IMAGE-BASED DEEPFAKE DETECTION SYSTEM

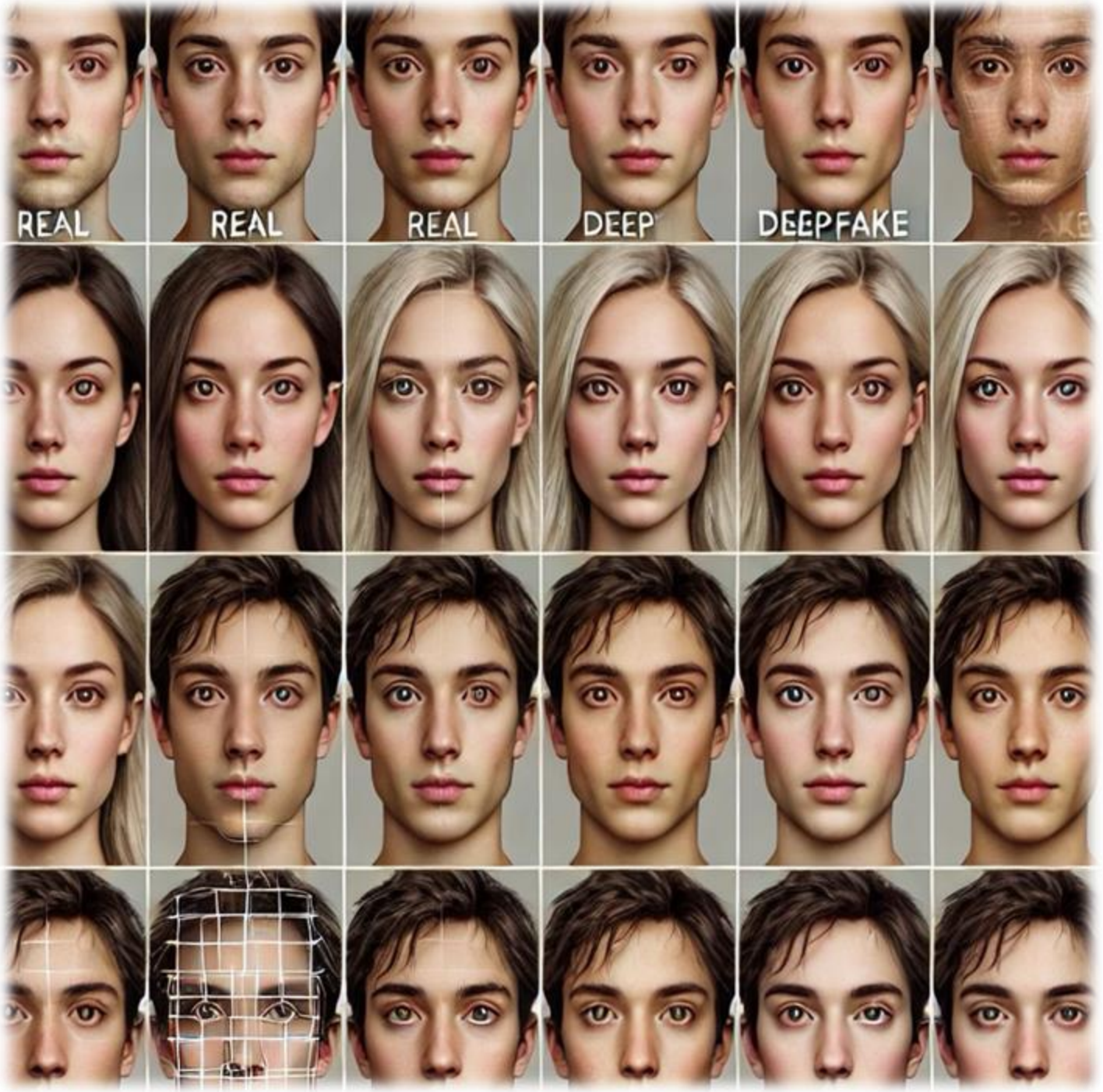
By Group 9:

Parth Malik

Shushant Ghosh

Suprajasai Konegari

Supriya Konegari



INTRODUCTION

Deepfake video of Facebook CEO Mark Zuckerberg posted on Instagram

'Scary': How a woman discovered deepfakes of herself

Deepfake scams have robbed companies of millions. Experts warn it could get worse

PUBLISHED MON, MAY 27 2024•10:20 PM EDT | UPDATED TUE, MAY 28 2024•5:31 AM EDT

AI-driven scams are expected to surge ahead of Black Friday.

AI-driven scams and deepfake technology are more enhanced than ever this holiday season.

Finance worker pays out $25 million after video call with deepfake 'chief financial officer'

INTRODUCTION

What is Deepfake Detection?

- Deepfakes create realistic fake content, blurring the line between real and manipulated media.
- They are used in scams, cybercrimes, and identity theft, like fake interviews or fraudulent transactions.

Challenges with Existing Solutions

- Traditional methods struggle with subtle, high-quality manipulations.
- Outdated techniques can't handle advanced deepfake technologies.

Necessity of Improved Solutions

- Vision and Swin Transformers offer better accuracy and adaptability for detection.
- Essential for ensuring authenticity in online banking, news, and social platforms.

