REPORT
Image Based Deepfake Detection System

Github link

Parth Malik, Shushant Ghosh, Suprajasai Konegari, Supriya Konegari

1. Problem Statement

As artificial intelligence technology advances, the creation of highly realistic manipulated media, commonly referred to as "deepfakes," has become increasingly prevalent. This technology, powered by machine learning models capable of generating hyper-realistic images and videos, enables the production of convincingly false media that can spread misinformation, damage reputations, and compromise privacy. In particular, the potential for deepfakes to target public figures and influencers heightens concerns around privacy, security, and information integrity. Given the sophisticated nature of today's manipulation techniques, traditional detection methods are often inadequate, struggling to identify subtle, high-quality fakes and keep pace with evolving generation methods.

This challenge is especially critical in politically sensitive contexts, such as elections, where deepfakes could manipulate public opinion, misrepresent candidates, and disrupt the democratic process. To counter these threats, our study investigates advanced deepfake detection techniques using Vision Transformers (ViTs), leveraging their self-attention mechanisms for fine-grained image analysis. We compare these results with Convolutional Neural Networks (CNNs), assessing model performance across an expanded dataset, including images generated by Generative Adversarial Networks (GANs). Our approach aims to enhance media verification processes by improving the detection of even the most intricate manipulations, supporting the integrity of information and reinforcing public trust in media content.

2. Methods

2.1. Dataset

The **FaceForensics++ dataset**, available on Kaggle, is a large-scale collection specifically curated for training and evaluating deepfake detection models. It includes real and manipulated facial images created using popular deepfake techniques, such as FaceSwap, Face2Face, and DeepFakes, which employ advanced facial manipulation methods like facial reenactment and face-swapping. This dataset comprises over 1,000 video sequences and around 13,000 labeled image frames, offering extensive data for supervised learning and model training.

To further enhance dataset complexity and provide a more challenging array of manipulated images, an additional 1,600 deepfake samples were created using celebrity images sourced from the **UTK Face Dataset**. These samples, generated using GANs, add a layer of diversity that simulates real-world deepfake scenarios more effectively. By combining GAN-generated images with traditional facial manipulation techniques, the dataset enables model training across broader spectrum of fake image types, ultimately boosting robustness and detection accuracy for deepfake detection. Figure 1 represents the dataset overview. In this report,

- **Original Dataset** refers to the data sourced exclusively from the FaceForensics++ dataset, consisting of real and manipulated images used for initial model training and evaluation.
- **Post-GAN Data** refers to the expanded dataset, which includes the original FaceForensics++ data supplemented with synthetic images generated using GANs.

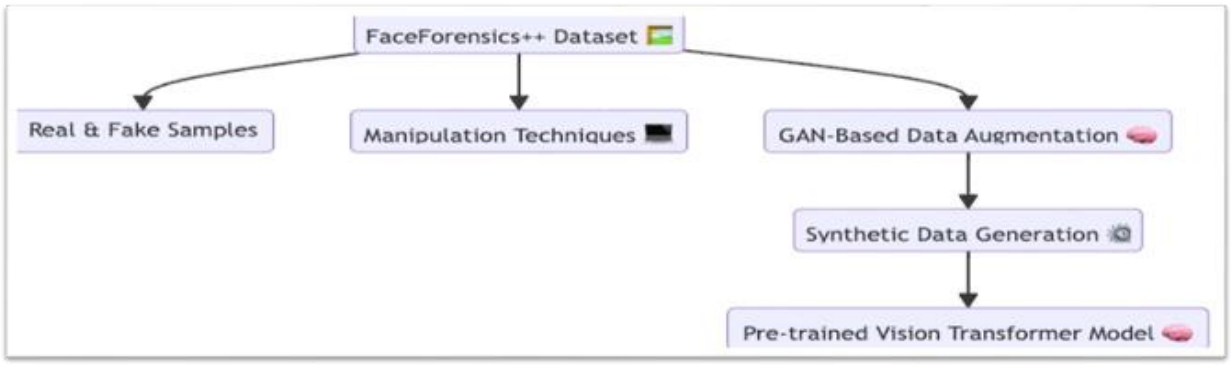

Fig 1. Dataset

2.2. Data Pre-processing

One of the primary challenges with the FaceForensics++ Dataset is the lack of consistent pairing between real and manipulated images. In some cases, certain real images do not have corresponding fake counterparts, leading to an imbalance in the dataset. This inconsistency can hinder the model's ability to effectively learn the distinguishing features between real and fake images. When a model is trained on an imbalanced dataset, it may become biased toward the more prevalent class, resulting in higher false negatives or false positives. As a result, the model's overall accuracy and reliability in detecting deepfakes can be compromised, making it crucial to address this imbalance during the training process.

