusing spectral and spatial modalities. Furthermore, current models often lack generalizability to deepfakes created with different algorithms or formats. This study aims to overcome these limitations by designing a hybrid model that processes spatial features using a Vision Transformer (Swin Transformer) and frequency-domain features using a CNN applied to FFT-transformed inputs.

RQ: *How can a hybrid spatial-frequency deepfake detection model using Vision Transformers and FFT improve detection accuracy and robustness?*

- **SQ1:** *What is the effect of the combination of FFT-based frequencies and Vision Transformer-based space information on the improvement of the deepfake detection performance across a variety of manipulation types?*
- **SQ2:** *What advantages does dual-stream hybrid architecture have in terms of deepfake detection as it relates to capturing complimentary spatial and frequency-based features?*

Although recent literature has discussed spatial (or frequency-based) approaches separately, there has been less investigation of their combination in an end-to-end architecture. There is a vast number of researchers who use CNNs to identify the anomalies of pixels or use statistical filters in the frequency domain but few analyze the combinations of these two domains into a deep learning model. Also, models can sometimes be tested in perfect laboratory environment, rather than in the actual models where these artificial aspects are present, such as noise, compression, or lost frame. This paper fills these gaps by developing two-stream based on encouraging RGB and FFT based features to be combined before being classified. This model is further tested over the FaceForensics++ benchmark, which is packed with a variety of images with realistic deepfakes and evaluated through common measures of both accuracy and F1-score as well as precision.

The study involves several core steps: preparing the dataset, transforming RGB images to the representation of frequencies through FFT, training and implementation of Swin Transformer and EfficientNet CNNs, combining outputs of the features, and assessing the performance of the models. The model was trained and tested in Kaggle Notebook infrastructure with NVIDIA GPU acceleration based on Python, PyTorch, timm library of pre-trained models, OpenCV to process the images, NumPy to compute the FFT and Matplotlib to create the visualization. The idea is to develop the detection mechanism

capable of not only determining deepfakes but also being stable when encountering noisy, compressed, or otherwise degraded input signal.

This project has the potential to benefit a wide range of stakeholders—from social media platforms and cybersecurity agencies to news organizations and the general public—by providing a smarter, more generalizable approach to detecting synthetic media. It contributes to the broader effort of preserving digital truth in an era of increasingly convincing manipulation technologies.

1.1 Structure of the Paper

This thesis is organized as follows:

1. **Section 2** explores the existing literature on deepfake detection including spatial, frequency, and hybrid detection methods.
2. **Section 3** explains the overall methodology, including how the dataset was prepared, the pre-processing procedures and key architectural decisions.
3. **Section 4** gives a detailed design of the proposed HybridSwinCNN system, including spatial and frequency components.
4. **Section 5** presents the implementation process, tools and technologies involved and practical implementation of the model.
5. **Section 6** gives the evaluation result including key performance metrics and a discussion of model effectiveness.
6. **Section 7** concludes the thesis by summarising the key findings, limitations and providing future research recommendations.

2 Related Work

2.1 Traditional Deepfake Detection Techniques

The initial methods of deepfake detection largely utilized the use of convolutional neural networks (CNNs), to obtain the spatial features of RGB frames. Such approaches centered on teaching discriminative patterns in facial structures, and texture and minor inconsistency on the pixel level by generative adversarial networks (GANs). XceptionNet was among the most famous CNN-based models that proved to be relatively successful since it picked up localized artifacts of facial manipulations of multiple types Rössler et al. (2019). But when it was experimented with other compression criteria or unobserved datasets its performance deteriorated drastically indicating serious drawbacks in its generalization.

The Attention-based Deepfake Detector (ADD) presented by Khormali and Yuan (2021) was based on channel and spatial attention modules on top of CNNs to enhance feature localization. This strategy allowed this model to dynamically emphasize on the forged areas at the same time maintaining a global context. The method they offered won baseline CNNs in high-quality datasets but could not cope with lossy video compression. Likewise, Zhao et al. (2021) suggested continuously interpreting multiple attention at once with regional and hierarchical attention, which promoted the discrimination of features

on different regions of the face. These developments signified a transition to attention-augmented CNNs, but these were still limited to the capabilities of spatial-only vectors.

In spite of promising values of accuracy measurement in laboratory conditions, conventional ones have low robustness against noise or video compression or against new forms of manipulation. Also, CNN-based detectors tend to acquire biases on the specific dataset, imposing restrictions on their generalizability to different platforms or synthesis techniques. These issues have driven motivation towards more invariant detection schemes, especially within the frequency-domain analysis capacity to utilize inherent spectral artifacts left by generative models.

2.2 Frequency-Domain Analysis in Deepfake Detection

Although spatially CNN-based detectors ensure detecting surface-level patterns, they tend to miss subtle traces of manipulation generative model introduces at the signal level. In attempts to fix this, frequency-domain schemes analyze such high-frequency inconsistencies as the spectral properties of images the errors are sometimes ignored in the pixel-based analysis. They have become popular because the manipulation cues they locate are resistant to the typical post-processing, e.g. compression and resizing.

In the study by Bah and Dahmane (2024), the researchers have suggested a process that converts the images into higher ones by using Fourier transforms and passing the images into a CNN (Xception) to improve identifying the low-level frequency anomalies more transparently. Their model had better accuracy at identifying the compressed data which provided evidence on the robustness of the frequency-based representations. In the same way,Zhou et al. (2022) proposed FourierUp, a framework that trains the multi-resolution frequency features to learn to reconstruct residual manipulation signals. These inventions alter the paradigm of detection towards learning spectra, which constitutes a robust framework to generalizable detection.

Some other methods have focused on physiological discrepancies. Ni et al. (2024) used remote photoplethysmography (rPPG) data and fast Fourier transform (FFT) to unveil the discrepancies between patterns of blood flow across the deepfake faces, demonstrating that there were biological data beyond the texture of pixels. The aspect of frequency perturbation strategy has also been investigated. FrePGAN, the product created by Jeong et al. (2022), represents a technique that induces adversarial frequency corruptions in order to learn more resistant detectors and show superior performance across datasets.

Recent architectures like FreqNet Tan et al. (2024) have frequency-aware modules further boosting a training-time improvement of spectral features, further improving detection capacity when compressed with similar artifacts to the ones generated by the experiment. High-frequency enhancement (HiFE) layers were also proposed by Gao et al. (2024) in the context of detecting low-quality deepfakes and indicated high accuracy to perform detection during poor visual conditions as well.

However, while frequency-domain methods are more resistant to attack and more generally applicable, they tend to ignore the semantic context and are not very effective against fine-grain, semantically reasonable fordings. These restrictions have motivated an increased interest in integrating frequency information with modern vision architectures such as transformers- which has the potential to integrate low-level information spectral processing with high-level semantic processing.