PHASE 1 - RECAP

Model Training and Dataset

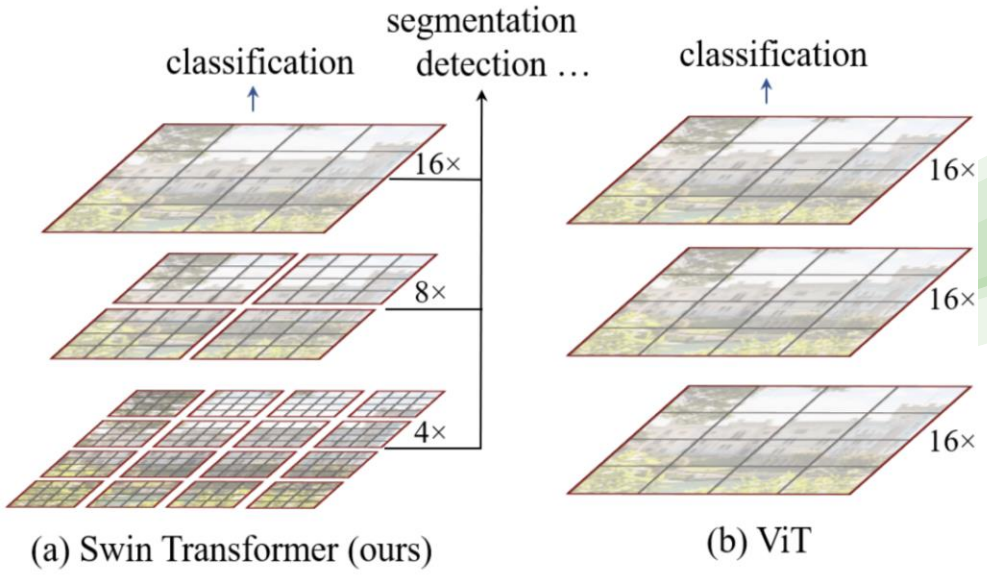
- Models Trained: CNN and ViT
- Dataset used:
  - *FaceForensics*++ for authentic and manipulated images
  - GAN-generated images

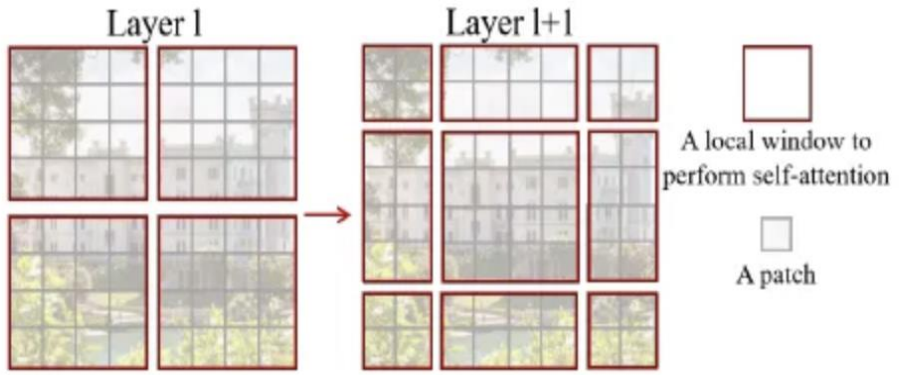
Results Overview

Metric	CNN	ViT	Better Model
Precision	**87.2%**	86.4%	CNN
Recall	86.1%	**92.1%**	ViT
F1-Score	86.65%	**89.2%**	ViT
ROC-AUC	0.95	**0.9642**	ViT

ViT vs SWIN TRANSFORMER

- **Hierarchical:** SWIN captures multi-scale features; ViT lacks hierarchy.
- **Shifted Attention:** SWIN refines context locally; ViT relies on global attention.
- **Multi-Scale:** SWIN detects subtle patterns; ViT struggles with fine details.
- **Context Sharing:** SWIN connects regions; ViT lacks cross-region interaction.
- **Integration:** SWIN balances features; ViT favors global patterns.





An illustration of the shifted window approach for computing self-attention in the proposed Swin Transformer architecture

Fig 2. ViT vs SWIN Transformer

SWIN TRANSFORMER - MODIFICATIONS

Frozen Early Stages (1 & 2)
Retains pretrained generic ImageNet features (edges, textures) & Preserves foundational knowledge, avoiding overfitting.

Fine-Tuned Later Stages (3 & 4)
Captures task-specific deepfake anomalies (e.g., subtle pixel distortions, unnatural texture transitions, or irregular lighting patterns) & Adapts model to detect nuanced manipulations.

Custom Binary Classifier
Optimized for real vs. fake detection. Improves prediction accuracy for deepfake tasks.