Deepfakes are synthetic media created by swapping or altering faces in images. We have used **InsightFace** library which plays a crucial role in detecting and aligning facial features, while **Generative Adversarial Networks (GANs)** enhance the realism of these altered faces.

2.2.1. InsightFace Architecture

InsightFace is an open-source deep learning library used for face recognition and alignment. Its key components include:

- **ArcFace Model (Face Recognition)**:
    a. Built on **ResNet.** ArcFace extracts facial features and generates embeddings that help map facial characteristics.
    b. The **Additive Angular Margin Loss (ArcFace loss)** ensures high accuracy in recognizing and comparing faces, which is useful for matching the source and target faces in deepfakes.
- **Landmark Detection & Alignment**:
    a. InsightFace uses a landmark detection network, like **FAN**, to locate facial key points (eyes, nose, mouth), aligning the source face with the target for smooth integration.
- **Face Segmentation**:
    a. Using **Fully Convolutional Networks (FCNs)**, InsightFace can segment key facial regions (eyes, nose, mouth) for better blending during face swapping.

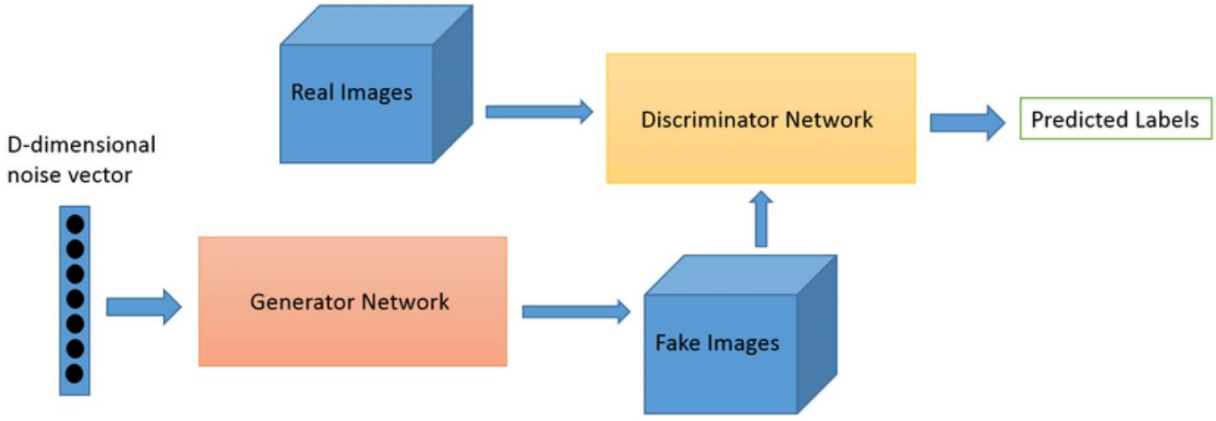

Fig 2. GAN Architecture

2.2.2. GAN Architecture

Generative Adversarial Networks (GANs) consist of two neural networks that compete with each other: the generator and the discriminator (Figure 2 shows the GAN architecture):

- **Generator:**
    a. Produces synthetic images, learning to mimic real data.
    b. It uses **Convolutional Neural Networks (CNNs)** with **transposed convolutional layers** (also known as deconvolution layers) to gradually up-sample the input and generate high-resolution images.
- **Discriminator**:
    a. Distinguishes between real and fake images, pushing the generator to improve.
    b. It uses **convolutional layers** to down-sample the input, extracting important features like edges, textures, and lighting patterns that help differentiate between authentic and generated images.

GAN-based architecture transfers realistic facial expressions from the source to the target face and ensuring that the swapped face blends naturally into the target scene.
By combining **InsightFace** for face detection and alignment with **GANs** for facial refinement, deepfakes achieve a higher level of realism. Incorporating such deepfakes into the training set would increase the robustness of the model and help it in identifying realistic deepfakes.