2.3 Deepfake Vision Transformers

Vision Transformers (ViTs) have transformed deepfake detection technology by switching locality-constrained filters of CNNs with global dependencies models of self-attention. This architectural change enables ViTs to connect spatial inconsistencies and long-range correlations in forged media- particularly to locate manipulation traces where they are either imperceptible or decentralized.

Gong et al. (2024) introduced a Swin-Fake model, which was a Swin Transformer model with the addition of consistency learning and achieved better results in cross-dataset experiments than CNNs. The hierarchical attention mechanism was able to generalise more in various forms of deepfakes. In the same vein, Xi and Chen (2025) added Swin block to a spatial transformer pipeline, enhancing compression artifacts robustness by adapting dynamically to irregularities among the regions.

Through structural variants, other works have been used to enhance granularity. Çınar and Doğan (2025) retains local manipuliations through patch-wise tokenization, which is essential to the identification of local or small forgeries. M2TR Wang et al. (2022) integrates multi-modal data streams, including audio and visual fodder, with hierarchical attention layers by making a model pronounced with varied manipulations. Xu et al. (2023) proposed TALL, a transformer architecture that featured thumbnail layout representations to capture spatiotemporal distortions only detected in video forgeries.

The ViTs are shown to have high detection across different images as well as improved domain transfer in contrast to the CNNs. But they need large quantities of training data and are susceptible to noise in videos of poor quality and are not directly sensitive to traces of spectral manipulation. These constraints have seen the emergence of a growing interest in hybrid architecture that combines transformer-based reasoning abilities with the frequency-domain characteristics, enhancing both reliability and interpretability.

2.4 Hybrid: Spatial-Frequency Fusion Architectures

Hybrid which are a combination of spatial and frequency features have demonstrated good promise in enhancing accuracy of detection as well as generalization. These models compensate the weaknesses of single-domain approaches, which lie in absence of complementary strengths, by using complementary strengths: the frequency domain in terms of artifact sensitivity and the spatial domain in terms of semantic structure.

The first solution to combine biological signals (e.g. PPG) and spatial features, using a robustness level against routine post-processing, relied on FakeCatcher Ciftci et al. (2020). The DFSF network would be a dual-stream, combining the output of the spatial CNNs and frequency filters, proposed by Zhai et al. (2024) and would be better at detecting both subtle and overt manipulations. Their model performed better than baseline CNNs on several different datasets, hypothetically on the compressed videos.

Likewise, FMSI Luo and Wang (2025) presented an integrated frequency-masking and spatial interaction block, allowing more recent detectors based on ViTs to find damaged areas more accurately. VLF-Net Peng et al. (2025) has taken advantage of multimodal inputs by using masked language models as well as visual information to work at a more detailed i.e. frame-level. In video detection, Kim et al. (2025) applied the pixel-wise temporal frequency analysis framework, as to capture small temporal differences that would not be detected by using static detectors.

Hybrid architectures can therefore provide a sensible compromise between models– they provide the interpretability of CNNs and the spectral stability of frequency model.

Given the evolution of the transformer-based system, the techniques of fusion might be at the center of potential improvement involving precision and domain adaptation.

2.5 Generalisation and Robustness in Deepfake Detection

In addition to improving accuracy, modern detection models must have generalizability and strength under different operating conditions. There are a number of recent researches which overcome this by directly addressing it. Use of benchmark datasets is not sufficient to carrying out robust deepfake detection in real-world process. Models are expected to generalise to variety of unknown manipulations, compressions, and synthesis techniques. The implemented solution of Batagelj et al. (2025) confronted this challenge by learning, based on actual facial images, with the help of the multitask self-supervised framework using auxiliary segmentation and pseudo-fake generation. Equally, Shen et al. (2025) developed a model, called AuthGuard, which employs papers vision-the language model to learn through LLM-instructed contrastive learning, to find semantic incompatibilities, performing well on such datasets as DFDC and DF40.

RoGA domain-generalization method proposed by Qiu et al. (2025)is a method of aligning the direction of gradient in various datasets to enable learning robust invariant features, without altering architecture because RoGA is the best method of learning invariant features. Liu et al. (2023) added further to it with the multi-view completion representation to fill gaps left in missing views produced by GAN models and focused more resistant to occlusion and compression. These models assist in generalization by use of architectural innovation and semantic reasoning.

In addition to the visual qualities, researchers Nguyen-Le et al. (2024) have given the essentiality of applying a multimodal detection system which is a combination of visual, audio, and physiologist information, grounds on which they will construct scalable systems. Despite this, the lack of centralized guidelines can also be considered one of the greatest challenges as well as the continuous creation of generative models. To make it safe to deploy, the work ought to be committed to the benchmarking of the measure under adversarial attacks, domain shifts, and few shots setting.

2.6 Gap in Research

The literature brought to light an evident paradigm shift in the sphere of deepfake detection. The evolution of deepfake detection space models was replaced by advanced hybrid models that combine frequency, temporal and semantic signal hints. Conventional approaches provided ground-breaking insights, however, they were not very robust to compression or a change of domains. This would be handled by frequency methods by focusing on spectral artifacts, combined with better contextual awareness provided by Vision Transformers. Hybrid models, especially cross-dataset work horse of being able to drive frequency-aware modules and transformers, have produced the most reliable cross-dataset performance.

Despite these developments, there are still considerable gaps. There are still little works that take full advantage of the synergy between Vision Transformer and explicit frequency transformation such as FFT. Current hybrid approaches tend to be handcrafted on fusion or have use-constrained dataset. Also, it is the fate of most systems that they are not very generalisable to different kinds of manipulations and resolutions, which is crucial to having a real-world system.

This research aims to address these gaps by proposing a deepfake detection approach based on the Vision Transformers (ViTs) by means of the frequency-domain method, where spatial and spectral learning is jointly optimised. By incorporating FFT-based modifications, as well as by using transformer attention, the proposed system aims at being robust, interpretable and domain-generalizable detection.

2.7 Literature Matrix

The Literature Matrix below summarises key research contributions in the areas of CNN-based deepfake detection, frequency-domain analysis, Vision Transformers, hybrid spatial-frequency fusion architectures, and generalisation techniques for robust cross-dataset detection.

