DEEP FAKE DETECTION USING VISION TRANSFORMERS

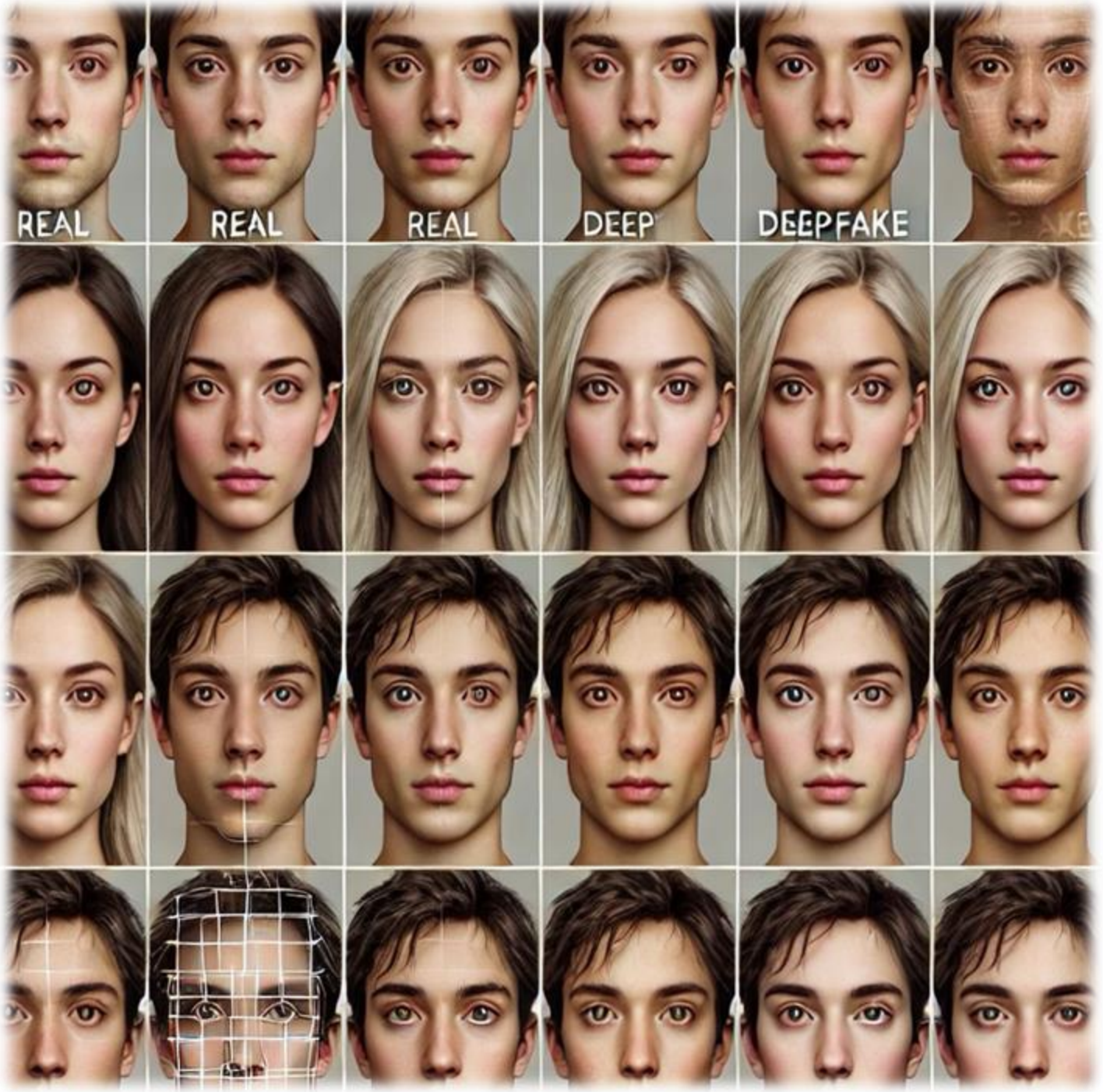
By Group 9:

Parth Malik

Shushant Ghosh

Suprajasai Konegari

Supriya Konegari



INTRODUCTION

Why is Deepfake Detection Important?

- Advanced AI enables highly convincing fake images.
- Deepfakes spread misinformation and harm reputations, especially of public figures.

Challenges with Existing Solutions

- Traditional tools miss subtle, high-quality fakes.
- Outdated methods struggle with advanced manipulations.

Necessity of Improved Solutions

- Vision Transformers provide flexibility and accuracy.
- Essential for media verification, especially during elections.

AI deepfakes a top concern for election officials with voting underway

Artificial intelligence could be weaponized on voters, feds warn

By Devin Dwyer and Sarah Herndon
October 18, 2024, 5:36 AM

AI fakes raise election risks as lawmakers and tech companies scramble to catch up

FEBRUARY 8, 2024 · 5:00 AM ET

HEARD ON MORNING EDITION

Shannon Bond

News | US Election 2024

'A lack of trust': How deepfakes and AI could rattle the US elections

At least 20 states have passed regulations against election deepfakes, but federal action remains stalled.

EMERGING TECHNOLOGIES ■ JUNE 12

Deepfake fraud directed at banks on the rise

Added friction, manual processes and better education is being used to combat the rise of digital clones from fraudsters

AI deepfakes threaten to upend global elections. No one can stop them.

As more than half the global population heads to the polls in 2024, AI-powered audio, images and videos are sowing confusion and clouding the political debate

COMMENTARY

Artificial intelligence, deepfakes, and the uncertain future of truth

John Villasenor
February 14, 2019

As social media guardrails fade and AI deepfakes go mainstream, experts warn of impact on elections

Politics Dec 27, 2023 5:27 PM EDT

AI-generated deepfakes are moving fast. Policymakers can't keep up

UPDATED APRIL 27, 2023 · 6:11 PM ET

HEARD ON MORNING EDITION

Shannon Bond

DATASET

FaceForensics++ Dataset
A large-scale collection of real and manipulated facial images using deepfake techniques like FaceSwap, Face2Face, and DeepFakes.

