National College of Ireland

From Deception to Detection: A Cybersecurity-Driven Approach to Deepfake Identification

MSc Research Project
MSc Cybersecurity

Sharan Nagaraj Kumar
Student ID: 23269839

School of Computing
National College of Ireland

Supervisor: Dr. Michael Prior

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Sharan Nagaraj Kumar
Student ID:	x23269839
Programme:	MSc Cybersecurity
Year:	2024-25
Module:	MSc Research Project
Supervisor:	Dr. Michael Prior
Submission Due Date:	11/08/2025
Project Title:	From Deception to Detection: A Cybersecurity-Driven Approach to Deepfake Identification
Word Count:	5606
Page Count:	17

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:	Sharan Nagaraj Kumar
Date:	1st August 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):	

From Deception to Detection: A Cybersecurity-Driven Approach to Deepfake Identification

Sharan Nagaraj Kumar
23269839

Abstract

The usage of deepfake technology has increased rapidly over the years, posing serious threats to businesses and people who are affected. It facilitates activities such as identity fraud, social engineering attacks, and misinformation, causing loss of reputation. This research work proposed a hybrid Deep Learning model that integrates Convolutional Neural Network, EfficientNet, and Vision Transformers to detect deepfake images. A custom dataset that includes a partial amount of the benchmarked public dataset (FaceForensics++, DFDC, and Celeb-DF-V2) and images from Google searches to cover the latest manipulated images is used to train the model. Hybrid architecture was validated against the established detection models, demonstrating an accuracy up to 93% and an Area Under the Curve (AUC) of 97.81%. Strengthening detection models and addressing challenges related to datasets like bias and high-quality data forgery were the focus of this work. This research work contributes to building more robust cybersecurity measures with the help of Artificial Intelligence to safeguard digital media integrity.

1 Introduction

1.1 History of Deepfake Technology

Over the last few decades, the integrity and authenticity of digital data have become important to prevent manipulation. Social networks made communication with multimedia easier (Heidari et al., 2023). At the same time, the spread of false information through social media platforms increased significantly. Deepfake is a type of fake content that is generated from the source media using Deep Learning (DL), such as Generative Adversarial Network (GAN), a combination of a generative network and a discriminative network. These two deep learning networks are complex and can generate images from input data which would be as realistic as the original data. Data is generated using an encoder and a decoder of a generative network, discriminative model is used to assess the authenticity of the probability between original and fake generated data. The term deepfake was first uttered in late 2017 by an anonymous Reddit user (Rana et al., 2022). Subsequently, the technology became more accessible, ease of use lowered the barrier for malicious purposes. Resemble AI report stated that in the first quarter of 2025, $200 million was lost due to deepfake-enabled fraud cases (Resemble AI, 2025). Artificial intelligence (AI) and DL were combined to replace an object in an image or video with something else that causes serious threats (Dami, 2022). The need to build a robust detection system is important.

1.2 Case Study of Cybersecurity Threats

Deepfakes can be used to impersonate a person, spread false information, and undermine trust in the digital environment. This deepfake provides an opportunity for digital scammers to create fake images or videos in their phishing and social engineering attacks. Additionally, Companies are facing problems where candidates use AI-generated profiles to create employment history and provide answers during interviews. According to Gartner, by 2028, 1 in 4 candidates in a company will be fake, which might lead to installing malware demanding ransom and stealing information (Son, 2025). $25 million in funds in a company were transferred to scammers due to video-enabled deepfake phishing. One notable case in Hong Kong is an employee from engineering firm Arup's finance department who received a phishing email from the Chief Financial Officer (CFO) to initiate transactions. Later, the employer joined a video call, where the scammers looked and sounded like the CFO and other employers. The victim was convinced that it was real and processed 15 wire transfers to scammer accounts (Young, 2024). Damage caused due to this technology goes beyond cybercrime, where insurance companies face fake accident claims.

Competitors can create fake images or videos about a brand to defame brand reputation and trust, massive loss in company values, and its stocks (Tummalapenta, 2024). Deepfakes cause National security threats by creating false videos or audio recordings of government officials, which leads to misinformation and misleads. The fake video of the Ukrainian President made troops surrender in 2022 (Kopecky, 2024). Verification of information using traditional methods like analyzing metadata and manual inspection, is outdated in detecting deepfakes generated by GANs. Collaborating with technical professionals like cybersecurity experts and regulatory bodies, and raising awareness were most important to address the challenges created by Deepfake (Alanazi et al., 2025).

1.3 Motivation and Significance of Study

The motivation of this research is to understand how deepfake creates an urgent need for proactive cybersecurity measures to prevent their spread. Unlike older vulnerabilities in code, it creates a severe threat to trust. As deepfake uses Artificial Intelligence (AI) models to create synthesized data, detection or prevention should be robust enough to detect manipulated data contents. Identifying synthesized images would help in various fields, such as Legal proceedings that need proof for each statement that is being presented. Providing evidence that the data is fake would be a turning point in the case. Forensic investigations can be strengthened when the authenticity of the data is established. Corporate Security helps to prevent impersonation fraud. Ultimately, it reflects on the brand's trust and reliability. Cybercrime prevention helps to detect phishing scams like vishing, business email compromises, and many others. This research contributes to the goal of developing a detection model that aligns with real-world needs for the detection and prevention of deepfake technology.

1.4 Research Aim and Objectives

This research aims to address the following research questions: design and evaluate a deepfake detection system.