Figure 3 shows the result of generating deepfakes using the above methodology.


Fig 3. Image swapping and GAN based realistic image generation.

2.2.3. ViT Data Augmentation

To enhance the robustness and generalization capabilities of the ViT model, a comprehensive data augmentation strategy was implemented during training. This augmentation pipeline incorporated various transformations, including horizontal and vertical flips, rotations, color jittering, affine transformations, and random resized crops, all aimed at simulating a wide range of real-world scenarios. These techniques were specifically designed to improve the model's adaptability to diverse input patterns while mitigating the risk of overfitting. Adjustments included increasing rotation angles up to 30 degrees, intensifying color variations, and introducing vertical flips with a probability of 0.2. Figure 4 shows an image at each step of the data augmentation pipeline.

Furthermore, affine transformations and random resized cropping were fine-tuned to provide greater variability in object sizes and positions. By enhancing the diversity of the training dataset, these transformations enabled the model to capture a broader range of image distortions and inconsistencies. As a result, the ViT model became more resilient and effective in generalizing to unseen data, ultimately improving its performance in distinguishing between real and fake images.

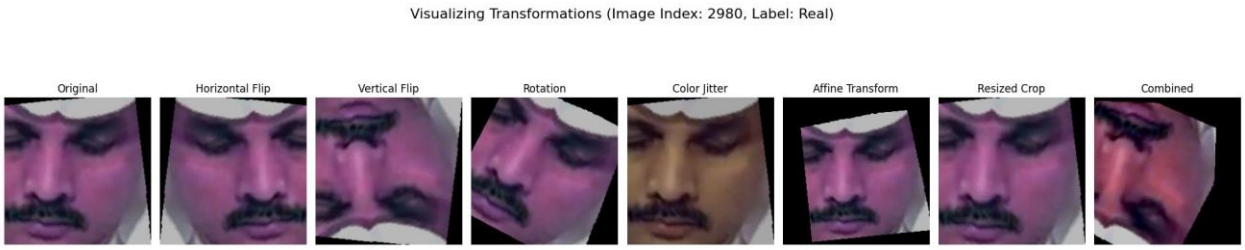

Fig 4. ViT Data Augmentation

2.3. Modeling

We opted for Convolutional Neural Networks (CNNs) due to their proven effectiveness in image classification tasks, particularly in extracting spatial features and identifying intricate patterns in visual data. On the other hand, we selected Vision Transformers (ViTs) for their ability to capture long-range dependencies and contextual relationships by processing images as sequences of patches.

2.3.1. CNN Architecture

The Convolutional Neural Network (CNN) architecture consists of several layers designed to effectively extract features from input images and classify them as "real" or "fake."

- **Convolutional Layers:** The architecture includes three convolutional layers, utilizing filters of sizes 32, 64, and 128, respectively. Each layer employs ReLU (Rectified Linear Unit) activation functions, which enhance non-linearity and facilitate efficient feature extraction from the input data.
- **Pooling Layers:** Following each convolutional layer, a max-pooling layer is implemented to reduce the spatial dimensions of the feature maps. This downsampling process helps to minimize computational complexity and retain the most significant features.
- **Fully Connected Layers:** The network comprises three fully connected (FC) layers. The first FC layer reduces the dimensionality to 512 units, followed by a second FC layer that further reduces it to 128 units. The final FC layer outputs 2 units, corresponding to the binary classification of images as either "real" or "fake."
- **Dropout Layer:** To mitigate the risk of overfitting, dropout is applied to the fully connected layers. This technique randomly sets a fraction of the input units to zero during training, promoting generalization of the model.
- **Output Layer:** The output layer utilizes a softmax activation function, which allows for the probability distribution of the two classes—"real" and "fake"—enabling effective binary classification.

Figure 5 illustrates the CNN architecture, detailing the arrangement of convolutional and fully connected layers, providing a clear visual representation of the model's structure and functionality.