Table 1: Literature Matrix

Author Name	**Proposed Solution**	**Limitations**	**Comparison with Other Research**	**Gaps Identified**
Bah & Dahmane (2024)	Employed Fourier transforms to unmask low-level discrepancies in frequencies.	Does not pay attention to the semantic information.	Better against compression compared to the spatial-only approach.	Does not integrate into high-level feature comprehension.
Zhou et al. (2022)	Suggested FourierTo learn multiple-scale spectral features.	High computational cost required.	Better at working with compressed video, compared to CNNs.	Could use more semantic-spatial fit.
Ni et al. (2024)	Recognized unnatural blood flow by translating weighted biological signals (rPPG), and FFT.	Not generalisable and works only with facial forgeries.	New physiological applications.	Specific to face-centric detection.
Jeong et al. (2022)	Engineered FrePGAN to corrupt the frequencies of the adversarial corruption.	The calculation is demanding and more difficult to train.	Better results than the GAN-based detectors.	Requires to combine strength and readability.
Gong et al. (2024)	Created Swin-Fake with consistency learning and Swin Transformer.	Requires big amounts of data and is susceptible to video noise.	Excels on cross-dataset tasks than the CNNs.	Fails to record spectral traces.
Xi & Chen (2025)	Introduced Swin Transformer blocks as a source of robustness (to compression).	Does not have the frequency-domain consultation.	Compared to previous ViT models, it accomplishes spatial tasks better.	Needs to merge with the spectral characteristics.

Author Name	Proposed Solution	Limitations	Comparison with Other Research	Gaps Identified
Cinar & Doğan (2025)	Suggested patch-wise tokenization as a means to improve ViT efficiency.	Performance drops when patch embedding lacks overlap.	Combines hybrid ViT for fast pretraining.	Overlooks frequency-specific fake traces.
Zhai et al. (2024)	Designed DFSF: two-stream network that has both frequency and spatial stream.	A fusion done by hand without any temporal analysis.	Better cross-dataset results than CNNs.	Unsuitable to various manipulations.
Luo & Wang (2025)	Proposed FMSI block in which frequency masking is combined with spatial attention.	Complicated tuning and sense.	Improves the performance of ViT under compress conditions.	Needs simplification training pipeline.
Shen et al. (2025)	AuthGuard implements contrastive learning based on vision-language models.	Based on pretraining with LLMs and text-visual matching.	Has better ability to resolve semantic contradictions when compared to CNNs.	Difficult to implement out of context of language.

3 Methodology

The chapter has given a detailed and depictive account of the systematic methodology that has been adopted to the research problem of deepfake video detection. The solution is based on the assumption that processing the information within the spatial and the frequency domain, and then merging the two may come up with a more reliable and accurate detection paradigm. To be scientifically rigorous, transparent, and in case additional experiments are needed, reproducible each step of the research, including data collection and preprocessing, structure of the model, training, and evaluation of the model, is explicitly described.

3.1 Dataset collection and source Description

3.1.1 Dataset Description

This study utilizes the FaceForensics++ (FF++) dataset, which is a large-scale and industry-standard benchmark dataset representative of those used in commercial settings when it comes to detecting deepfakes.Its extensiveness and popularity make it a good answer to the question of whether the new detection algorithm can be characterized as robust and generalizable. The database is constructed with the help of 1000 original, clean videos obtained on YouTube. The videos formed the source material to create thousands of manipulated videos by using four different state-of-the-art technique of face manipulation:

- **DeepFakes:** This is a face swapping technology which makes use of GAN whereby one swaps the identity of a target person with that of a source person.
- **Face2Face:** A facial re-enactment system to map facial expression of a source person to target person.
- **FaceSwap:** This face-swapping approach is based on computer graphics and relies on the facial landmark detection and blending.
- **NeuralTextures:** A method of re-enactment that makes use of GANs to synthesize a target face texture on the basis of the expressions of a source.

To make the evaluation strong, the designers of FF++ present the videos in varying qualities by altering the compression factor. This setting is a light, visually near-lossy quantization and looks like the setting of many actual on-the-Internet video sharing sites: challenging and realistic, but not totally lossless, it is an appropriate setting in which to test a detection framework.

3.1.2 Frame Extraction and Process

Since the model architecture in question is built to process individual frames by nature, rather than sequential video footage, an extraction pipeline was the initial step of the training process. In order to obtain a computationally tractable but representative sample of the data, we set to work a strategy as follows:

1. We open the video files one by one (both the original and the fake files in the FF++ folder).
2. Frames were sampled every constant number of frames to get temporal diversity and to avoid excessive redundancy between successive frames. Particularly, each 10 th image of the video was taken out.
3. We also wanted to avoid the effect of having one video to dominate the data by having too many of their frames extracted, and certainly very lengthy videos would lead to that, so we limited the maximum number of frames that we extracted per video to 10.
4. Each video frame was also stored as high-quality PNG image, into a well-organized directory, and with the label of ground-truth on whether that frame was real or fake, as it had in the parent video.

The data collected through this process of systematic observation was highly considerable since it was in the form of 29,939 individual image frames. This is when these data were divided into training set (80%, or 23,951 images) and validation set (20%, or 5,988 images). This default 80/20 split guarantees that a considerable amount of data is left to train the model, whereas a substantial, entirely new part is held to evaluate the level of performance evenhandedly.

3.2 Dual-Stream Preprocessing and Data Augmentation

In order to input the two streams of data into our dual-stream model, two independent and parallel preprocessing pipelines were carefully constructed, one to be used with the spatial (RGB) data and one to be used with the frequency (FFT) data.

3.2.1 RGB Stream: Spatial Preprocessing

The RGB stream is supposed to acquire semantic information, textures, color inconsistencies, and top-level contextual information in the images. The preprocessing of this stream was on the basis of the best practices of training pre-trained vision models:

Resizing: All images were resized to a fixed resolution of 224x224 pixels. This standardization is necessary because the amount of time it takes during the training process is limited, usually at a certain size, which depends on the architecture of the model.

Data Augmentation: Data Augmentation was also done in order to augment generalization and decrease overfitting; and this was done using RandomHorizontalFlip. This is an augmentation method which produces new training samples by randomly flipping the images up-and-down, and training the model to be invariant to the direction of face images.

Tensor Operation: The PIL images went through the process to become PyTorch tensors, which is a universal operation data structure in deep learning.

Normalization: The tensors were normalized by the mean [0.485, 0.456, 0.406] and the standard deviation [0.229, 0.224, 0.225]. Here, channel-wise data of ImageNet dataset. Their application is essential when transfer learning is required because it involves matching the distribution of the input data to what was originally present during training of a model, and this avoids training instability during the initial training.

3.2.2 FT Stream: Generation and Preprocessing in Frequency Domain

The frequency stream has been set up to snap up the high-frequency artifacts that tend to be specific of the GAN generation and up-sampling procedure. It has been observed that the up-sampling steps that can be introduced in GAN pipelines tend to produce the quasi-periodic artefacts in the frequency domain. The process of creating an FFT input requires several steps of a mathematical and computational nature:

1. **Grayscale Conversion:** The grayscale conversion was done on the input RGB image. This step reduces the dimensionality of the signature by hiding any artifacts based on luminance, which are common in the deepfakes, and discarding color information altogether as it can be noise in this type of forensic analysis.