- **RQ1:** How can a hybrid deep learning model be used to accurately detect and

attribute image-based deepfakes to improve cybersecurity measures against identity fraud, misinformation, and visual impersonation?

- **RQ2:** What other performance measures can be considered for the model to fine-tune it and check how it works in real-world implementation?

This system supports to enhance the integrity of data by accurately identifying synthesized information using deep learning techniques. The research objectives are as follows:

- Exploring the challenges and threats of deepfake technologies in cybersecurity. This has been done by covering real-world case studies and their impact.
- Utilizing three public datasets such as FaceForensics++, DFDC, Celeb-DF-V2 and combining with custom created dataset (web-collection).
- Developing a hybrid model, integrating Convolutional Neural Network (CNN), Efficient Net, and Vision Transformers (ViT) to detect whether the input image is real or synthesized.
- Assessing the system performance by accuracy, recall, precision, F1-Score, Confusion Matrix, and ROC Curve. This helps to compare the developed model with baseline models and fine-tuning.

2 Related Work

This Chapter discusses about the existing deepfake detection techniques and their limitations, the use of deep learning in classification tasks, and the limitations of current models related to cybersecurity. The rapid growing of technology created an easy way for advancement of deepfake technology. Understanding how the deepfake works would helps to build a detection model to defend against it.

2.1 Benchmarked Datasets for Deepfake Detection

FaceForensics++ is a large dataset that contains face manipulations using techniques such as DeepFakes, Face2Face, FaceSwap, and many others. It contains videos of three different qualities, which contain 1000 YouTube videos and 4000 fake videos. This dataset lacks a colour-blending process and a facial landmark mismatch. The Deepfake Detection Challenge (DFDC) is the largest facial dataset developed by Facebook, Microsoft, Amazon, and other institutions. Few of the videos in this dataset are from paid actors, and fake videos are generated using Artificial Intelligence (AI). Multiple deepfake generation techniques were used in this dataset creation, leading to the problem of differences in source and target faces. Celeb-DF-V2 is developed with 590 original videos and 5639 manipulated content using FaceSwap and DFaker. The main drawback of this dataset is that videos are used from YouTube, which contains a small amount of diversity (Asian faces) (Gong and Li, 2024). (Naitali et al., 2023) noticed that presently available benchmarked datasets lack comprehensive coverage of all aspects and focus only on basic forgery detection.
In summary, these datasets have advanced research and each of its have its characteristics, lacks in diversity. This highlights the need for traning the system with generalized data.

2.2 Deepfake Detection Models

Earlier studies have relied on developing detection models using CNN to detect deepfake images. (Rafique et al., 2021) proposed a detection system using CNN, which contained two models, such as Alex Net and Shuffle Net. After pre-processing the data, Error-Level analysis has been done to identify the compression ratio of fake and real images. Using a dataset named "Real and Fake Face detection" they achieved 86.8% accuracy via Alex Net and 88.2% via Shuffle Net, but it works fragile in real-world implementation. CNN helps to identify the distortion in the artifacts matching process (Chang et al., 2020). Researchers used CNN and Recurrent Neural Network (RNN) as their base model in research to extract inconsistencies, lighting, and textures that are unnoticed by humans (Tipper et al., 2024). It helps to extract frame-level features from the image. Detection of synthesized images can be done in two ways such as the Conventional process, which depends on handcrafted features, and DL, which works based on learned features. The drawback of using conventional methods is that the facial key points around the face show ambiguous behaviour (Seow et al., 2022). CNN architectures such as DenseNet169, DenseNet121, DENSNET201, VGG16, VGG19, VGG-Face, and ResNet50 were used to detect deepfake images generated using GAN methods. This research proved that augmented CNN-based approaches performed well and suggested efficient models improve better detection (Ahmed et al., 2022). The table below discusses the state-of-the-art methods

Table 1: Comparison of Deepfake Detection Methods

Method	**Dataset**	**Results**	**Limitations and Future Scope**
CNN-based LRCN method that detects images using temporal changes (Eye blinking) Li et al. (2018)	CEW	CNN provided good results up to 0.99 AUC	Limited to visible eye regions, future work may include other facial dynamics.
DeepD discriminator developed to detect GAN-generated images (Hsu et al., 2018)	CelebA	Precision: 0.947	Works only for GAN-based synthesized images, and the model's accuracy is only 90%.
DenseNet Pairwise learning (Hsu et al., 2020)	CelebA	Precision: up to 0.909	Proposed CFFN model handles images generated by various GAN methods, future scope includes detecting deepfake videos
CNN and LSTM (Bappy et al., 2017)	NIST16, IEEE Forensics Challenges, and Coverage	AUC: 0.7641	Lack of fixed patch-based structure, edge-dependency, and video capability

C2RNet (C-CNN + R-CNN) (Xiao et al., 2020)	CASIA, COLUMB, and FORENSICS	Precision: 0.581, 0.804, and 0.367	Focuses on local tampering regions, but lacks in single forged area per image, and its patch-based analysis lowers resilience
CNN, CNN-GRU, CNN-LSTM, and TCN (Sohail et al., 2025)	FaceForensics++	Precision: 0.917 (Without GAN) and 0.935 (With GAN)	Vulnerable to real-world complexities such as heavy compression, noise, and dataset imbalance
Deep Learning based ICP node model along with Blockchain technology (Choi and Kim, 2023)	FaceForensics++ and DFDC	AUC: 0.9263	Vulnerable to real-world complexities such as heavy compression, noise, and dataset imbalance
Lightweight CNN model (Lamichhane et al., 2022)	DFDC	Precision: 0.727 (VGG-19), 0.788 (Inception-ResNet-V2), 0.946 (Xception), and 0.923 (Model-B)	Evaluation is done only with the DFDC dataset, which lacks generalization and new manipulation methods
Deep learning model using the AdaBoost algorithm (Thirumaleshwari Devi and Rajasekaran, 2025)	DFDC	Accuracy: 86.5%	AdaBoost assigns high weights to wrongly classified images, creating more noise, which also affects the model's performance. Using AdaBoost in a hybrid model will help to overcome the above issue.