- **Samples:** Over 1,000 video sequences and 13,000 labeled image frames for supervised learning.
- **Techniques:** Includes various methods, such as facial reenactment and face-swapping.

GAN-Generated Images
Combined with GAN-generated images to enhance dataset complexity.

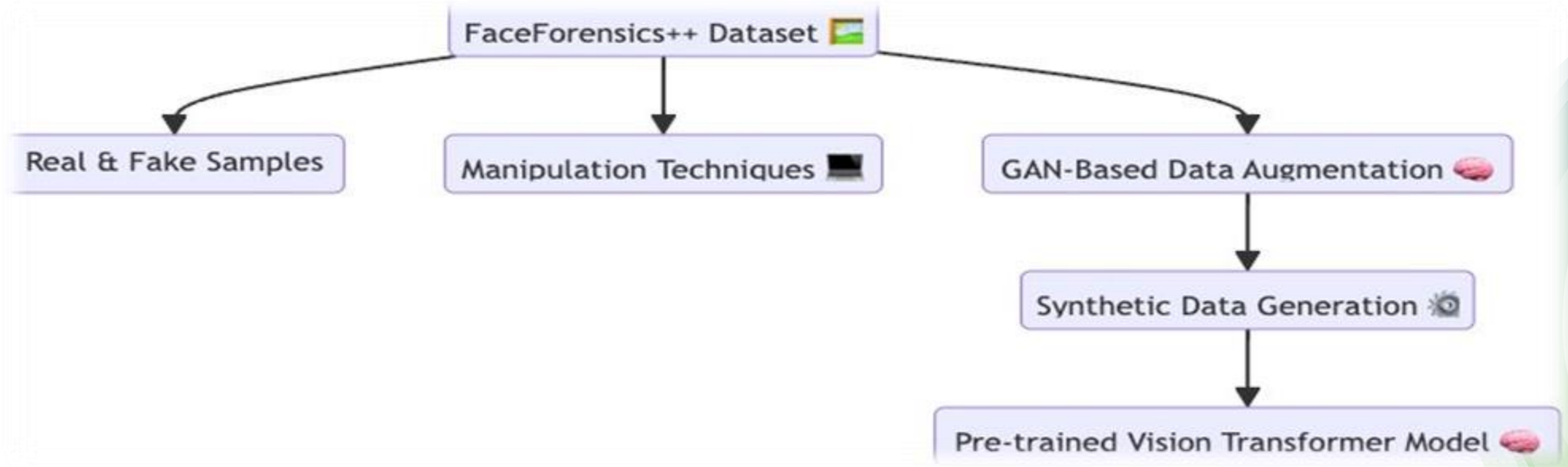


Figure 1: Dataset

PRE-PROCESSING STEPS

Balancing the Dataset

In deepfake detection, a balanced dataset of paired real and fake images is essential for model accuracy. Initially, some real images lacked fake counterparts, risking model bias.

To resolve this, we removed real images without fake pairs and generated 1.6k deepfakes using celebrity images from the UTK Face Dataset (Figure 3).

These GAN-generated images, created with the InsightFace library, were added to the original dataset for training.

Figure 2: Generated Deepfakes of Celebrity Dataset

PRE-PROCESSING STEPS

Image Generation using GAN & InsightFace

InsightFace was used for face swapping. It's key components are :

- Face recognition built on ResNet to extract facial features.
- Landmark Detection network like FAN for identifying the key points (eye, mount, nose).

GANs were employed to create high-quality synthetic images. The GAN architecture included:

- A Generator: uses CNN to transform input image to a more realistic and high quality synthetic image.
- Discriminator: also uses CNN to detect if the generated image is real or fake.