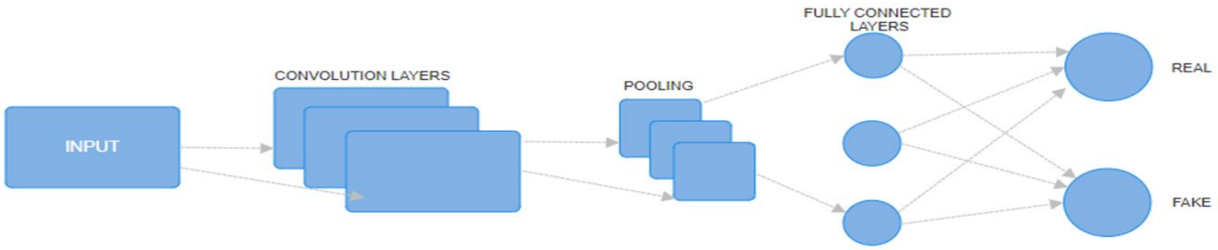

Fig 5. CNN Architecture

2.3.2. ViT Training

To effectively utilize the Vision Transformer (ViT) architecture for deepfake detection, we employed a partial fine-tuning strategy through transfer learning. Figure 6 illustrates this process. Here are the key points outlining our approach:

- **Utilization of Pretrained Model**: We started with a ViT model pretrained on large-scale datasets, leveraging its existing knowledge.
- **Output Layer Modification**: The output layer was adjusted from handling 1,000 classes to just 2 classes—real and fake—suitable for our binary classification task.
- **Initial Freezing of Layers**: Most of the model's layers were initially frozen, allowing only the classifier to be trained. This method is beneficial when working with limited data, as it prevents overfitting.
- **Selective Unfreezing of Encoder Layers**: To enhance the model's ability to learn specific features from deepfake images, we selectively unfroze the last six encoder layers (layers 6-11). This adjustment allows these layers to update their weights based on our dataset.
- **Preserving General Features**: The first six encoder layers (0-5) remained frozen to maintain their capability to extract general, low-level features applicable across various tasks.
- **Hierarchical Learning Strategy**: This hierarchical unfreezing approach enables the model to retain broad patterns from the initial layers while adapting to deepfake-specific features in the deeper layers.

.

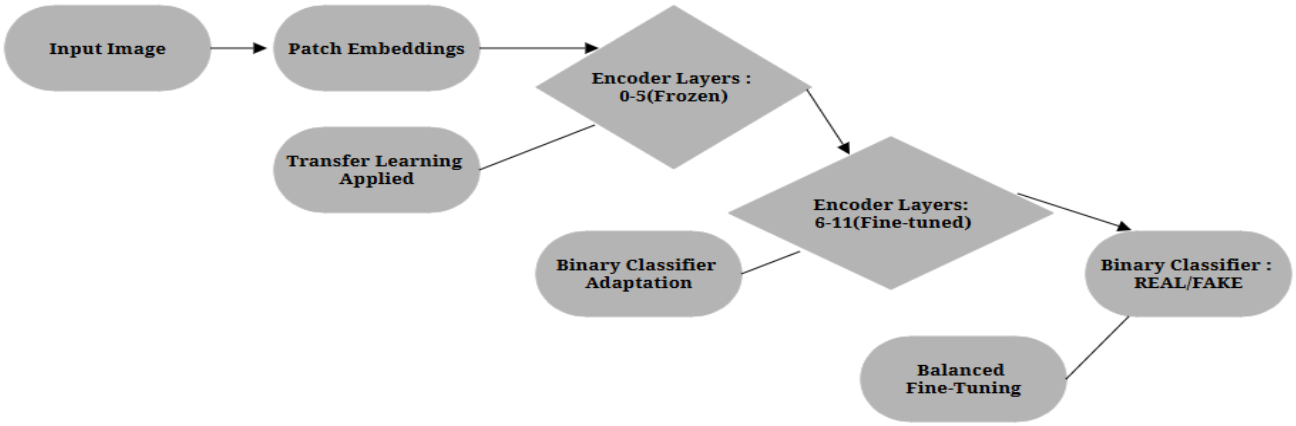

Fig 6. ViT Training

3. Results

In our project, we focused on evaluating the performance of our deepfake detection system using critical metrics such as precision, recall, and F1-score.