Table 1: Comparison of Deepfake Detection Methods

2.3 Deep Learning Models

Dominantly deep learning based models are used because of its ability to understand complex data (Coccomini et al., 2022) stated that recent detection methods using Vision Transformers are providing better results when it is combined with Convolutional networks. As summarized in the state-of-the-art table mentioned above, most of the models are dependent on CNN very heavily for extracting features, but there is a need for an end-to-end system designed explicitly for detecting synthesized images, but CNN focuses on local features and struggles to capture global contextual relationships, which is crucial

for face detection. Through research, (Khormali and Yuan, 2022) found that CNN is used as a backbone, where a vision transformer-based architecture would fill the research gap.

2.4 Summary and Research Gap Identifications

From the above survey, we can observe that the current benchmarked datasets have a rough quality of images or videos, along with low resolutions, which can be detected by the naked eye, resulting in a weak detection system, whereas deepfake image quality has increased significantly. The performance metrics used to test the model did not test under common real-world situations. For example, not every image will have high resolution. Testing the developed model under such circumstances will be useful to fine-tune the model for real usage. Using Vision Transformers along with CNN-based models helps to address the issue of both local and global feature extraction from facial images. This research work addresses this issue by creating a hybrid CNN-EfficientNet-ViT based model trained on generalized data, which helps in improved robustness and leverages the strength of both CNN and transformer.
As we discussed in Section1, the lack of robustness evaluation against real-world problems and not much focus on public awareness about these types of threats created a need to develop both a model and ways to effectively enrich the audience about the digital threats.

3 Methodology

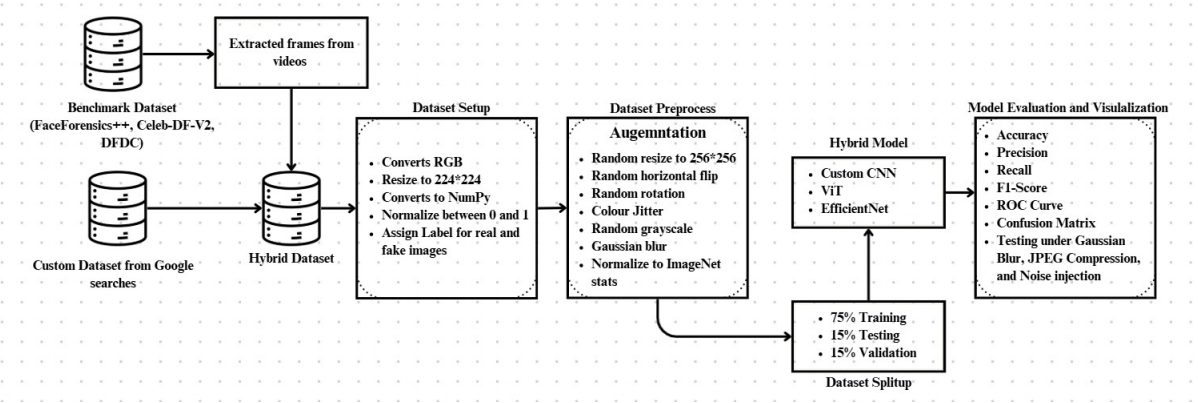

Figure 1: Architecture Diagram

3.1 Dataset Selection

This research involves multiple publicly benchmarked datasets along with images collected from Google searches to create a custom hybrid dataset. Three public datasets are used, such as FaceForensics++, DFDC, and Celeb-DF-V2. FaceForensics++ consists of 1000 original videos that are manipulated using different methods. (Rössler et al., 2019). DFDC is developed by various companies and research institutes that contain a video-based dataset to create a robust detection model (Brian Dolhansky, 2019). Celeb-DF-V2 contains synthesized videos that have a quality similar to that of those online (Li et al.,

2020). By combining all of these datasets, we can create a hybrid dataset that overcomes the overall limitations of each dataset. In addition to that, images from Google searches using keywords 'deepfake images', 'AI images', 'synthesized images', 'fake human images', and 'GAN images' were used to create a custom image dataset, which is combined with the hybrid public dataset. The images collected through web sources are manually reviewed, used only for research purposes, and labeled as real and fake. No personal information was stored. In total, 6,395 real and 6,790 fake images were used to train and test the hybrid model.

3.2 Frame Extraction

As we are using a hybrid dataset, due to computational constraints, only a subset of the dataset is used in this work. All three datasets are video-based datasets; every 5 seconds, a frame is captured and saved as an image. This approach ensured diversity across each scene, including lighting conditions. Later, these images are combined with images that are saved using Google searches.