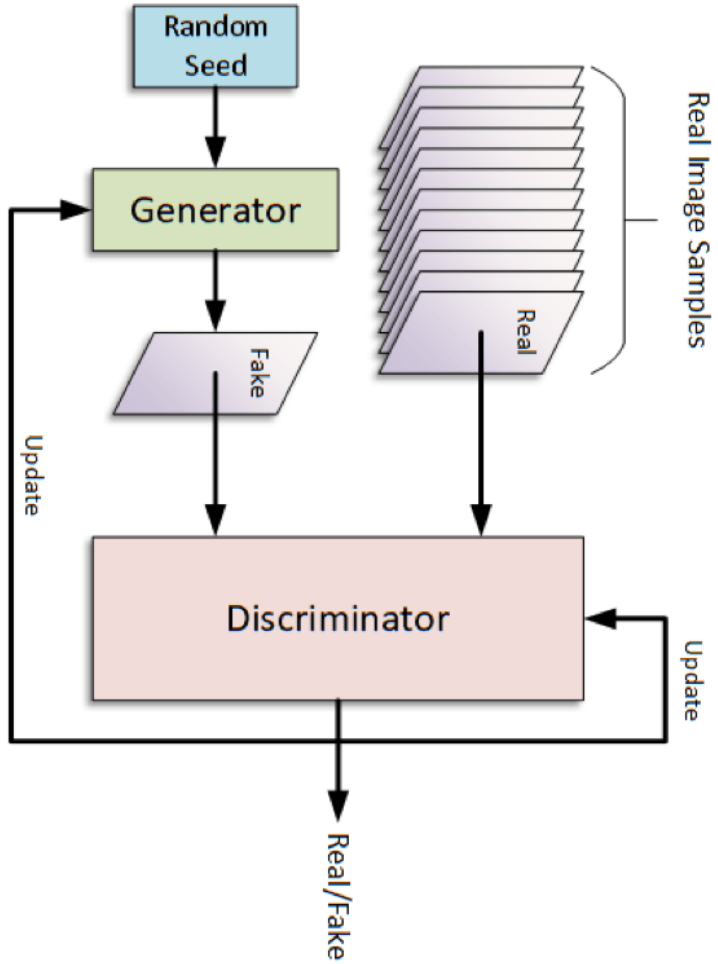


Figure 3: GAN Architecture

PRE-PROCESSING STEPS

Image Generation Results

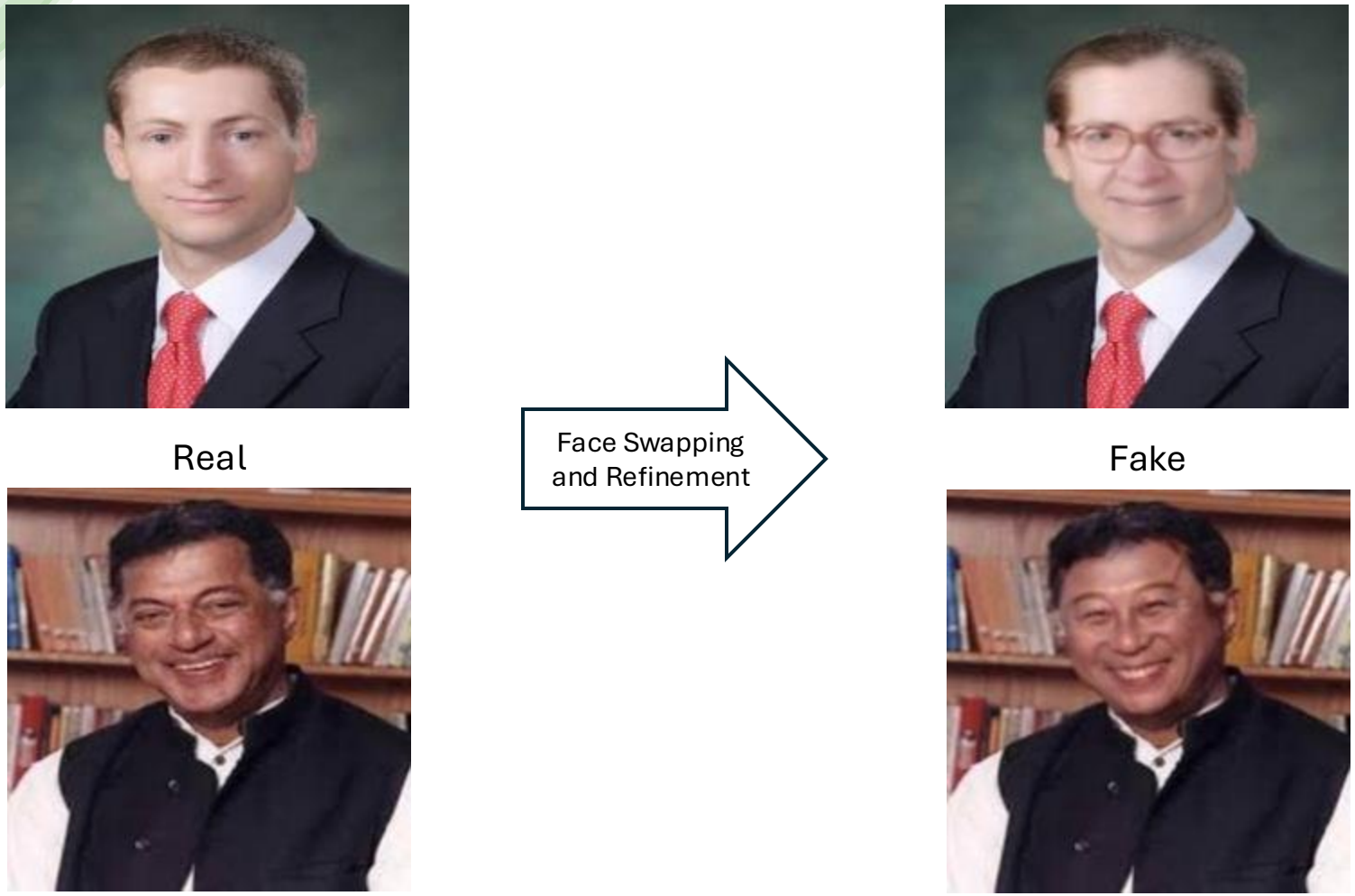


Figure 4: Results of image generation using GANS

MODELING

CNN Architecture

- **Conv Layers:** 3 convolutional layers with filters (32, 64, 128) using ReLU activation for feature extraction.
- **Pooling Layers:** Max-pooling after each convolution to reduce spatial dimensions.
- **Fully Connected Layers:**
  - **FC1:** Reduces to 512 units
  - **FC2:** Reduces to 128 units
  - **FC3:** Outputs 2 units for "real" or "fake" classification
- **Dropout Layer:** Applied in FC layers to prevent overfitting.
- **Output Layer:** Softmax activation for binary classification