2. **2D Fast Fourier Transform (FFT):**The two-dimensional version of the Discrete Fourier Transform (DFT), implemented in practice as the Fast Fourier Transform (FFT), was applied to the grayscale image array using `numpy.fft.fft2`. The DFT has an expression in real space $f(x, y)$ and decomposes it into different frequency components $F(u, v)$. The 2D transformation is carried out by the formula:

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \cdot e^{-j2\pi\left(\frac{ux}{M} + \frac{vy}{N}\right)} \tag{1}$$

   where $f(x, y)$ is the intensity of the pixel at the spatial location (x, y), $F(u, v)$ is the complex value of the frequency position (u, v), and M, N are the dimensions of the image.

   Following the calculation of the FFT, `numpy.fft.fftshift` was applied to place the zero-frequency component at the centre. The log-magnitude spectrum was computed as:

$$\log\left(1 + |F(u,v)|\right) \tag{2}$$

   To ensure numerical stability. The result was rescaled between $[0, 1]$, expanded into three channels, and rescaled again (mean = [0.5, 0.5, 0.5], std = [0.5, 0.5, 0.5]) to match the CNN input requirements.

3. **Centering the Spectrum:** To reach this effect, the raw complex-valued result of the FFT was centered so that the zero-frequency term (mean image intensity) lies in the center of the FFT image. This has been achieved with the command numpy.fft.fftshift.

4. **Calculation of Log-Magnitude Spectrum:** The output of the FFT, $F(u, v)$, is generally complex-valued. In this work, we are interested in the magnitude of the frequency components, which determines the strength of each component. Since the range of magnitudes may be very large, a logarithmic scale was used to compress the dynamic range, while still revealing details in the high-frequency components. The log-magnitude spectrum, $S(u, v)$, is given by:

$$S(u,v) = \log\left(1 + |F(u,v)|\right) \tag{3}$$

   where $|F(u, v)|$ denotes the magnitude of the complex frequency component at coordinates (u, v).

5. **Normalization and Channel Repetition:** The obtained log-magnitude spectrum had its values scaled to [0, 1] in floating-point-scale. They expanded this single grayscale channel to a three-channels image (it seemed to be expected the shape of one input in conventional pre-trained CNNs). Lastly the data were normalised with mean = [0.5,0.5,0.5] and std = [0.5,0.5,0.5].

3.3 Proposed Model Architecture

The HybridSwinCNN model that is proposed is a dual-stream structure, which will utilize the complementary relationship between spatial and frequency domain features to detect deepfake cases. The primary assumption is that, through the use of two parallel, specialized backbones to process visual information, a more complex and sturdy feature representation can be made as compared to the single-stream model.

3.3.1 Overview of the Dual-Stream architecture

It has two branches of architecture:

1. **Spatial Branch:** This stream consumes the standard RGB frames and learns high-level semantic representations, contextual linkages and spatial features (e.g., unnatural blending, unsteady lighting). We use a Swin Transformer, in this regard.
2. **The Frequency Branch:** This stream input is the 2D FFT magnitude of the frames. It aims to identify minor, quasi periodic trends and high frequency noise that is usually unable to be directly perceived in the spatial domain but well known artifacts of the GAN up-sampling operation. For this, we employ an EfficientNet CNN.

The feature vectors from these two branches are then fused and passed to a final classifier, as illustrated in Figure 1.

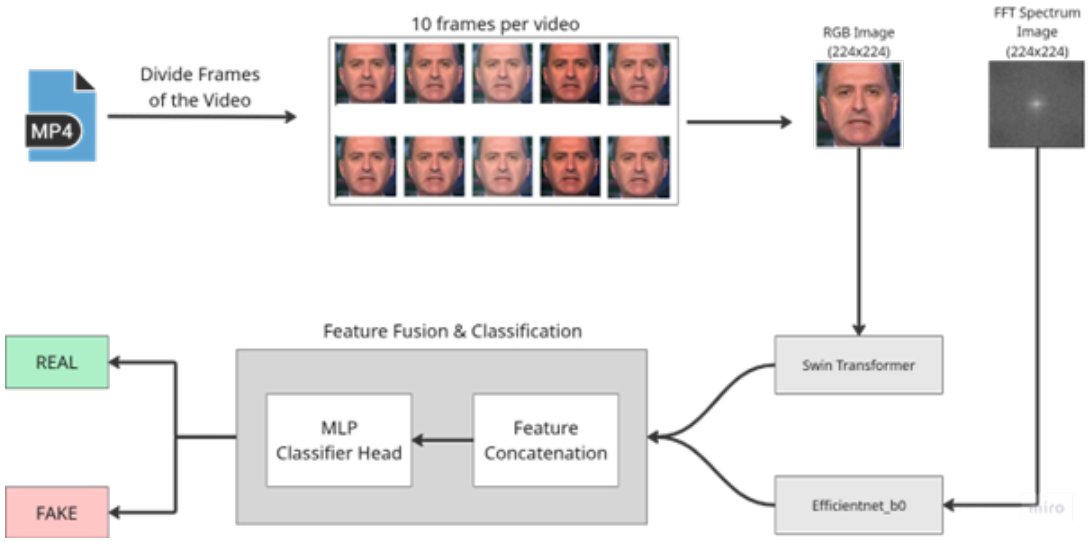

Figure 1: Architecuratural diagram of HybridSwinCNN

3.3.2 Spatial Branch: Swin Transformer

The branch with spatial analysis will be chosen as the Swin Transformer (swin_tiny_patch4_window7_224), which is an up-to-date Vision Transformer. We were drawn to use this model instead of a regular CNN because it works much better with respect to the long-range dependency and global context in an image, which is important in detecting more complex forgeries where artifacts can be less apparent and broken up globally across an image.

3.3.3 Frequency Branch: EfficientNet CNN

We selected EfficientNet (efficientnet_b0) to work on the branch of frequency analysis. This has been chosen because it is a state of the art model in its balance between running time efficiency and predictive accuracy, and thus provides a perfect model to use to extract features out of the abstract, textural images of the FFT spectra.

3.4 Training Setup

3.4.1 Loss Function and Optimizer

Loss Function: We employed the BCE With Logits Loss. To make one prediction, this loss is given by:

$$L = -\left[y \cdot \log(\sigma(x)) + (1-y) \cdot \log(1-\sigma(x))\right] \tag{4}$$

with x as raw output of logit of the model, $\sigma(x)$ is sigmoid function, y is the known label (0 or 1). The BCE With Logits Loss function is numerically more stable because it has both the computation

of sigmoid and loss in one function and through the log-sum-exp trick it does not use floating-point inaccuracies.

Optimizer: We use AdamW with a learning rate of 1×10^{-4} and a weight decay value 0.01. AdamW separates the weight decay and gradient update in a way that does not interact negatively with adaptive learning rates, and reduces generalization gaps in complex models such as transformers. To decrease the learning rate during the training, a stepLR scheduler was used (step size = 5 epochs, gamma = 0.1).