3.3 Model Configuration

The proposed deepfake detection to address the authenticity of digital images contains a hybrid model that consists of CNN, EfficientNet, and Vision Transformers. CNN is used to capture the local features of the images that are manipulated. In simple terms, CNN is used as the backbone. EfficientNet is used to extract high-level semantics from images. Vision Transformers (ViT), a pre-trained model that requires fewer resources compared to CNN-based models, helps split images into smaller sets, similar to Natural Language Processing methods, which are fed to transformers.

3.4 Performance Metrics

As we have seen in the literature survey, the proposed models are evaluated with either accuracy or precision for image-based models. In this paper, our model is evaluated using several factors to fine-tune it.

- Accuracy: The measure of a model's classification of correctly classified images among all testing samples. Provides a way to fine-tune the model based on class imbalance.

- Precision: It indicates the number of images predicted, synthesized images, in comparison with the total number of fake images. This measurement helps to reduce false positives.

- Recall: Identifying the actual fake images detected by the model. This measure ensures synthesized content is not overlooked.

- F1-Score: Harmonic mean between Precision and Recall, helps to maintain the model's balance between false positives and false negatives.

- Confusion matrix: A visual of true positive, true negative, false positive, and false negative. Enables handling of specific misclassification patterns.

- ROC curve: Graphical representation of a model's performance across various thresholds. High ROC curve, better performance.

Robustness Evaluation: To evaluate the model's performance under real-world scenarios, image degradation was conducted under certain conditions.

- Gaussian Blur: A kernel size of 5 x 5 Gaussian blur filter is applied to images to create the effect of motion blur and defocus.

- JPEG Compression: A Few of the images spread on social media would have low quality. This method of compression reduces the image quality factor by 30.

- Noise Injection: To support low-light conditions and add noise in the images, a 0.05 standard deviation is added to the pixel value.

The robustness evaluation will help to address the **RQ2** and other metrics address the problems encountered in **RQ1** while developing a detection model to solve cybersecurity challenges. The experiment is executed in the Google Colab Pro platform with an NVIDIA Tesla T4 GPU and 51 GB of RAM.

4 Design Specification

The detection system was trained using multiple datasets to address generalization issues and used three algorithms to classify images. The complete end-to-end process of this model is discussed below:

- Initially, the custom dataset using Google searches is collected, saved in a folder. Later, benchmarked datasets such as FaceForensics++, DFDC, and Celeb-DF-V2 were used to extract images from videos. The custom hybrid dataset is prepared for data pre-processing and training the model.

- Both real and synthesized images underwent data pre-processing steps to create a uniform input to the model for better training. A detailed explanation of data pre-processing is discussed in the Implementation section.

- The pre-processed dataset is divided into training, testing, and validation to train our hybrid model. After that, model definitions for CNN, EfficientNet, and ViT were defined.

- The input for the CNN, EfficientNet, and ViT is processed separately. Feature vectors of these are concatenated before passing to the final layer for classification.

- After training the model, using the performance metrics, the hybrid model is fine-tuned multiple times and tested under robustness conditions as defined above.

- The trained model is hosted on the HuggingFace website for public use, which will address the gap identified in cybersecurity measures and can be used as evidence in legal proceedings.

5 Implementation

The hybrid model was developed using a Google Colab Pro environment, which begins with frame extraction, data pre-processing, data splitting, and model training. The detailed process of developing is discussed below:

5.1 Dataset Generation

Frames were extracted from benchmarked datasets such as FaceForensics++, DFDC, and Celeb-DF-V2. Using OpenCV, every 5 seconds, pictures are captured and saved in a folder accordingly. Using a metadata file present along with the dataset, appropriate fake video frames are also captured. A custom dataset was created using Google searches and saved in the same folder where benchmarked images are saved. To avoid the data imbalance, almost equal numbers of real and synthesized images are used in the dataset. Finally, the dataset is converted to a ZIP file and saved to Google Drive.

5.2 Data Pre-Processing

To create a unified training dataset for the model, the following pre-processing steps are followed:

- Pre-processing steps such as converting to RGB, resizing to 224x224, and normalization between 0 and 1 were done.
- Using NumPy, the pre-processed images and their labels are stored to ensure efficient loading during model training and reduce the time for processing raw images. 13,185 images are used to train and validate the model.
- To reduce overfitting, an augmentation process for testing data, such as random horizontal, random flip 20°, colour jitter techniques like brightness, contrast, saturation, and hue, applied Gaussian blur, and ImageNet normalization values are applied as per PyTorch recommendation.
- The split was made into three parts, 'train_data', 'test_data', and 'val_data' with 70%, 15%, and 15%, respectively.
- For testing and validation parts, only resizing and normalization are applied as an augmentation process to test the performance of the model.