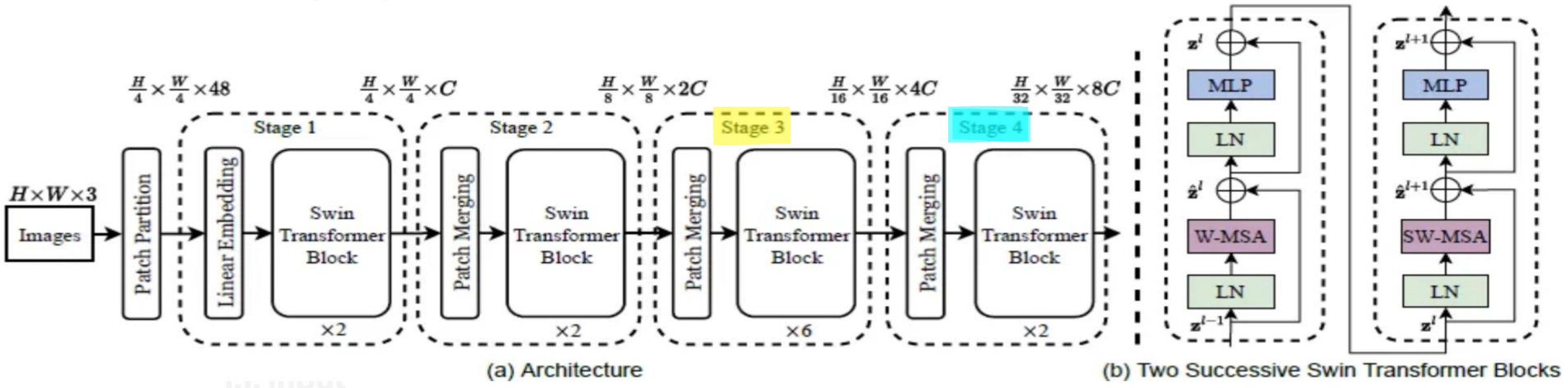


Fig 3. SWIN Transformer architecture

TRAINING : CNN Vs ViT Vs SWIN

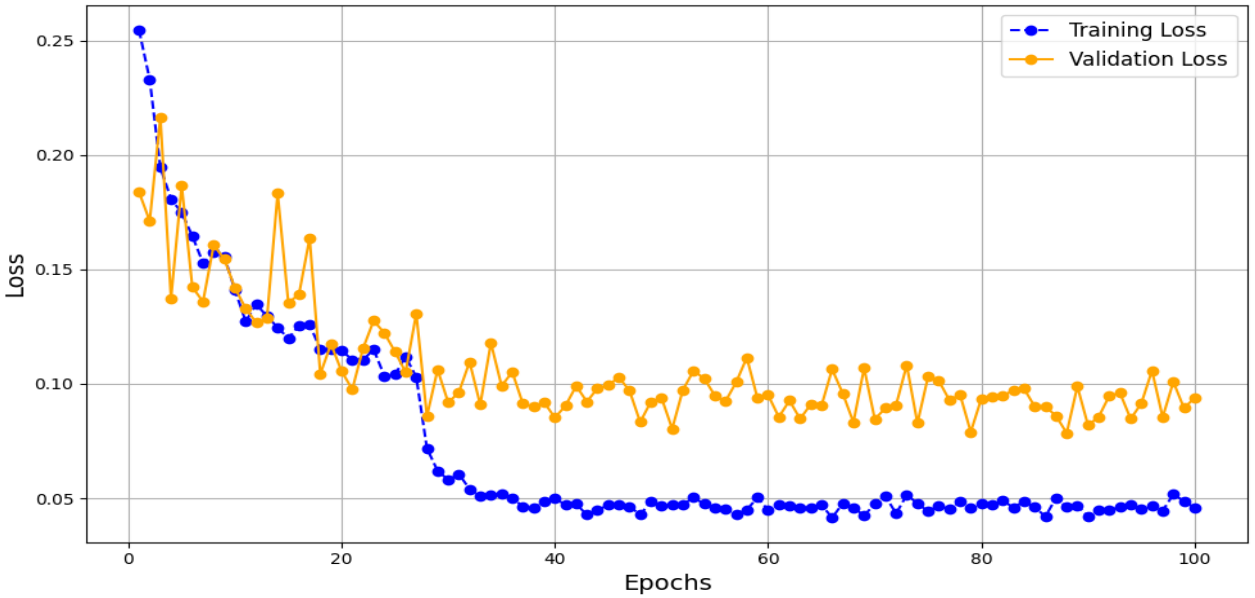


Fig 4.SWIN Training vs validation loss

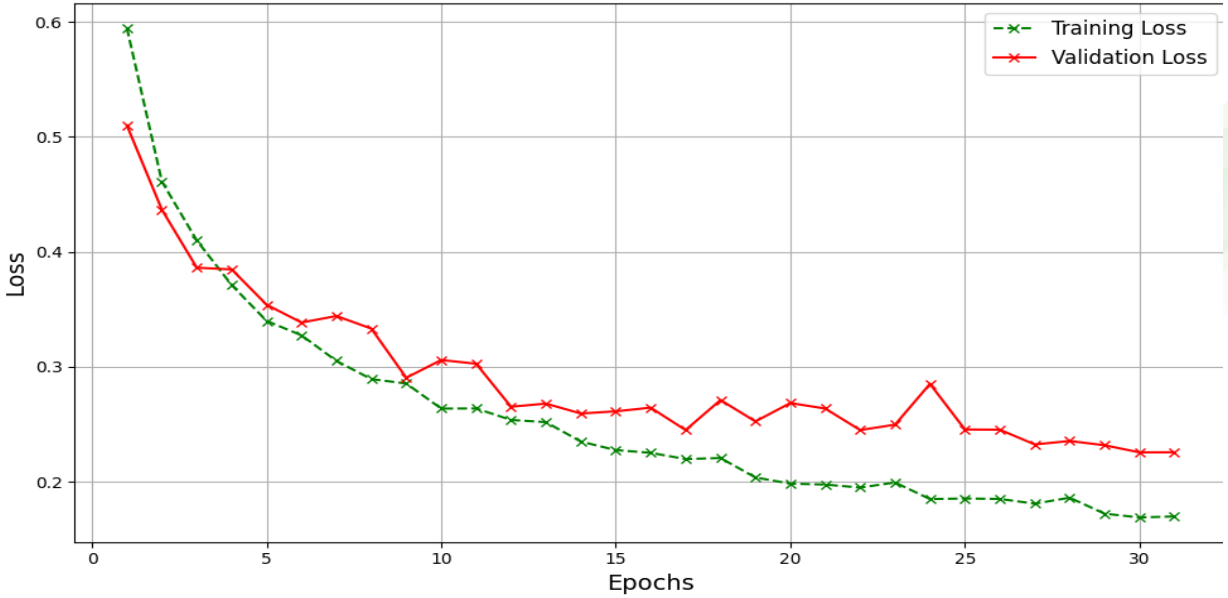


Fig 5.ViT Training vs validation loss

Insights:

- **SWIN:** Optimized F1 & Recall, stable validation loss.
  Outperforms others but **3x longer** training time.
- **ViT:** Optimized for Recall, generalizes well.
- **CNN:** Optimized for Recall, shows signs of overfitting.
  Unstable validation performance.