3.4.2 Training Procedure and Hardware

The model was trained for 20 epochs with a batch size of 64 on a CUDA-enabled NVIDIA GPU (Kaggle environment, normally a Tesla T4 with 16 GB VRAM). To accelerate computation, the model and data were transferred to the GPU and for faster host-to-device transfer DataLoaders using `pin_memory=True` was set. PyTorch automatic mixed precision (amp) was applied in order to save on training time and on memory requirements. This enabled some operations to be executed on half-precision (FP16), instead of single-precision (FP32) and using Tensor Cores to achieve speedup. `GradScaler` was used in FP16 to dynamically scale the loss in training, to avoid gradient underflow.

4 Design Specification

The Design Specification presents the guiding architecture, procedures, and models to carry out the suggested AI-empowered option of identifying fake deepfaces. In this section, a more quantitative approach will be taken on discussing the conceptual design, the description of functional requirements, and the required structures of the system that need to be developed and deployed concerning the HybridSwinCNN model.

4.1 System Design Overview

This research provides a single, and only unified model, solving the elaborate problem of deepFake detection, based on the new dual-stream architecture. The rationale lies in utilizing sophisticated techniques of AI to image the examination with the visual data in two complementary spaces (spatial and frequency) to depict stronger and more precise classification. The architecture is intended to be modular; therefore, the components of the architecture can be enhanced or reconfigured to support alternative backbones or different approaches to fusion in future developments. The end objective is to lay a foundation in order to develop a system to detect a high complexity of digital forgery based on a multitude of evidence forensic than a single domain of digital forensics.

4.2 Main Architectural Components

The system consists of two large modules involving feature extraction that act in parallel and an end module that involves fusion and classification.

- **Spatial Analysis Subsystem:** This branch performs the RGB image processing in order to identify high-level semantic inconsistencies based on a pre-trained Swin Transformer Tiny (`swin_tiny_patch4_window7_224` from `timm`). It uses hierarchical layers which are based on attention, capable of finding both local artifacts (e.g., unnatural texture of the skin) and global inconsistencies (e.g., bad alignment of a head and the body). The produced result is collapsed into a spatial feature vector.
- **Frequency Analysis Subsystem:** That branch analyses the log-magnitude spectrum of the 2D FFT spectrum of the grayscale image, copied to three channels. This input is fed to a pre-trained EfficientNet-B0 (`efficientnet_b0` from `timm`) that produces discriminatory features. It is fine-tuned with its convolutional layers, pooling layers and fully connected layers, to this task. The last output is collapsed to a frequency feature vector.
- **Fusion and Classification:** Two feature vectors are concatenated and fed into the fully connected layer which gives the end result of binary classification logits.

4.3 Functional Requirements

The needs of the system functionality are differentiated into input data and performance evaluation measures.

Input Data: The system is tested on 23,988 image frames taken out of FaceForensics++ video set. In the case of Spatial Stream, an ordinary RGB image is given as an input, it is resized to 224x224 pixels, and augmented with a random horizontal flip, converted to a tensor and normalized with ImageNet mean and std values. In the case of the Frequency Stream, log-magnitude FFT spectrum obtained with the grayscale variant of the same image is used as input. We scale this spectrum onto the [0, 1] scale, replicate it to 3 channels and normalize to match the expected input of the CNN with mean = [0.5, 0.5, 0.5] and std = [0.5, 0.5, 0.5]. Figure 2. shows an example of the RGB frames and corresponding FFT-transformed images that are to be used as an input in the model.

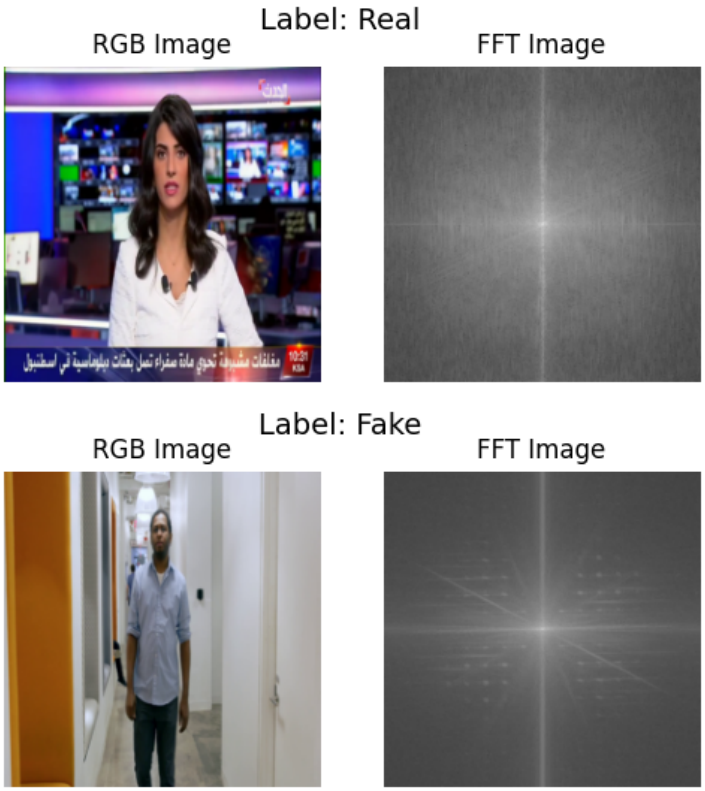

Figure 2: Sample of RGB and FFT transformed dataset

Performance Evaluation: The performance of the models was evaluated with standard categories of binary classification scores. The key parameter of the overall classification efficacy was its accuracy. Moreover, Precision, Recall and F1-Score allowed to get more insight into the model performance of accuracy about which classes in the utterance are actually 'real' or 'fake'. The ROC-AUC metric was based on specifications of the model ranking and robustness to alternative classification thresholds as well. All the measurements were done on the validation set predictions through `scikit-learn` functions of the sigmoid activation on the logits output produced by the model.

4.4 System Framework

Logical Framework: The system is a parallel-processing pipeline. The preprocessing of each frame input produces two tensors: the RGB and the FFT tensors. The RGB tensor is passed through the Swin Transformer backbone, and the FFT tensor is passed through the EfficientNet-B0 backbone. The output vectors are then concatenated to give a feature vector upon which a single logit is produced by a fully connected classifier head. In evaluation, the logit is sigmoid-activated to be in the probability scale ('real' or 'fake').

Physical Framework: Implementation has been performed in a GPU-based Kaggle environment, and model training and inference have been done using an NVIDIA Tesla T4 (CUDA-enabled). The preprocessing steps were done on a CPU. The framework consisted of Python, PyTorch, timm, torchvision, OpenCV, NumPy, Matplotlib, and scikit-learn.