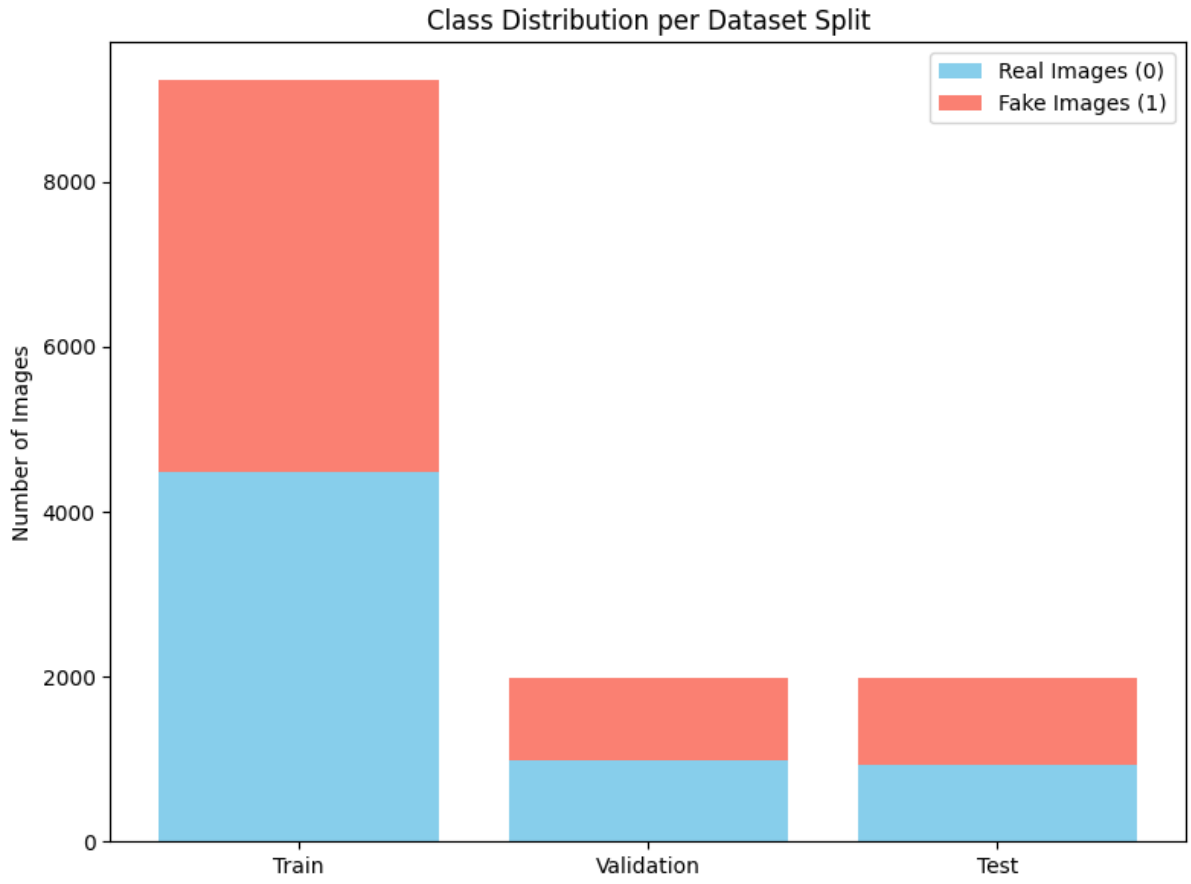

Figure 2: Image count in each part of the dataset after pre-processing

5.3 Model Definition and Training

5.3.1 Convolutional Neural Network

The CNN architecture contains 4 layers, trained from scratch without pre-trained parameters. The input for this model is normalized to 256x256. The layers of CNN include a set of convolutional layers, ReLU activation, max pooling, and dropout. Other technical information about the model is defined in the Table 2. Decision to use CNN as a baseline model because it can focus on deepfake-specific patterns. In addition to that, batch normalization layers were included after the convolution to accelerate training.

5.3.2 Vision Transformers and EfficientNet

The pre-trained model ViT named 'vit_base_patch_16_224' and EfficientNet-B0 are used in this research. Both of them are imported through the Timm library. ViT model processes the images in batches and captures global context dependencies and spatial relationships. All three of them produce an embedded output vector which is then processed with a Relu activation layer with a dropout value of 0.3, and finally softmax layer is applied for classification.
The hybrid model was trained using 50 epochs with a batch size of 64. Function named 'ReduceLROnPlateau' is used to reduce the learning rate (1×10^{-4}), if the validation loss does not improve for 3 epochs. To ensure generalization, the best model is captured in every epoch by comparing it with the previous best model.

Outputs of the model are passed through the fusion layer using Adam. This ensures that the hybrid system decides its results by collecting features such as spatial, contextual, and semantic before providing results.