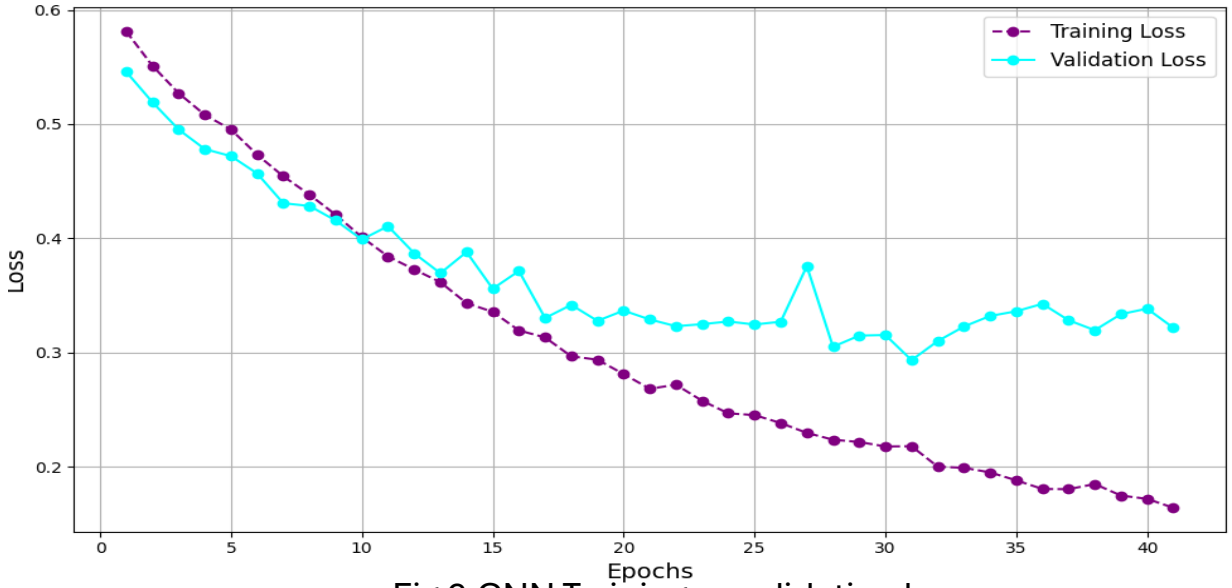


Fig 6.CNN Training vs validation loss

RECALL vs F1 OPTIMIZED SWIN RESULTS

Metrics	Recall Based Swin	F1 Based Swin
Recall	**.9798**	**.9637**
Precision	.9034	.9628
F1	**.9401**	**.9632**
Accuracy	.9375	.9632
ROC	.9900	.9957

Predicted label

	Real	Fake
Real	1784	201
Fake	48	1937

Fig 7. Recall Optimized Confusion Matrix

Predicted label

True label

	Real	Fake
Real	1914	71
Fake	64	1921

Fig 8. F1 Optimized Confusion Matrix

Insight:

The **F1-Optimized Swin Model** is the better choice for its balanced, high-performance metrics across all categories.