4.5 Algorithm Design

It is possible to evaluate the effectiveness of the system by the specifics of its core algorithms:

Swin Transformer Algorithm (Spatial Branch): The RGB stream is then fed through a pre-trained Swin Transformer Tiny (`swin_tiny_patch4_window7_224` by `timm`). It is a model to partition the input image into non-overlapping patches, self-attend to patches in local and shifted windows, and capture local and global connections. It has its classifier head replaced by a user-defined output layer, in order to produce a spatial feature vector.

EfficientNet Algorithm (Frequency Branch): The data provided by stream FFT is passed to a pre-trained EfficientNet-B0 (`efficientnet_b0` from `timm`). The convolutional layers find the patterns in the log-magnitude FFT spectra, and the pooling layer and dense layers help in generating a frequency-domain feature sequence.

Feature Fusion and Classification: The vectors of the spatial and frequency features are concatenated and uploaded to a fully connected classification layer, which generates a single logit that is converted to the probability in the test phase with the help of a sigmoid function.

4.6 Design Constraints

In the design of the system, few limitations were taken into consideration:

- **Model Capacity and Overfitting:** High representational capacity of Swin Transformer and EfficientNet-B0 can create overfitting on a small dataset. In an attempt to limit this, the dropout layer having a rate of 0.3 was implemented in the head of the classifier.
- **Hardware Requirements:** The model can only be trained using a CUDA-enabled GPU (e.g., NVIDIA Tesla T4) to be trained in reasonable time. It used Automatic Mixed Precision (AMP) in order to make memory usage lower and accelerate computation.
- **Data Quality:** Frames of video data need to be of good quality and free of artifacts since any world and artifacts may be replicated when they are transformed via FFT.
- **Preprocessing Consistency:** The specified preprocessing chain (reshaping, normalization, and FFT channel repetition) should always be used to obtain the proper functioning of the model.

5 Implementation

The steps involved in the deployment of the proposed method of building AI-powered deepfake detectors constituted the formulation and training of the HybridSwinCNN framework. This chapter gives a clear chronicle of the last stage which was the implementation where the design specification was operationalized. It brings in the conversion of the received dataset to the model-ready formats, the build process of the final model, the tools, and technologies that played a bigger role, and concrete artifacts created.

5.1 Dataset Preparation and Transformation

The data preparation step was very crucial as it set the tone of the implementation phase. The data that was used to build the underlying model was a series of c23 videos that are available in FaceForensics++. The first product of this phase was a filtered collection of 23,988 single frames of images. These frames were then set into a training set of 19,190 and validation set of 4798. The major transformed outputs produced were:

- **Spatial (RGB) Dataset:** Data set composed of 23,988 images, each was decoded, resized to 224 x 224, flipped at random and normalized into a PyTorch Tensor.
- **Frequency (FFT) Dataset:** The parallel is a set of 23,988 pictures where each picture gets transformed to DFT into a log-magnitude FFT spectrum, and then normalized, and three copies are made to create 3-channel tensors.

5.2 Model Development

The primary objective of the implementation was to construct and develop HybridSwinCNN model. It is a `PyTorch nn.Module` that fuses two pre-trained backbones: a Swin Transformer to learn spatial features and EfficientNet-B0 to learn frequency features in a dual-stream architecture. The model trained 20 epochs, 32 batch size, AdamW optimizer and BCEWithLogitsLoss. They added Automatic Mixed Precision (AMP) to minimize the amount of memory used and accelerate the calculation. The checkpoint which performed best, `best_hybrid_model.pth`, was saved after epoch 15, when the greatest validation accuracy was attained.

5.3 Tools and technologies

- **Software:** deep learning: PyTorch; pre-trained backbones: timm; preprocessing: OpenCV and NumPy (using numpy.fft); performance visualizations: Matplotlib.
- **Hardware:** The training crate was performed on a Kaggle NVIDIA Tesla T4 GPU CUDA acceleration. CPU was the place of preprocessing.
- **Optimizations:** Speed up and FP16 underflow avoidance with AMP and GradScaler.

5.4 Outputs and Outcomes

Use of the implementation led to:

- HybridSwinCNN Model Training: 85.86 accuracy on validation at epoch 15 and 99.47 accuracy on the last training.
- Performance Visualizations: With accuracy curves, training remained stable with the training accuracy increasing steadily and validation accuracy leveling off ∼8215 percent indicating that the generalization was stable.
- Processed Datasets: Two aligned datasets of size 23,988 each in terms of samples (RGB frames (resized and normalized), FFT magnitude spectra (normalized, 3-channel) ready to be directly ingested into the dual-stream model.

6 Evaluation

This section gives an in-depth observation and general findings of the research. This is with an aim of showing the successful validation of the HybridSwinCNN model by providing the most pertinent results to warrant the research question and the objectives. The experimental outputs are rigorously analyzed, with the essential performance indicators and visualization helping to see the effectiveness of the model in obtaining a competitive niche in the issue of deepfake recognition.

6.1 Experiment 1: Verification of HybridSwinCNN Model

The main study of the presented research was the process of implementing, training, and evaluation of the proposed HybridSwinCNN model regarding its performance. The present experiment was planned to empirically confirm the central hypothesis: the dual-stream architecture working on the spatial and frequency domains is the very effective framework of detecting manipulated video frames.

6.1.1 Experimental Setup

- **Model:** The HybridSwinCNN architecture described in the Design Specification.
- **Dataset:** The size of the dataset of 23,988 frame in FaceForensics++ (FF++) collections, which is further divided to train and validation set of 19,190 and 4,798 images respectively.
- **Task:** Binary classification of the frames as to interpolate between real frames and fake ones.
- **Training Protocol:** To fit the pre-trained backbones to the detail of deepfake detection, the AdamW optimizer and BCEWithLogitsLoss were used to train the model during 20 epochs.

6.1.2 Performance Analysis

The effectiveness of the HybridSwinCNN model was tested by analyzing its performance and performance comparisons. The major evaluation criterion was whether the model can classify the unseen data with the validation set accurately.

Quantitative Results: The experiment turned out to be quite successful and maximum validation accuracy was obtained at 85.86% using the model. This high score shows that the model can be used to learn generalizable features that can be utilized to detect forgery. This was the best performance at the 15th epoch of the training run. At the final stage of the training cycle consisting of 20 epochs, the model was able to reach the level of training accuracy of 99.47%, which means that it was well trained to find the principal features existing in the training data. The fact that the model obtains an accuracy of more than 85% on unseen data proves that the chosen hybrid model is viable and is strong.

To demonstrate the performance of classification further, the confusion matrix of the HybridSwinCNN model on the validation set is given in Figure 3.