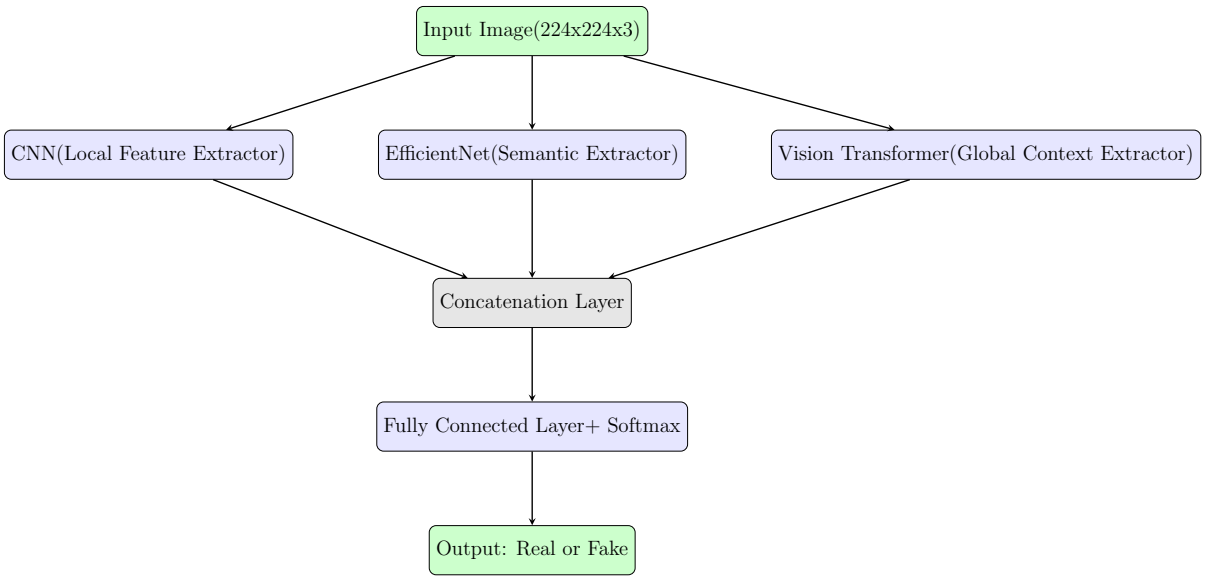

Figure 3: Hybrid Model Architecture for Deepfake Detection.

Parameters	**CNN**	**ViT**	**EfficientNet-B0**
Mini-batch size	32	32	32
Learning rate	0.001	3e-5	1e-4
Hidden Layer	[32,64,128,256]	768	1280
Activation	Relu	Gelu	Swish
Output Layer	Softmax	Softmax	Softmax
Loss Function	CrossEntropy	CrossEntropy	CrossEntropy
Optimizer	Adam	AdamW	Adam

Table 2: Model Configuration

6 Evaluation

The performance of our model is evaluated with various metrics, which are used in various deep learning methods. As per the Table 3, the accuracy of our model outperformed the current state-of-the-art, even when we had diversity of datasets, such as variation in image quality and multiple augmentation techniques. In addition to that, the Precision and Recall scores were also high, 92.75% and 93.2% respectively, which implies the model performs its best under false positives and false negatives.

Accuracy	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
92.72%	94.54%	91.22%	92.85%	97.53%

Table 3: Performance Metrics of the Hybrid Model

The ROC curve, 97.81% showed that the model was able to identify and classify the images effectively. This indicates that a hybrid model trained using diverse images is working properly. Additionally, we evaluated the model's performance in the presence of low-quality images, noise, and Gaussian blur. Most of the papers surveyed in the Table 1 did not have these metrics. This helps to test our model as it works in the real world. The model was tested under two different values in each technique, as shown in Table 4.

Gaussian Blur	Noise Injection	JPEG Compression
49.22% (kernel size=5)	54.47% (std=0.02)	49.32% (quality=70)
49.22% (kernel size=9)	54.52% (std=0.05)	49.32% (quality=30)

Table 4: Results of Robustness Evaluation

Reduction in accuracy when Gaussian Blur and other techniques are applied is due to the inability of the model to understand the motion artifacts, which can be improved in future works. Despite these degradations, the model performed at its best, which depicts generalization and considers real-world problems.