3.1 CNN

3.1.1 CNN Results: Post-GAN Generated Data

- **Precision**: The precision of **CNN model is 0.8722**, indicating that when the model predicts an image as "fake," it is correct approximately 87.22% of the time. This high precision suggests that the model effectively identifies genuine deepfake images without misclassifying too many real images as fake.
- **Recall**: The **recall of the model stands at 0.8609**, meaning that it successfully identifies about 86.09% of all actual deepfake images. This metric highlights the model's capability to capture the majority of true positives, which is crucial in minimizing the risk of false negatives—failing to detect a manipulated image.
- **F1-Score**: The **F1-score, calculated at 0.8665**, represents the harmonic mean of precision and recall. This score balances both metrics, providing a single performance measure that reflects the model's overall effectiveness in detecting deepfakes.
- **Confusion Matrix**: Figure 7 presents the confusion matrix, which visually represents the model's classification performance. In this matrix, the **true positives (TP) for "real" images are 832**, while there are **180 false negatives (FN)**, indicating the number of real images incorrectly classified as fake. Conversely, **the true positives for "fake" images are 882**, with **91 false positives (FP)**, representing the number of fake images misclassified as real. The confusion matrix illustrates a strong ability of the model to distinguish between real and fake images, with a notable balance between the two classes.
- **ROC Curve**: Figure 7 showcases the Receiver Operating Characteristic (ROC) curve. The area under the **ROC curve (AUC) is 0.95**, indicating excellent model performance. An AUC value close to 1 suggests that the model is highly effective in distinguishing between real and fake images, reinforcing the robustness of our deepfake detection approach.

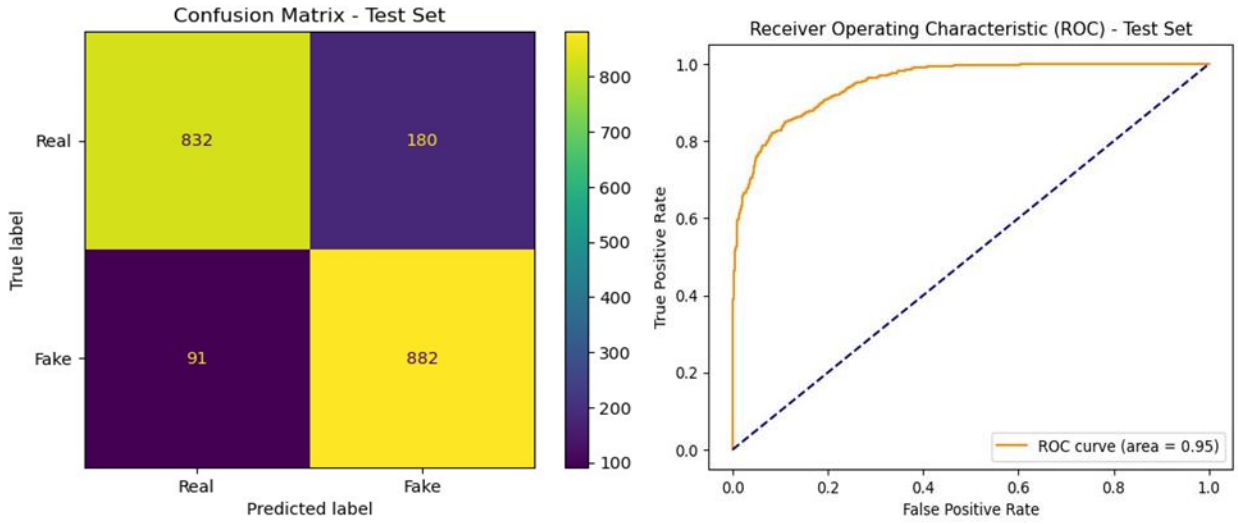

Fig 7. Confusion Matrix and ROC-AUC curve of CNN Model

3.1.2 CNN Model Predictions (After Data Augmentation) (1 = Fake, 0 = Real)

Fig 8. False Positives: Real images detected as fake

Figure 8 represents false positives indicating real images that were detected as fake by the CNN model. The first image's low quality and bluish tint, the second image's high contrast and deep shadows, and the third image's washed-out appearance can all create artificial looks that lead the model to misclassify them as manipulated.

Fig 9. False Negatives: Fake images detected as real

As seen in Figure 9, the first image’s intense red overlay, the second image’s unnatural yellow hue, and the third image’s strong red tint can all present artificial qualities that cause the model to misclassify them as real. These misclassifications suggest that the model struggles to detect certain color abnormalities, lighting inconsistencies, and exaggerated features in synthetic images, highlighting areas for improvement.