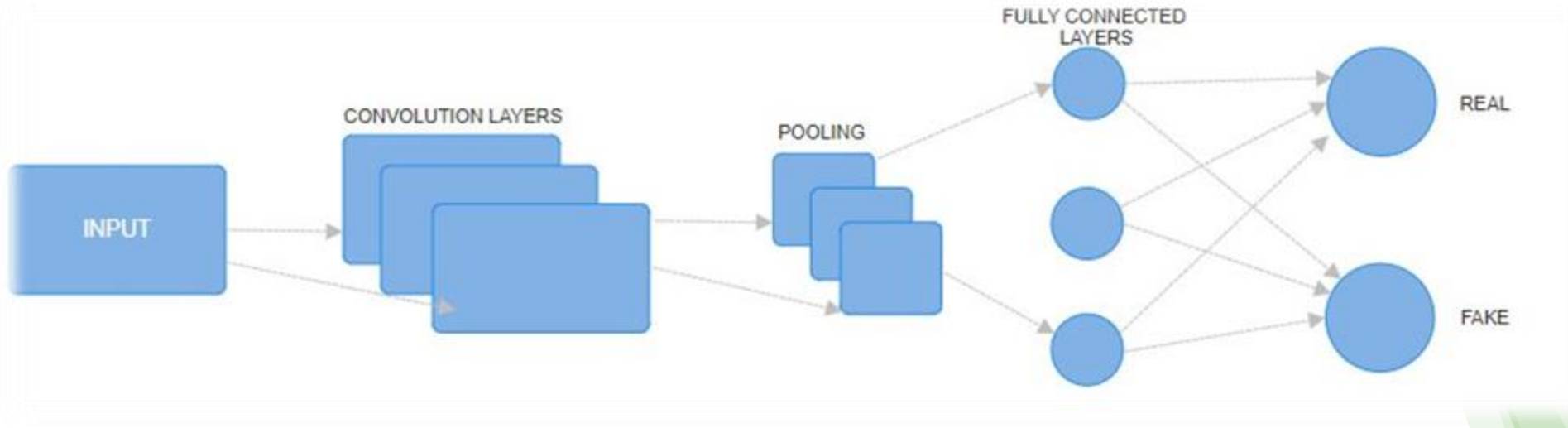


Figure 5: CNN Architecture

CNN RESULTS : POST-GAN GENERATED DATA

Precision: 0.8722

Recall: 0.8609

F1-Score: 0.8665

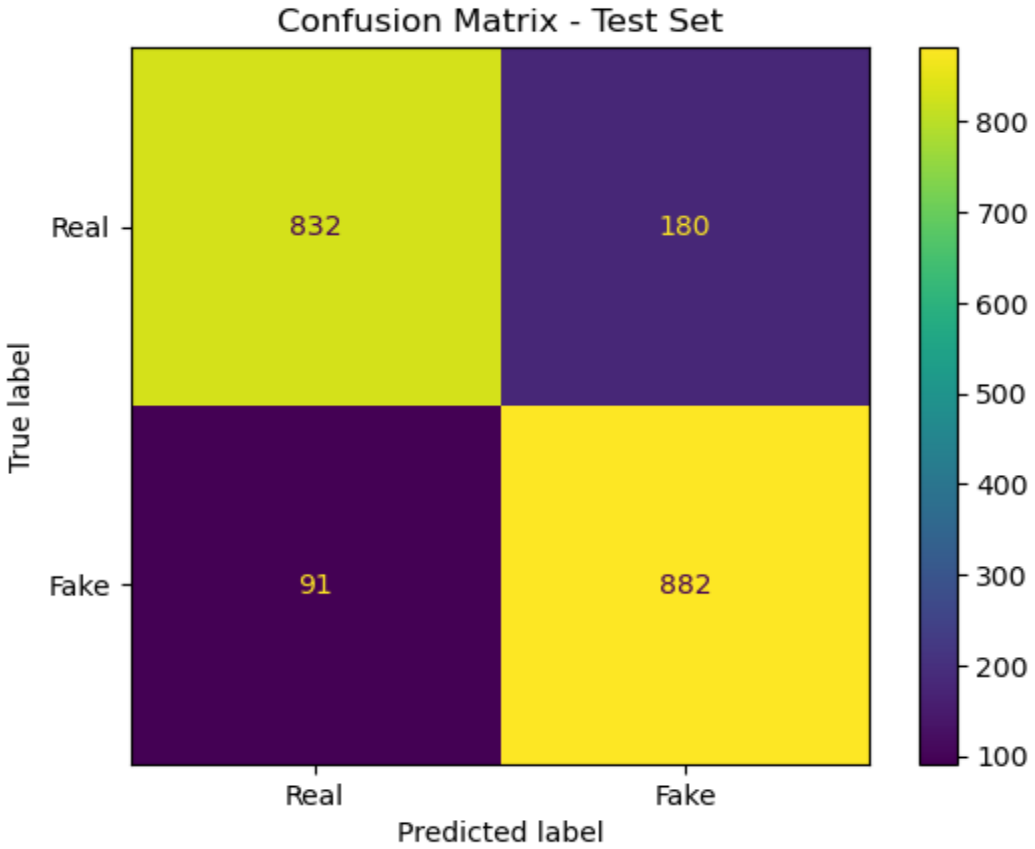


Figure 6 : Confusion Matrix

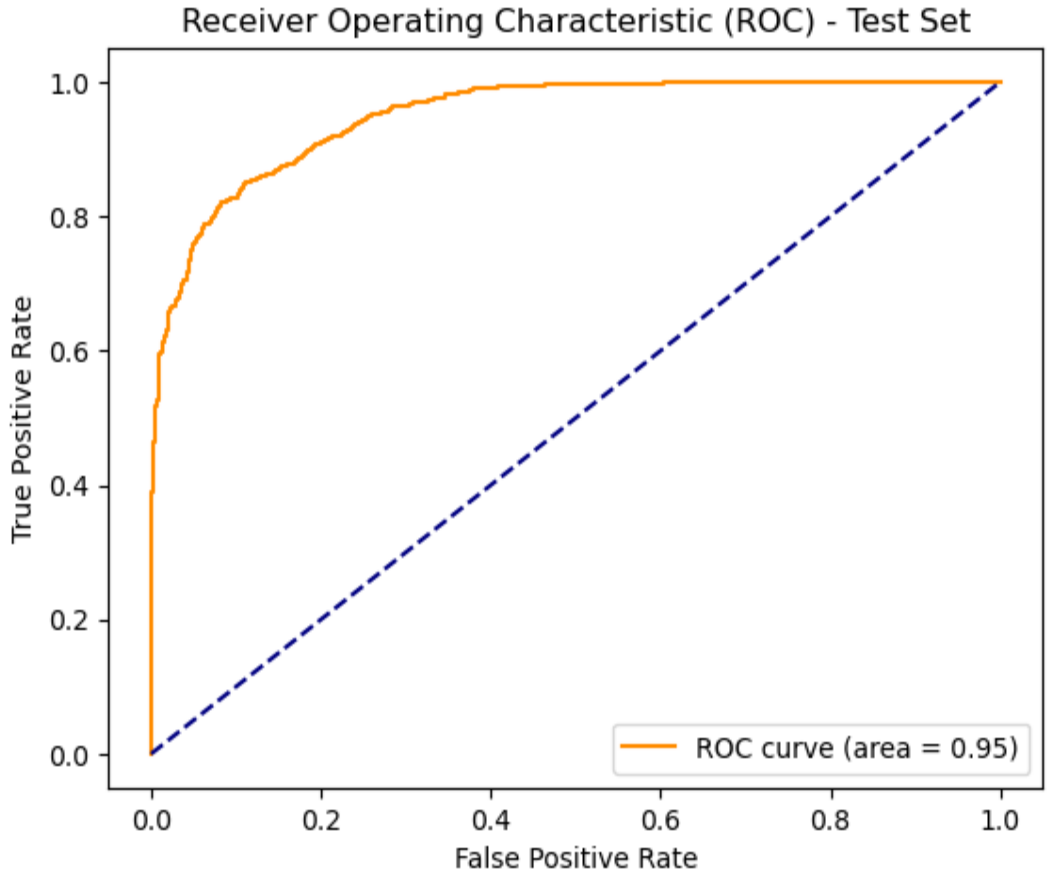


Figure 7 : ROC-AUC Curve

ViT ARCHITECTURE

Key Modifications

• **Transfer Learning Applied**: Fine-tuned a pre-trained ViT model.

• **Binary Classifier Adaptation**: Output layer changed from **1000 classes to 2 classes (Real/Fake).**

• **Partial Layer Unfreezing**:
- **First 6 encoder layers (0-5)**: Kept frozen to retain general features from the original model.

- **Last 6 encoder layers (6-11)**: Unfrozen to learn specific features from deepfake images.

• **Balanced Fine-Tuning**: Combines general knowledge with task-specific adaptation to improve accuracy and prevent overfitting.