6.1 Performance Metrics

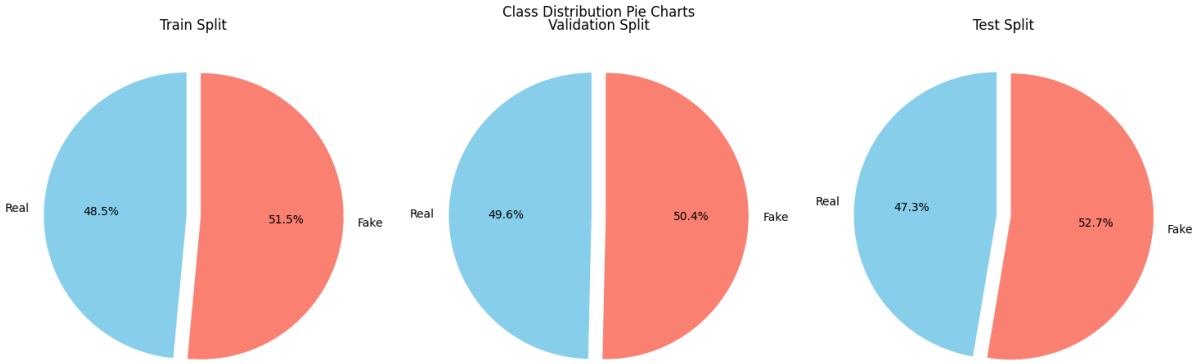

Figure 4: Class Distribution of Dataset

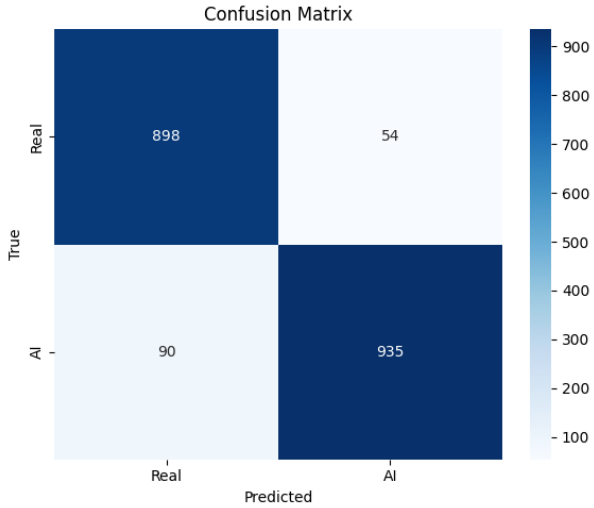

Figure 5: Confusion Matrix of Hybrid Model

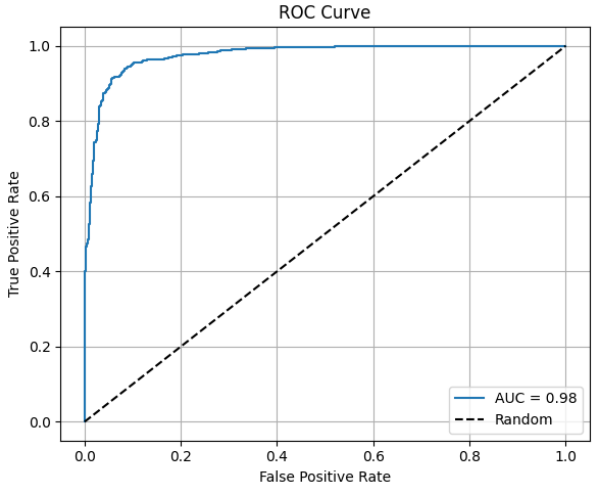

Figure 6: ROC Curve

Figure 7: Grade-CAM highlights the model's focus on facial feature regions

The total time taken to run the project, from data pre-processing to robustness evaluation, is approximately 3 hours. Figure 5 shows that the model misclassifies less than 10% of the fake images as real. Figure 7 visually explains where the hybrid model focuses during detection. This image is generated using the Grad-CAM heatmap technique. As we can see, our model focuses on parts where meaningful patterns are available, such as facial features.

6.2 Discussion

The results of this research work highlight that the hybrid model combines Convolutional Neural Network, Vision Transformers, and EfficientNet effective in classifying deepfake images. The model achieved the highest accuracy when compared to Table 1, and also the evaluation metrics of our model added an advantage of fine-tuning the model for better performance. The Grad-CAM heatmap contributed to correctly identifying the place where our model focuses. Initially, the model was not performing and classifying well due to fewer images for training. Further, the dataset increased from around 2,000 to 13,000 for better accuracy.
Limitations such as storage and computational issues, the dataset is converted into images and used to train our model. In addition to that, facial expressions such as blinking eyes and actions in objects are missed. Benchmarked datasets that are currently present are mostly outdated because of the evolving multiple technologies of creating deepfake images or videos.

6.3 Cybersecurity Relevance and Defensive Measures

The hybrid model directly addresses the cybersecurity threats by detecting the AI images by leveraging the strength of deep learning algorithms. Integrating these types of models in applications like Instagram, Facebook, YouTube, and other social media platforms for real-time detection of synthesized content helps in mitigating phishing and brand impersonation attacks. Before posting content on public platforms, integrity checks must be done before publishing it. Most of the victims are not aware of this type of threat, way to identify it, and take action against it. Government and Organizations should take awareness-related actions to reduce the impact and number of victims in the future by conducting sessions, defining methods to report these activities.
Whenever a person receives content that looks suspicious or needs to verify the authenticity, there are several platforms available online that work based on AI-based models to identify deepfake-generated images. Similar to that, the model developed through this research work is hosted on the HuggingFace website for public use. This helps to collaborate with other researchers to enhance the model to defend against all types of deepfakes.

7 Conclusion and Future Work

The main aim of this work is to propose a detection model that addresses the issues related to deepfake cybersecurity threats. Initially, a custom hybrid dataset is created using 3 benchmarked datasets and a web collection. Frames were extracted from videos, saved as images to train the model. Several pre-processing techniques were applied to create a uniform input to the model for training. This hybrid model contained a Convolutional

Neural Network and pre-trained Vision Transformers and EfficientNet. After developing the model, the performance of the model is evaluated using multiple factors to fine-tune and create a robust model. The results provided by the hybrid model were more than the present one, up to 93% accuracy and 97.53% of AUC. The robustness evaluation results showed how the model performs when there is image degradation, which reflects the real world. Based on these results, the model is fine-tuned several times. However, only images were used to train the model due to computational issues. Future works can focus on this area, and also use a database to save the images if they are new to the model.

References

Ahmed, S. R., Sonuç, E., Ahmed, M. R. and Duru, A. D. (2022). Analysis survey on deepfake detection and recognition with convolutional neural networks, *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1–7.

Alanazi, S., Asif, S., Caird-Daley, A. and Moulitsas, I. (2025). Unmasking deepfakes: A multidisciplinary examination of social impacts and regulatory responses, *Human-Intelligent Systems Integration* .
URL: *https://doi.org/10.1007/s42454-025-00060-4*

Bappy, J. H., Roy-Chowdhury, A. K., Bunk, J., Nataraj, L. and Manjunath, B. (2017). Exploiting spatial structure for localizing manipulated image regions, *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 4980–4989.

Brian Dolhansky, Russ Howes, B. P. N. B. C. C. F. (2019). The deepfake detection challenge (dfdc) preview dataset.

Chang, X., Wu, J., Yang, T. and Feng, G. (2020). Deepfake face image detection based on improved vgg convolutional neural network, *2020 39th Chinese Control Conference (CCC)*, pp. 7252–7256.