Fig 10. Correctly classified samples

From Figure 10, representing the correctly classified samples, the true positives (TP: 1) include images with visual artifacts such as unnatural lighting, high contrast, and color distortions, characteristics commonly found in manipulated images. The model successfully identified these as fake, suggesting it can recognize synthetic patterns that deviate from natural facial appearances. Conversely, the true negatives (TN: 0) show real images where the model accurately detected genuine facial features without being misled by lighting variations or shadows.

3.2 ViT

3.2.1. ViT Results: Original Dataset

The initial evaluation of the Vision Transformer (ViT) model on the original dataset (without any GAN-augmented data) showed moderate performance across key metrics. The model achieved an **accuracy of 78.9%**, a **precision of 77.4%**, and a **recall of 81.7%**, resulting in an **F1-Score of 79.5%**. This indicates a reasonably balanced performance between precision and recall.

The **ROC-AUC score of 86.3%** as shown in Figure 11, suggests that the model has a fair ability to differentiate between real and fake images, though there is room for improvement. Notably, the **test loss was recorded at 0.4893**, hinting at potential areas for refinement, particularly in enhancing recall. In deepfake detection, minimizing false negatives is crucial to avoid failing to identify manipulated images, which could have serious implications. Overall, the results indicate that while the model shows effectiveness in distinguishing classes, there remains a clear opportunity to enhance recall and reduce test loss for greater accuracy in deepfake detection.

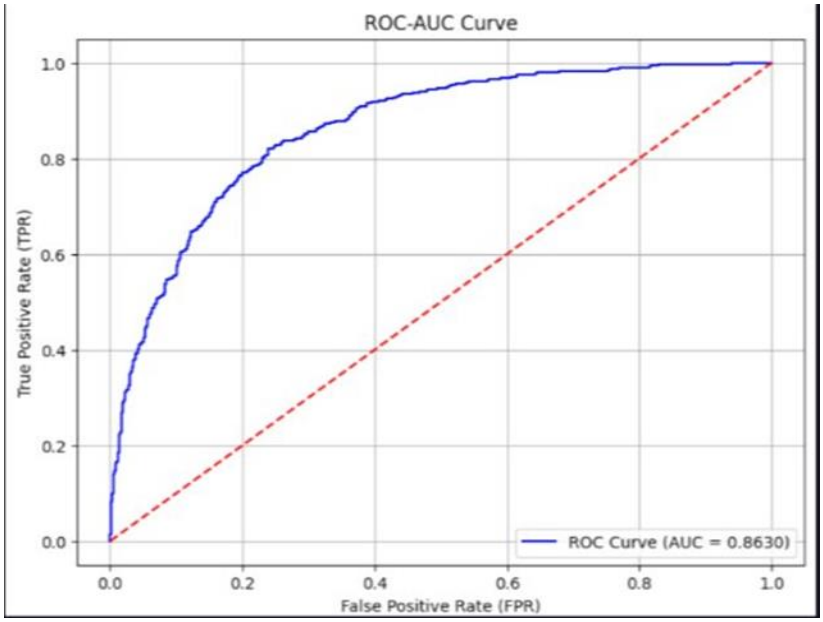

Fig 11. ROC-AUC Curve of ViT on original dataset

3.2.2. ViT Results: Post-GAN Generated Data

The introduction of synthetic data generated by GANs led to significant improvements in the model's performance. The model's **accuracy rose to 88.8%, a notable increase of 10.1%**.

- **Precision**: The ViT model's **precision reached 86.4%**, showing that 86.4% of images classified as "fake" were indeed deepfakes. This high precision underscores the ViT's effectiveness in accurately detecting deepfake images without misclassifying a substantial number of real images.
- **Recall**: With a **recall increase to 92.1%**, reflecting an 11.5% improvement from the original dataset performance, the ViT model successfully identified a majority of deepfake images. This elevated recall is essential in deepfake detection as it minimizes false negatives, ensuring that most manipulated images are identified.
- **F1-Score**: The **F1-score of 89.2%** represents a balanced performance, harmonizing precision and recall. This value emphasizes that the model maintains accuracy in both identifying deepfakes and reducing errors, supporting robust detection capabilities.
- **Test Loss**: The ViT model's **test loss dropped significantly to 0.2491**, marking a 49.1% reduction. This improvement in test loss suggests greater model reliability and accuracy due to the inclusion of GAN-generated synthetic data.
- **Confusion Matrix**: As shown in Figure 12, the ViT model's classification performance is represented by the following results: **1,697 true positives** for "real" images and **288 false negatives** (real images misclassified as fake), while for "fake" images, there are **1,836 true positives** and **149 false positives**. This matrix demonstrates the model's strong ability to distinguish between real and fake images, effectively balancing detection across both classes.