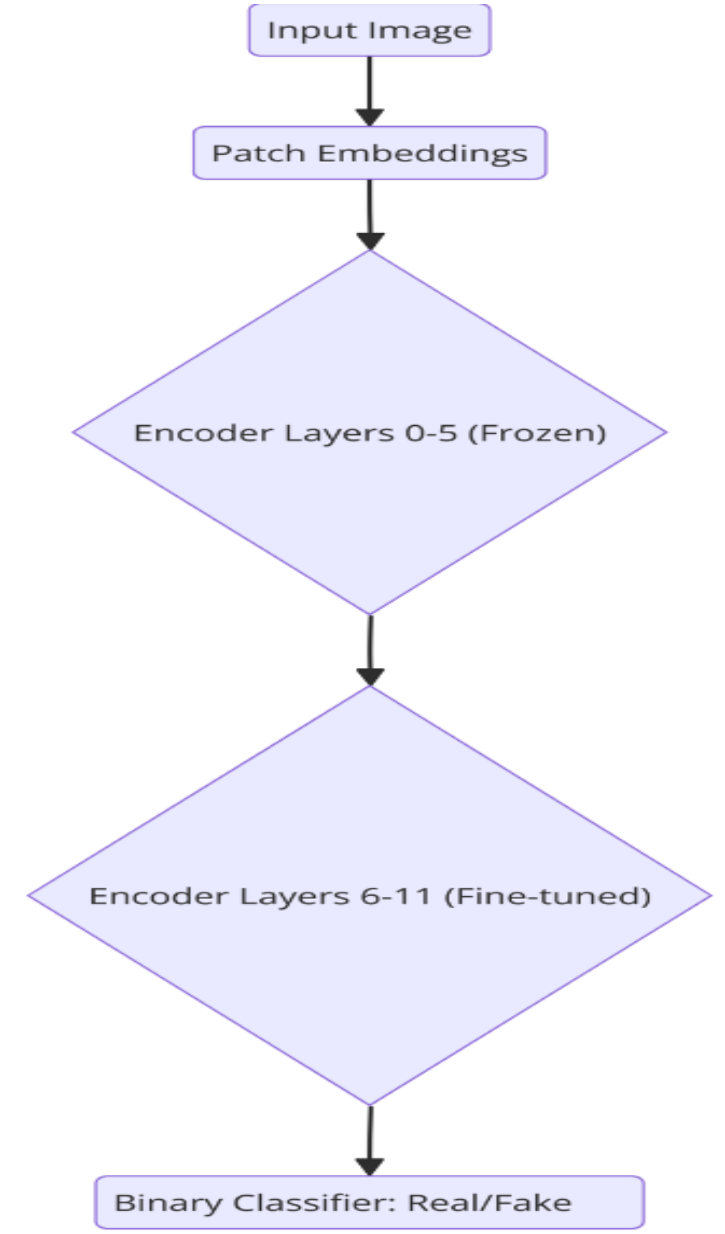


Figure 8: ViT Architecture

DATA AUGMENTATION FOR ViT TRAINING

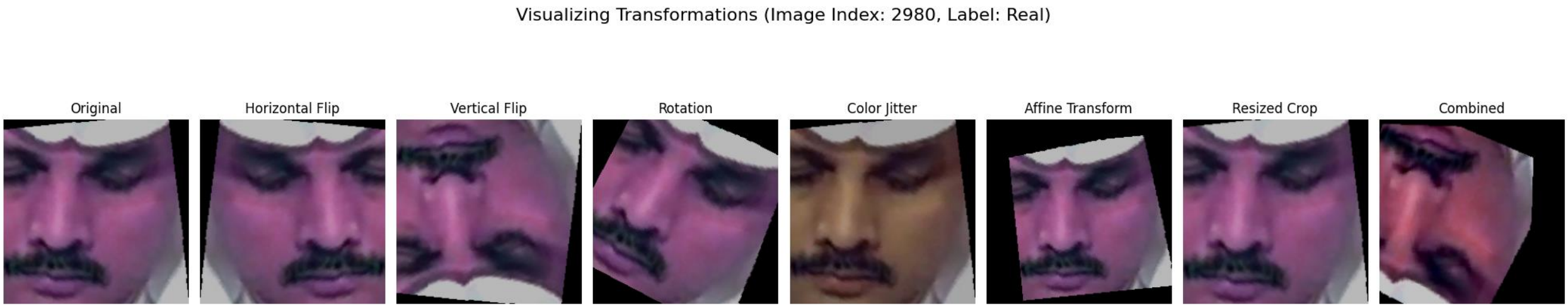


Figure 9 : Various steps in Data Augmentation

Final Output of a Transformation on an Input Image

- Introduces transformations like flipping, rotation, color jitter, affine transform, and resized crop to simulate real-world variations.

- Enhances model robustness, prevents overfitting, and improves generalization.

ViT RESULTS : ORIGINAL DATASET

Initial Results (Pre-GAN):