Choi, N. and Kim, H. (2023). Dds: Deepfake detection system through collective intelligence and deep-learning model in blockchain environment, *Applied Sciences* **13**(4).
URL: *https://www.mdpi.com/2076-3417/13/4/2122*

Coccomini, D. A., Messina, N., Gennaro, C. and Falchi, F. (2022). Combining efficientnet and vision transformers for video deepfake detection, *in* S. Sclaroff, C. Distante, M. Leo, G. M. Farinella and F. Tombari (eds), *Image Analysis and Processing – ICIAP 2022*, Springer International Publishing, Cham, pp. 219–229.

Dami, L. (2022). Analysis and conceptualization of deepfake technology as cyber threat.

Gong, L. Y. and Li, X. J. (2024). A contemporary survey on deepfake detection: Datasets, algorithms, and challenges, **13**: 585.
URL: *https://www.mdpi.com/2079-9292/13/3/585*

Heidari, A., Navimipour, N. J., Dag, H. and Unal, M. (2023). Deepfake detection using deep learning methods: a systematic and comprehensive review, **14**. Publisher: Wiley-Blackwell.

Hsu, C.-C., Lee, C.-Y. and Zhuang, Y.-X. (2018). Learning to Detect Fake Face Images in the Wild , *2018 International Symposium on Computer, Consumer and Control (IS3C)*, IEEE Computer Society, Los Alamitos, CA, USA, pp. 388–391.
URL: *https://doi.ieeecomputersociety.org/10.1109/IS3C.2018.00104*

Hsu, C.-C., Zhuang, Y.-X. and Lee, C.-Y. (2020). Deep fake image detection based on pairwise learning, *Applied Sciences* **10**(1).
URL: *https://www.mdpi.com/2076-3417/10/1/370*

Khormali, A. and Yuan, J.-S. (2022). Dfdt: An end-to-end deepfake detection framework using vision transformer, *Applied Sciences* **12**(6).
URL: *https://www.mdpi.com/2076-3417/12/6/2953*

Kopecky, S. (2024). Challenges of deepfakes, pp. 158–166. Publisher: Springer International Publishing.

Lamichhane, B., Thapa, K. and Yang, S.-H. (2022). Detection of image level forgery with various constraints using dfdc full and sample datasets, *Sensors* **22**(23).
URL: *https://www.mdpi.com/1424-8220/22/23/9121*

Li, Y., Chang, M.-C. and Lyu, S. (2018). In ictu oculi: Exposing ai created fake videos by detecting eye blinking, *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7.

Li, Y., Sun, P., Qi, H. and Lyu, S. (2020). Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics, *IEEE Conference on Computer Vision and Patten Recognition (CVPR)*, Seattle, WA, United States.

Naitali, A., Ridouani, M., Salahdine, F. and Kaabouch, N. (2023). Deepfake attacks: Generation, detection, datasets, challenges, and research directions, **12**: 216.
URL: *https://www.mdpi.com/2530758*

Rafique, R., Nawaz, M., Kibriya, H. and Masood, M. (2021). Deepfake detection using error level analysis and deep learning, *2021 4th International Conference on Computing Information Sciences (ICCIS)*, pp. 1–4.

Rana, M. S., Nobi, M. N., Murali, B. and Sung, A. H. (2022). Deepfake detection: A systematic literature review, **10**: 25494–25513.

Resemble AI, R. (2025). Q1 2025 deepfake incident report: Mapping deepfake incidents.
URL: *https://www.resemble.ai/wp-content/uploads/2025/04/ResembleAI-Q1-Deepfake-Threats.pdf*

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J. and Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images, *International Conference on Computer Vision (ICCV)*.

Seow, J. W., Lim, M. K., Phan, R. C. and Liu, J. K. (2022). A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities, *Neurocomputing* **513**: 351–371.
URL: *https://www.sciencedirect.com/science/article/pii/S0925231222012334*

Sohail, S., Sajjad, S. M., Zafar, A., Iqbal, Z., Muhammad, Z. and Kazim, M. (2025). Deepfake image forensics for privacy protection and authenticity using deep learning, *Information* **16**(4).
URL: *https://www.mdpi.com/2078-2489/16/4/270*

Son, H. (2025). Fake job seekers are flooding u.s. companies that are hiring for remote positions, tech CEOs say.
URL: *https://www.cnbc.com/2025/04/08/fake-job-seekers-use-ai-to-interview-for-remote-jobs-tech-ceos-say.html*

Thirumaleshwari Devi, B. and Rajasekaran, R. (2025). Deepfake video detection using ada-boosting on the dfdc dataset, *Procedia Computer Science* **258**: 1091–1101. International Conference on Machine Learning and Data Engineering.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050925014462*

Tipper, S., Atlam, H. F. and Lallie, H. S. (2024). An investigation into the utilisation of cnn with lstm for video deepfake detection, *Applied Sciences* **14**(21).
URL: *https://www.mdpi.com/2076-3417/14/21/9754*

Tummalapenta, S. (2024). New wave of deepfake cybercrime.
URL: *https://www.ibm.com/think/insights/new-wave-deepfake-cybercrime*

Xiao, B., Wei, Y., Bi, X., Li, W. and Ma, J. (2020). Image splicing forgery detection combining coarse to refined convolutional neural network and adaptive clustering, *Information Sciences* **511**: 172–191.
URL: *https://www.sciencedirect.com/science/article/pii/S0020025519308977*

Young, K. (2024). Cyber attack case study: Deepfake scammers con company.
URL: *https://coverlink.com/case-study/cyber-attack-case-study-deepfake-scammers-con-company/*