RECALL vs F1 OPTIMIZED SWIN RESULTS

RECALL OPTIMIZED

Fig 9. Correctly classified fake images

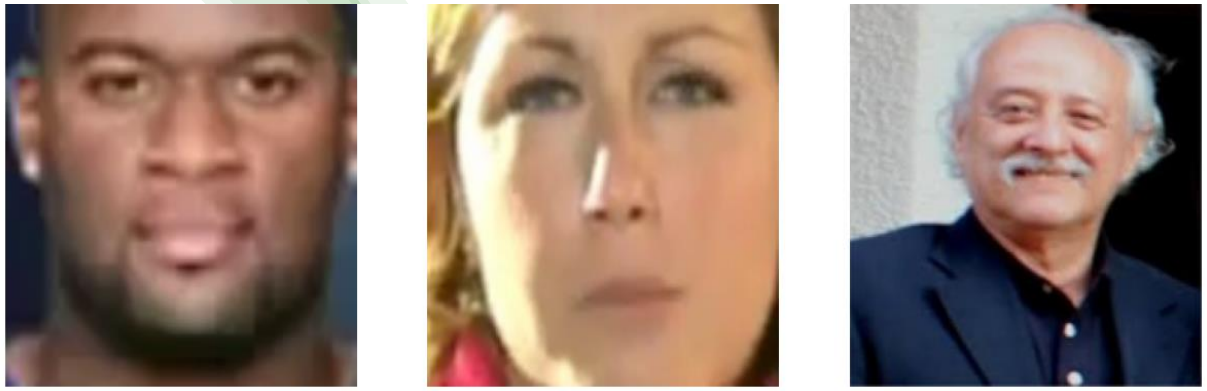


Fig 11. Misclassified real images

F1 OPTIMIZED

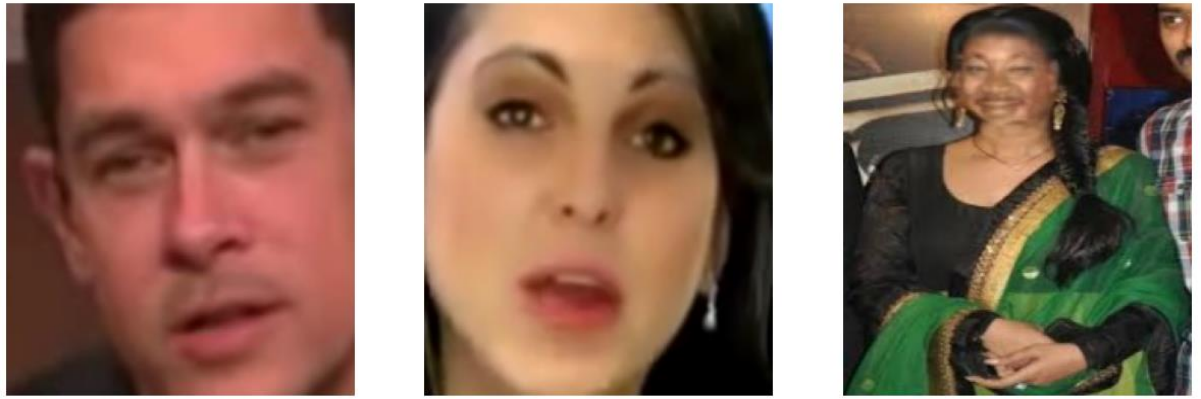


Fig 10. Correctly classified fake images

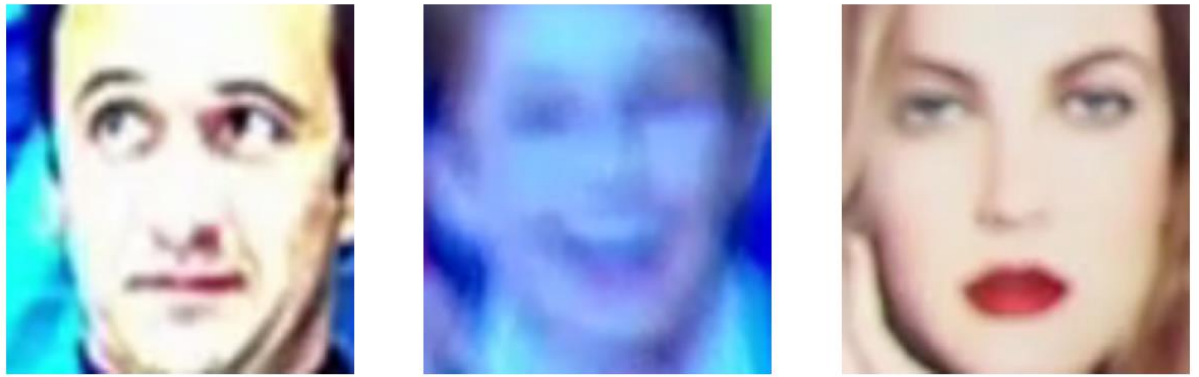


Fig 12. Misclassified real images

RECALL vs F1 OPTIMIZED SWIN RESULTS

RECALL OPTIMIZED

Fig 13. Correctly classified real images

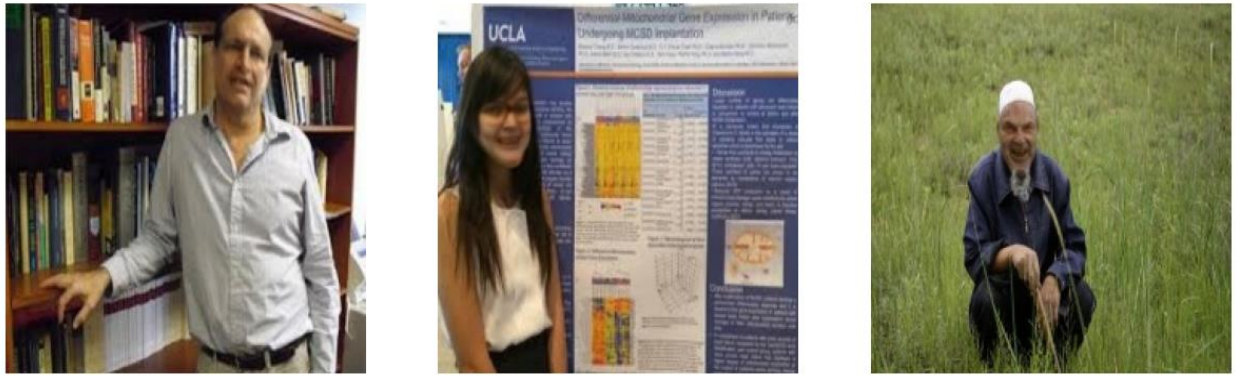


Fig 15. Misclassified fake images

F1 OPTIMIZED

Fig 14. Correctly classified real images

Fig 16. Misclassified fake images

PHASE 1 to PHASE 2: ViT vs SWIN – PERFORMANCE UPTICK

Metrics	ViT	SWIN
Recall +(4.27%)	**.9798**	**.9637**
Precision +(9.88%)	**.9034**	**.9628**
F1 +(7.12%)	**.9401**	**.9632**
Accuracy +(7.52%)	.9375	.9632
ROC +(3.17%)	**.9900**	**.9957**
Test Loss -(60.29%)	0.2491	0.0989