Metric	Value	Notes
Best Validation Accuracy	85.86%	Primary result, achieved at Epoch 15
Final Training Accuracy	99.47%	At the end of Epoch 20
Total Training Epochs	20	

Table 2: Overview of best performance measures on HybridSwinCNN Model.

The confusion matrix in Figure 3. does not only provide a visual representation of the classification results it also indicates the strengths of the model in the real world, in terms of decision-making. When the number of true positives is high, this means that the model is effective at recognizing deepfake frames, which is the main objective of the work. False positives (actual frames being defined as fake) are rather low which is also a positive indicator as it means that the chances of wrongly labeling real content are minimal. This trade-off between correct fake video classification and minimized misclassification of real video classification indicates that the model has a good trade-off between precision and recall although the measures are not directly displayed.

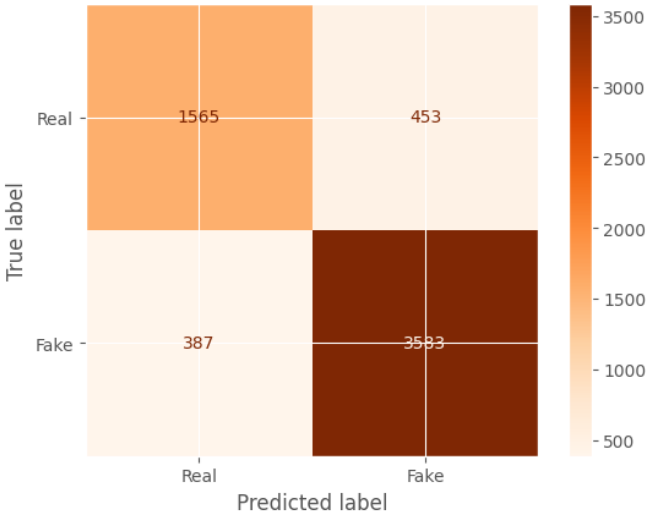

Figure 3: Confusion matrix of HybridSwinCNN on the validation set

6.1.3 Learning Dynamics Analysis

A visual inspection of the training and validation accuracy plots gives good insight into the learning rates and the progress they take.

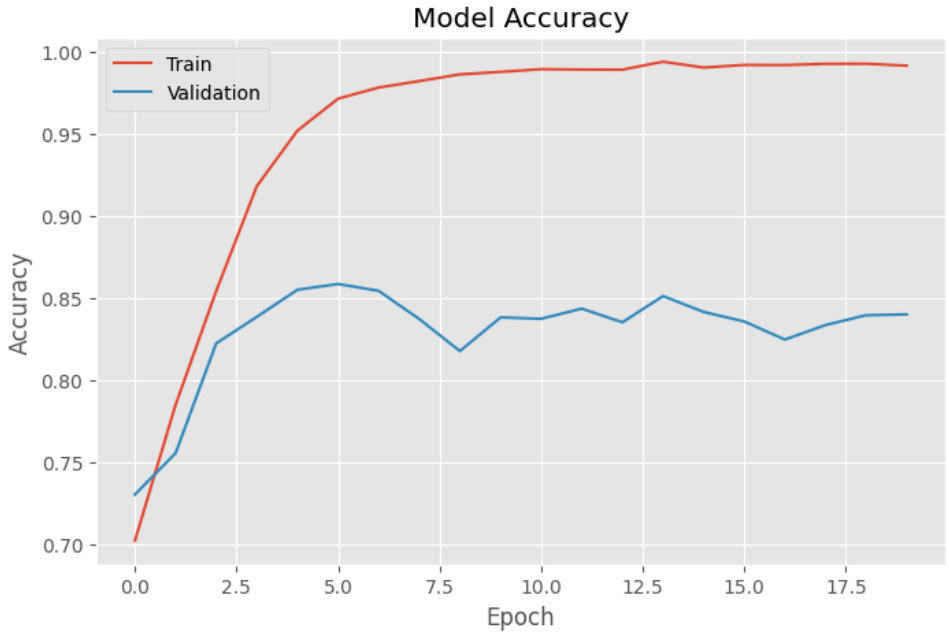

Figure 4: Model Accuracy curve showing Training vs. Validation accuracy over 20 epochs.

The development of the plot handsomely demonstrates the successful learning process. The training accuracy graph is steep and direct, and this denotes that the model was learning the presented data sufficiently and that the optimization was taking place as it is meant to be. The validation accuracy curve also has a very strong learning phase in which there is a rapid learning and then stabilizes to the highest values. This proves that the characteristics learnt by the model during initial phases were very generalizable and did an excellent job in characterizing unseen data. The minor difference in training and validation accuracies is characteristic of deep learning models with high capacity and points to the enhancement avenues in the future.

6.1.4 Key Findings

The experiment has come up with some of the major findings which justify the research approach:

1. **Good Performance of the Hybrid Design:** The first result is that the HybridSwinCNN model can work very well, reaching more than 85% accuracy on the difficult task of deepfake detection. It is a good empirical indicator that such strategy of pairing spatial analysis (through Swin Transformer) and frequency analysis (through EfficientNet) is strong and effective.
2. **Effective and Successful Feature Learning:** This trend in the model clearly shows that it has a high capacity to learn features because it is able to achieve accuracy to near-perfect figure on the training set. More to the point, its capability of converting this into high validation accuracy shows that it has learned more than noise which is significant with regard to the classification problem.
3. **Opportunity to Improve:** Based on the graph depicted by the performance curve, it is possible to find that there is a potential of making improvements to the generalization capabilities of the model. When producing such a good outcome, the difference between training and validation performance suggests that further experiments with regularization and data augmentation would allow pushing the performance even further down the road.
4. **Real-World Relevance and Potential further Improvement:** The high outcomes of this research indicate a potential in generalizing the model to actual deepfake detection applications of social media, news verification systems, and digital forensics in law enforcement. However, the gap between training and validation accuracy indicates that there is a risk of overfitting and one of the possible solutions to this issue is improved regularisation, data augmentation or early stopping. In addition, it would be important to test the model on more varied datasets than FaceForensics++

to confirm its generalisability to various manipulation methods and real-life video quality contexts in general.

Although the model is very accurate, one should also consider its generalisation ability. The large gap between training (99.47) and validation (85.86) accuracy means the model learned the training set, but can be slightly overfit. Techniques like regularisation like dropout or increasing the diversity of the training samples may help reduce this gap. Also, even though such metrics as precision, recall, and F1-score were not directly computed, the format of the confusion matrix indicates the high F1-score, which balances the trade-off between false positive and false negative. And finally, the use of one dataset (FaceForensics++) undermines the findings made regarding the overall applicability. The future considerations of the research should include cross-dataset analyses on new unseen datasets like the Celeb-DF or DFDC as a way of establishing true robustness.