- Accuracy: 78.9%, Precision: 77.4%, Recall: 81.7%
- F1-Score: 79.5%, ROC-AUC: 86.3%
- Test Loss: 0.4893

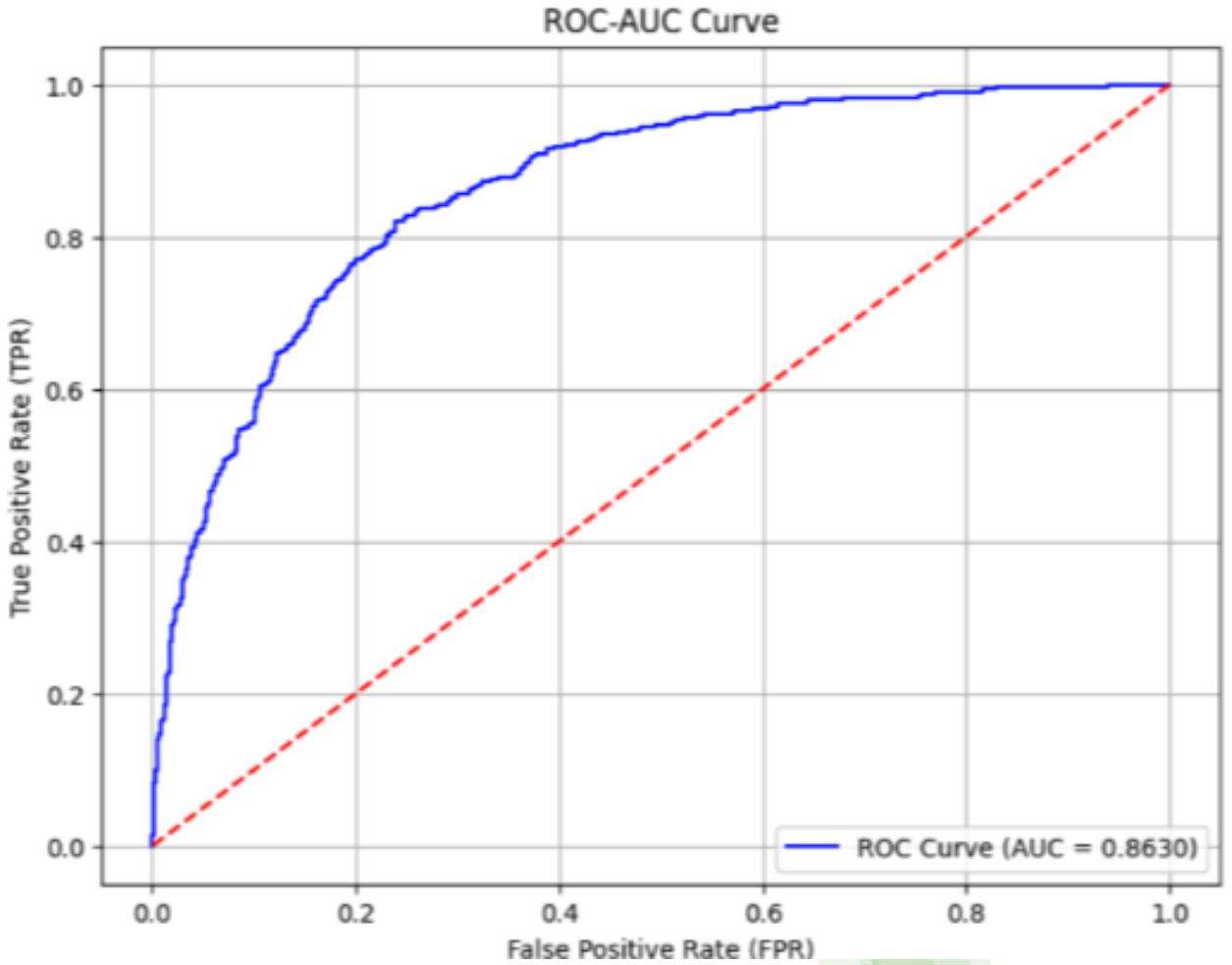


Figure 10 : ROC-AUC Curve

ViT RESULTS : POST-GAN GENERATED DATA

- Accuracy improved to 88.8% (+10.1%), Precision: 86.4% (+8.7%), **Recall: 92.1%** (+11.5%)

- F1-Score: 89.2% **(+10.1%)**, ROC-AUC: 96.4% **(+8.7%)**

- Test Loss reduced to 0.2491 **(-49.1%)**

Insight: Combining GAN data and introducing data augmentations significantly boosted recall and AUC, improving robustness.