Insights:

- False Negatives & Positives significantly decreased
- True Negatives & Positives increased

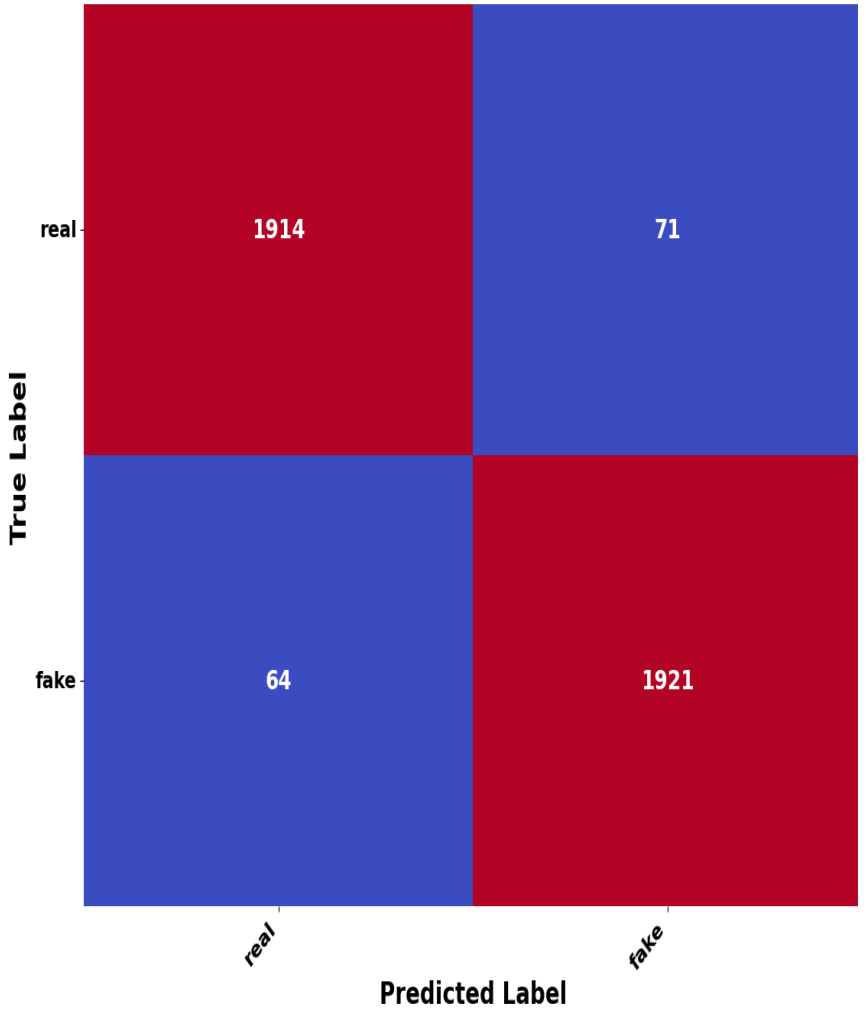


Fig 17. SWIN Confusion Matrix

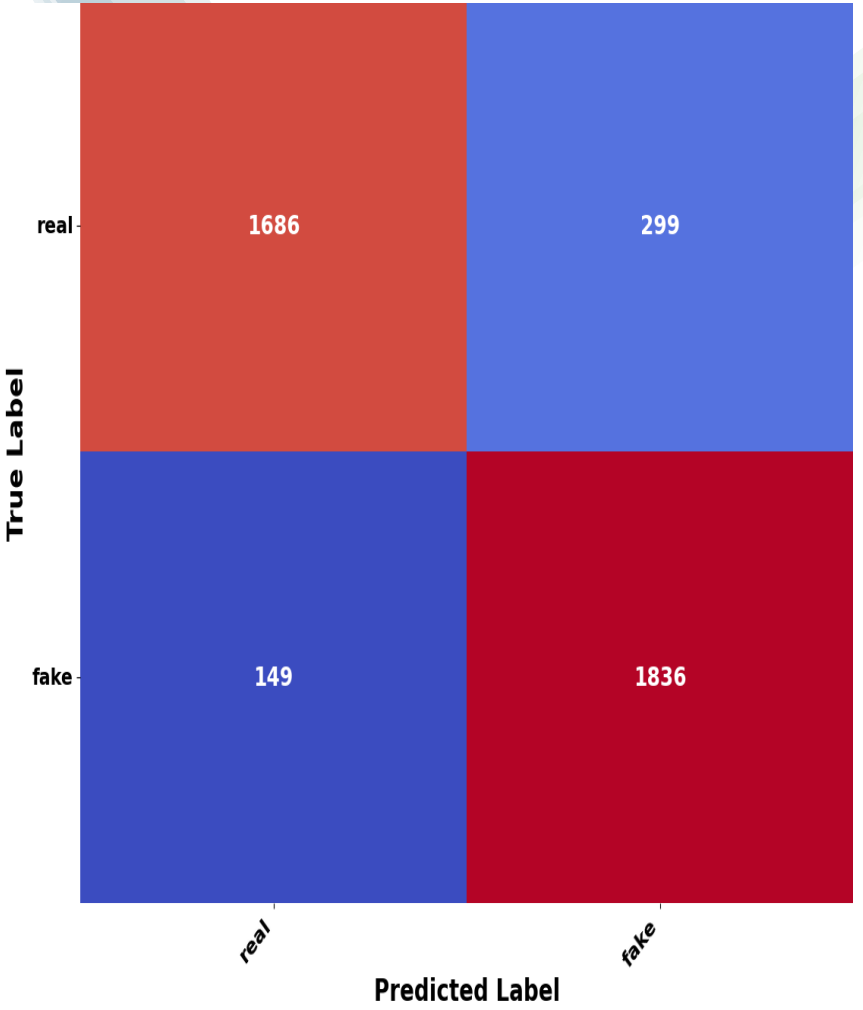


Fig 18. ViT Confusion Matrix

COMPARATIVE ANALYSIS

Strengths

CNN

- **Precision (87.2%)**: Reliable in identifying fakes with fewer false positives, better than ViT but lower than Swin.
- **ROC AUC (0.95)**: Strong at distinguishing real vs. fake, but lower than both ViT and Swin.

ViT

- **Recall (92.1%)**: Captures more fake images than CNN but fewer than Swin.
- **F1-Score (89.2%)**: Balances precision and recall better than CNN, but not as well as Swin.
- **ROC AUC (0.9642)**: Stronger than CNN, but not as strong as Swin, at distinguishing real vs. fake.

Swin

- **Precision (96.28%)**: Most reliable with fewest false positives.
- **Recall (96.37%)**: Best at capturing fake images.
- **F1-Score (96.32%)**: Best balance of precision and recall.
- **ROC AUC (0.995)**: Exceptional at distinguishing real vs. fake.
- **Accuracy (96.32%)**: Highest overall classification accuracy.

COMPARATIVE ANALYSIS

Weaknesses

CNN

- **False Positives (180)**: Misclassifies real images as fake, fewer than ViT but more than Swin.
- **Recall (86.1%)**: Misses more fake images than both ViT and Swin.
- **Precision (87.2%)**: Lower than Swin, though slightly higher than ViT.

ViT

- **False Positives (288)**: Misclassifies more real images as fake than both CNN and Swin.
- **Precision (86.4%)**: Lower than CNN and Swin, leading to more false positives.
- **Recall (92.1%)**: Still lower than Swin, missing some fake images.

Swin

- **False Positives (71)**: Misclassifies real images as fake, but fewer than ViT and CNN.