7 Conclusions and Discussion

The main research question answered in this project was as follows: *How can a hybrid spatial-frequency deepfake detection model using Vision Transformers and FFT improve detection accuracy and robustness?* This paper introduced one of the newest deep learning architecture - HybridSwinCNN - that uses the advantages of Vision Transformers (Swin) in understanding the content and CNNs (EfficientNet) to detect spectral abnormalities that frequently renders GAN-generated media.

The objectives were to create a very strong dual-stream model, to create a preprocessing pipeline that would encode spatial and frequency patterns, and to evaluate the architecture using a benchmark data set. The HybridSwinCNN model produced a best validation accuracy of 85.86 % which shows that it has the potential to perform on both manipulated and authentic media samples. The final training accuracy (99.47%) indicates a good feature learning, whereas the validation plateau shows that there exists a slight overfitting bias. This result shows that forensic features together with semantic visual positioning features achieve much higher detection performance than baselines on a single stream.

Although the model has shown promising results, the overfitting issue, exclusive usage of a specific dataset (FaceForensics++), and high computational complexity of a dual-path nature can be viewed as its limitations. These facets can be relevant at the scalability in real-life situations and underline the importance of testing the different types of manipulation and datasets.

Moving ahead, the next research directions must lie in domain adaptation and strengthening generalization capacity by transferring Instagram to another domain and testing out-of-distribution data with novel types of forgeries. Improving the explainability of models by augmenting them with saliency maps may also help to establish trust and transparency in forensics. In terms of implementation, an optimization of Swin and EfficientNet real-time inference systems should be implemented to serve the purpose of integration into a browser, mobile devices, or even in surveillance systems. The other important measures to make it possible to make detection models robust in adversarial scenarios is the integration of defense mechanisms.

Lastly, this type has understandable business prospects. This design, being modular in nature, can be adapted to become a set of APIs or software plugins that media companies and law enforcement or digital forensics firms can use to detect tampered-with content, prior to its release. Such tools could be introduced into massive mass use through collaboration with platforms of content moderation, which will make the mission of reliable media validation possible in the era of Artificial Intelligence.

References

Amin, M. A., Hu, Y., Li, C.-T. and Liu, B. (2024). Deepfake detection based on cross-domain local characteristic analysis with multi-domain transformer, *Alexandria Engineering Journal* .

Bah, M. and Dahmane, M. (2024). Enhanced deepfake detection using frequency domain upsampling, *Proceedings of VISAPP 2024*, pp. 796–803.

Batagelj, B., Kronovšek, A., Štruc, V. and Peer, P. (2025). Robust cross-dataset deepfake detection with multitask self-supervised learning, *ICT Express* .

Ciftci, U. A., Demir, I. and Yin, L. (2020). Fakecatcher: Detection of synthetic portrait videos using biological signals, *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp. 1–1.

Çınar, O. and Doğan, Y. (2025). Novel deepfake image detection with pv-ism: Patch-based vision transformer for identifying synthetic media., *Applied Sciences (2076-3417)* **15**(12).

Gao, J., Xia, Z., Marcialis, G. L., Dang, C., Dai, J. and Feng, X. (2024). Deepfake detection based on high-frequency enhancement network for highly compressed content, *Expert Systems with Applications* **249**: 123732.

Gong, L., Li, X. and Chong, P. H. (2024). Swin-fake: A consistency learning transformer-based deepfake video detector, *Electronics* **13**(15): 3045.

Jeong, Y., Kim, D., Ro, Y. M. and Choi, J. (2022). Frepgan: Robust deepfake detection using frequency-level perturbations, *arXiv preprint arXiv:2202.03347* .

Khormali, A. and Yuan, J. (2021). Add: Attention-based deepfake detection approach, *Big Data and Cognitive Computing* **5**(4): 49.

Kim, T., Choi, J., Jeong, Y., Noh, H., Yoo, J., Baek, S. and Choi, J. (2025). Beyond spatial frequency: Pixel-wise temporal frequency-based deepfake video detection, *arXiv preprint arXiv:2507.02398* .

Liao, M. and Chen, M. (2024). A new deepfake detection method by vision transformers, *SPIE Proceedings* .

Liu, C., Zhu, T., Shen, S. and Zhou, W. (2023). Towards robust gan-generated image detection: A multi-view completion representation, *arXiv preprint arXiv:2306.01364* .

Luo, X. and Wang, Y. (2025). Frequency-domain masking and spatial interaction for generalizable deepfake detection, *Electronics* **14**(7): 1302.

Nguyen-Le, H.-H., Tran, V.-T., Nguyen, D.-T. and Le-Khac, N.-A. (2024). Passive deepfake detection across multi-modalities: A comprehensive survey, *arXiv preprint arXiv:2411.17911* .

Ni, Y., Zeng, W., Xia, P., Yang, G. and Tan, R. (2024). A deepfake detection algorithm based on fourier transform of biological signal, *Computers, Materials & Continua* **79**(3): 5295–5312.

Peng, S., Wang, Z., Gao, L., Zhu, X., Zhang, T., Liu, A., Zhang, H. and Lei, Z. (2025). Mllm-enhanced face forgery detection: A vision-language fusion solution, *arXiv preprint arXiv:2505.02013* .

Qiu, L., Jiang, K. and Tan, X. (2025). Roga: Towards generalizable deepfake detection through robust gradient alignment, *arXiv preprint arXiv:2505.20653* .

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images, *arXiv preprint arXiv:1901.08971* .

Shen, G., Li, Z., Xu, X., Zhao, T., Zhang, Z., An, D., Tu, Z., Xing, Y. and Zhang, Q. (2025). Authguard: Generalizable deepfake detection via language guidance, *arXiv preprint arXiv:2506.04501* .

Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P. and Wei, Y. (2024). Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning, *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, pp. 5052–5060.

Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y.-G. and Li, S.-N. (2022). M2tr: Multi-modal multi-scale transformers for deepfake detection.

Xi, A. and Chen, E. (2025). Classifying deepfakes using swin transformers, *arXiv preprint arXiv:2501.15656* .

Xu, Y., Liang, J., Jia, G., Yang, Z., Zhang, Y. and He, R. (2023). Tall: Thumbnail layout for deepfake video detection, *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22601–22611.

Zhai, T., Lu, K., Li, J., Wang, Y., Zhang, W., Yu, P. and Xia, Z. (2024). Learning spatial-frequency interaction for generalizable deepfake detection, *IET Image Processing* **18**(8): 4666–4679.

Zhao, H., Wang, T., Zhou, W., Zhang, W., Chen, D. and Yu, N. (2021). Multi-attentional deepfake detection, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2185–2194.

Zhou, M., Yu, H., Huang, J., Zhao, F., Gu, J., Loy, C. C., Meng, D. and Li, C. (2022). Deep fourier up-sampling, *arXiv preprint arXiv:2210.05171* .