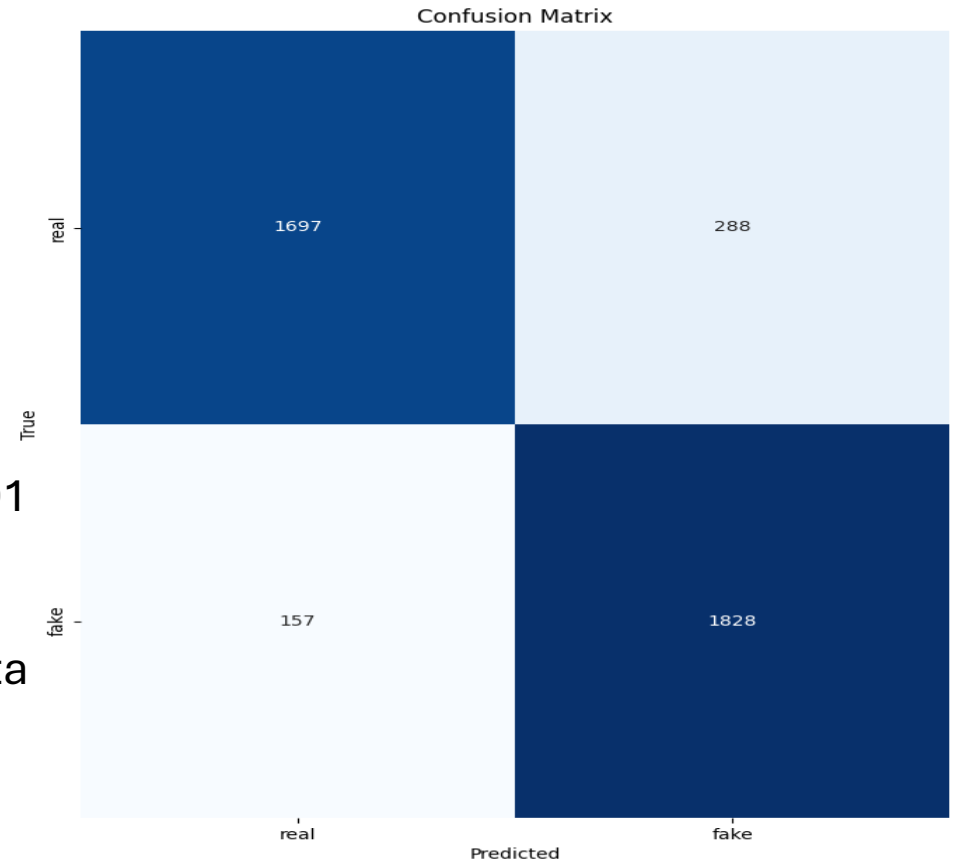


Figure 11: Confusion Matrix

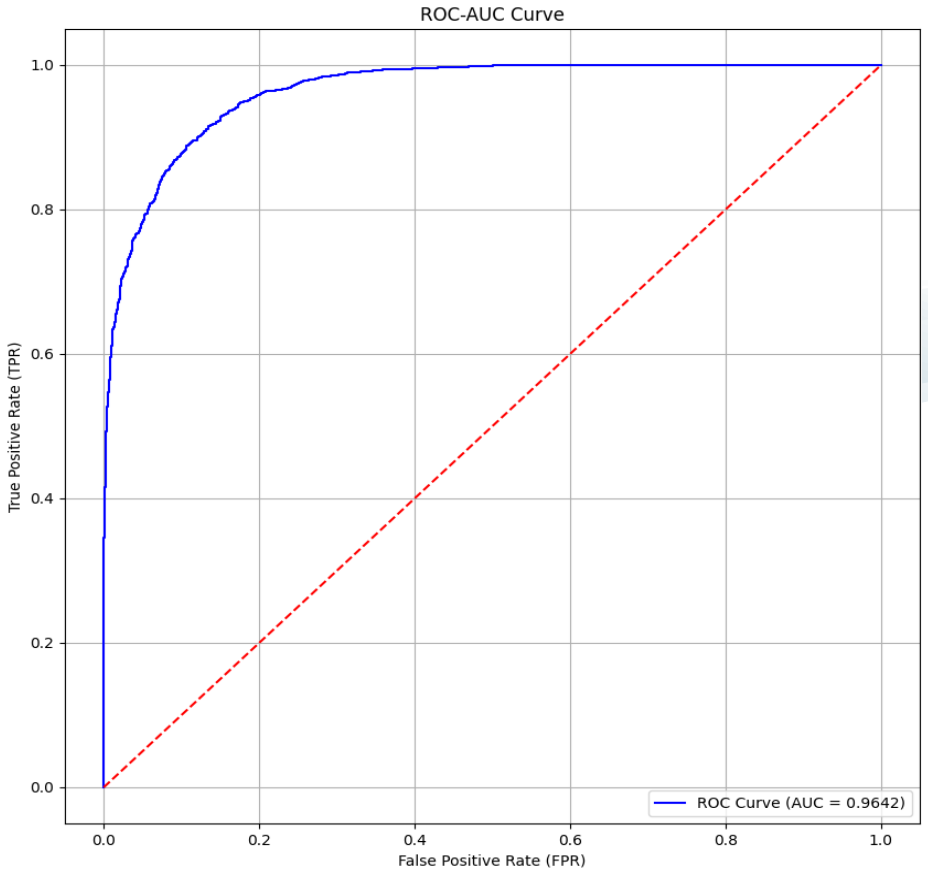


Figure 12: ROC-AUC Curve

METRICS – PRIORITY: RECALL

- **Accuracy**: Proportion of correct predictions for real and fake images but may be misleading in imbalanced datasets.
- **Precision**: Percentage of predicted fakes that are truly fake, reducing false positives.
- **Recall**: Proportion of actual fakes correctly identified; critical to minimize false negatives.
- **F1-Score**: Balances precision and recall, offering a combined view of detection accuracy.