- **Precision (96.28%)**: Best out of the other two models, but not perfect.
- **Recall (96.37%)**: Despite being the best, some fake images may still be missed.

COMPARATIVE ANALYSIS

Model Comparison

Metric	CNN	ViT	SWIN	Best Model
Recall	86.1%	92.1%	96.37%	SWIN
Precision	87.2%	86.4%	96.28%	SWIN
F1-Score	86.6%	89.2%	96.32%	SWIN
ROC AUC	0.95	0.9642	0.9957	SWIN

Best Model

The **Swin Transformer** model is the best performer, excelling in recall , F1-score, and ROC AUC . Its ability to minimize misses while maintaining high precision and accuracy makes it the most reliable for detecting fake images.

INFERENCE PIPELINE OVERVIEW

Purpose

- Processes input images through selected models (CNN, ViT, or Swin Transformer) to classify images as "real" or "fake."

Steps involved :

Image Preprocessing

- Resizes images to model-compatible dimensions (224x224 pixels).
- Normalizes pixel values using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]).

Model Loading

- Loads pre-trained weights for CNN, ViT, and Swin Transformer.
- Ensures models are in evaluation mode and run on GPU/CPU for efficient processing.

INFERENCE PIPELINE OVERVIEW

Model Inference

- Preprocessed images are passed through the selected model.
- Raw outputs (likelihood vectors) are generated, representing probabilities for each class.

Post-Processing

- Applies a softmax function to generate a probability distribution.
- Selects the class with the highest probability:
  - **"Real" (1):** If the likelihood of being real is highest.
  - **"Fake" (0):** If the likelihood of being fake is highest.

DEPLOYMENT

Frontend:

- Built with React for an interactive user interface.
- Styled using Material UI for a modern and responsive layout.

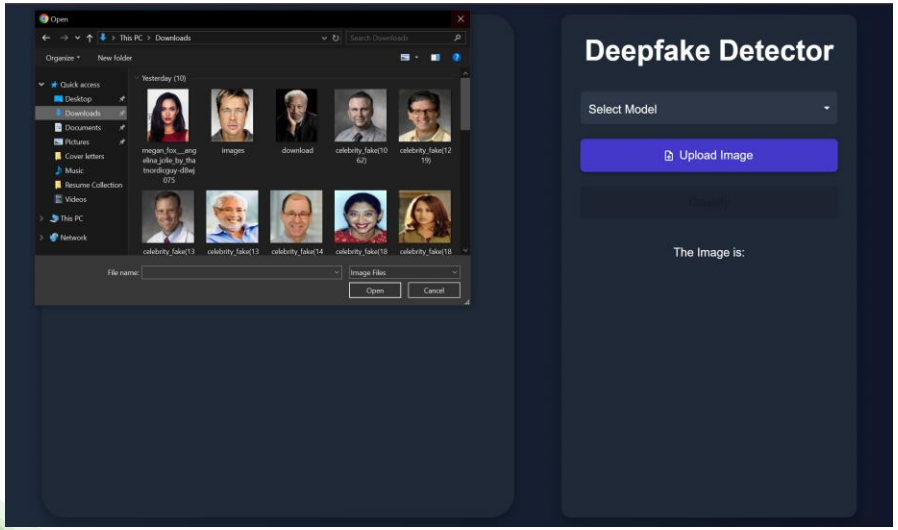
Backend:

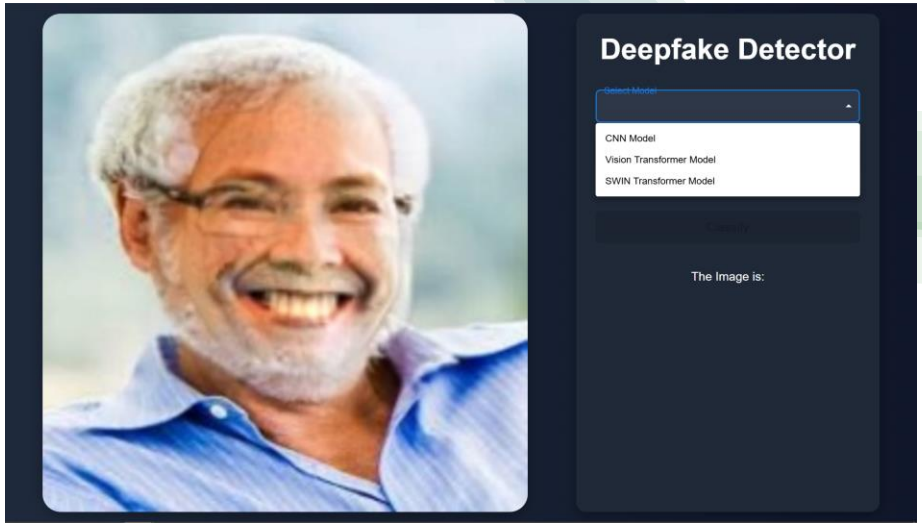
- Developed with Flask to handle image processing.
- Utilizes pre-trained models (CNN, ViT, and Swin) via REST APIs for deepfake classification.

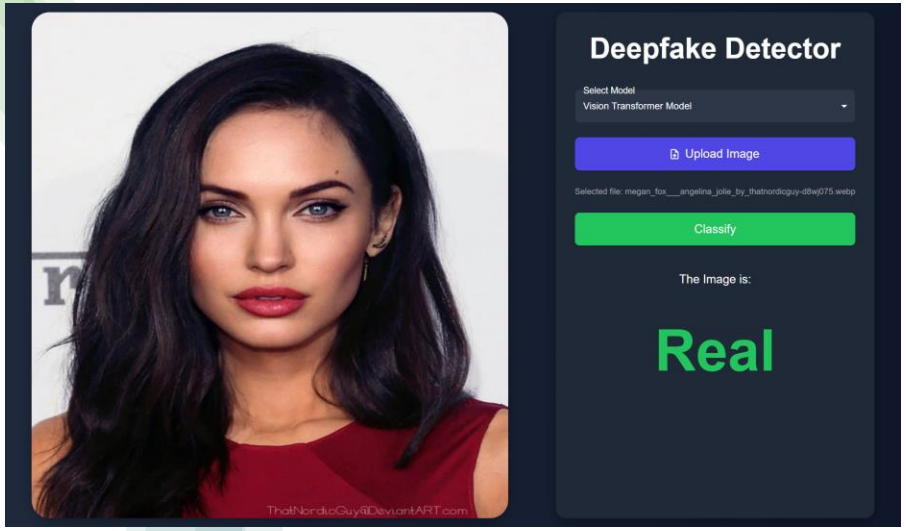
Functionality:

- Users can upload an image and select a model (CNN, ViT, or Swin).
- The app provides a "Real" or "Fake" prediction based on the selected model's output.

SNAPSHOT OF THE UI







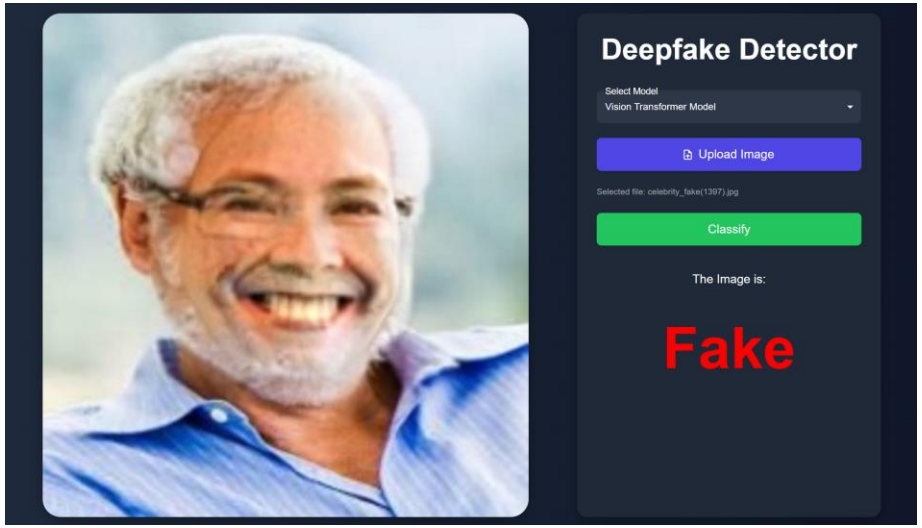


Fig 19. UI Screenshots

FUTURE SCOPE

- Could expand detection capabilities to video sequences, not just images.
- Could explore additional methods for improving detection accuracy and robustness.
- Could explore parallel GPU compute structures for training model on high resolution images.

THANK YOU