TRUE POSITIVES Fake detected as fake – good! **High True Positives (TP)** Means Our system is correctly identifying most deepfakes, making it effective at catching fake content.	**FALSE POSITIVES** Real detected as fake – bad. **High False Positives (FP)** Leads to unnecessary flagging of real content, which could cause trust issues in real-world applications.
FALSE NEGATIVES Fake detected as real – very bad! **High False Negatives (FN)** The most concerning, as it means deepfakes are slipping through undetected, defeating the purpose of the detection system.	**TRUE NEGATIVES** Real detected as real – good! **High True Negatives (TN)** Shows that the system can trust real content and won't over-classify it as manipulated.

CNN MODEL PREDICTIONS

False Positives: Real images detected as fake

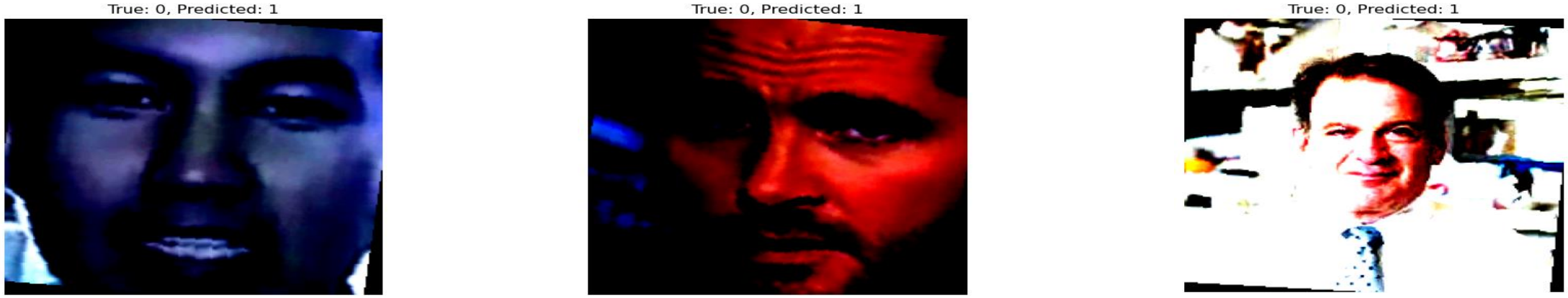


Figure 13 : False Positives

False Negatives: Fake images detected as real

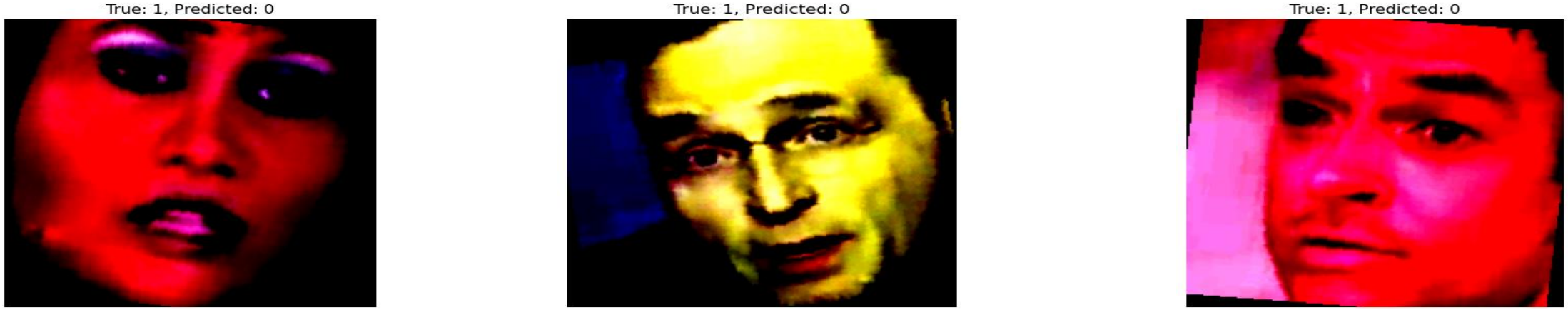


Figure 14 : False Negatives

(1 = Fake , 0 = Real)

CNN MODEL PREDICTIONS

Correctly Classified Samples

Figure 15 : Correctly classified Samples

(1 = Fake , 0 = Real)

DISCUSSION

Strengths and Limitations

CNN Model

Strengths:

- **Precision (87.2%)**: Reliable in identifying fakes with fewer false positives.
- **ROC AUC (0.95)**: Strong at distinguishing real vs. fake.

Limitations:

- **Recall (86.1%)**: Lower recall than ViT, missing some fake images.
- **False Positives**: 180 real images misclassified as fake, potentially impacting trust.

ViT Model

Strengths:

- **Recall (92.1%)**: Effectively captures fake images.
- **F1-Score (89.2%)**: Balanced precision and recall.
- **ROC AUC (0.9642)**: Strong at distinguishing real vs. fake.

Limitations:

- **Precision (86.4%)**: Slightly lower than CNN, with more false positives.
- **False Positives**: 288 real images misclassified as fake, showing some limitations.

DISCUSSION

Model Comparison

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

Best Model

ViT, due to higher recall, F1-score, and ROC AUC, making it more reliable for detecting fakes with minimal misses.

THANK YOU