- **ROC Curve**: Figure 12 illustrates the ROC curve, with an area under the curve **(AUC) of 0.964**, underscoring the ViT model's exceptional performance in distinguishing between real and fake images. An AUC close to 1 indicates a reliable and robust detection system, highlighting the model's ability to correctly classify images.

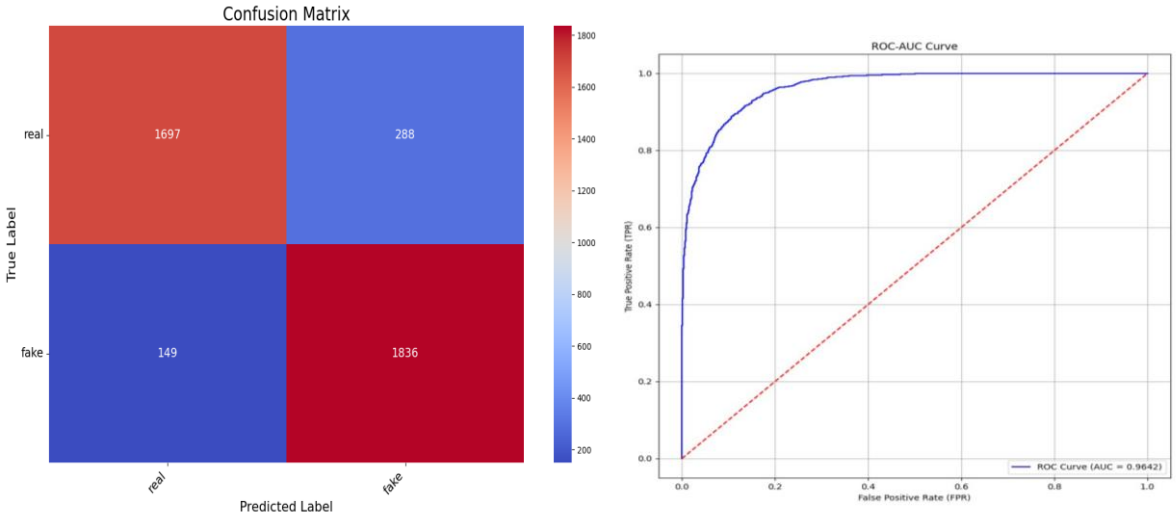

Fig 12. Confusion Matrix and ROC-AUC Curve of ViT Model

3.2.3 ViT Model Predictions (After Data Augmentation)

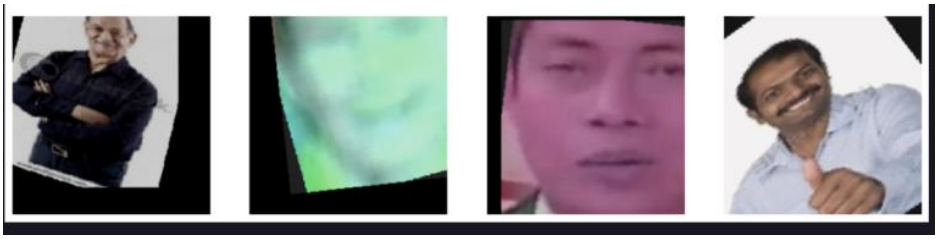

Fig 13. False Positives: Real images detected as fake

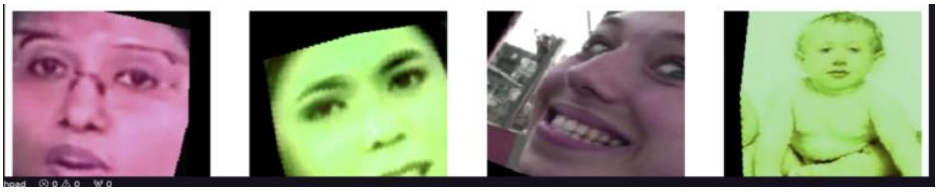

Fig 14. False Negatives: Fake images detected as real

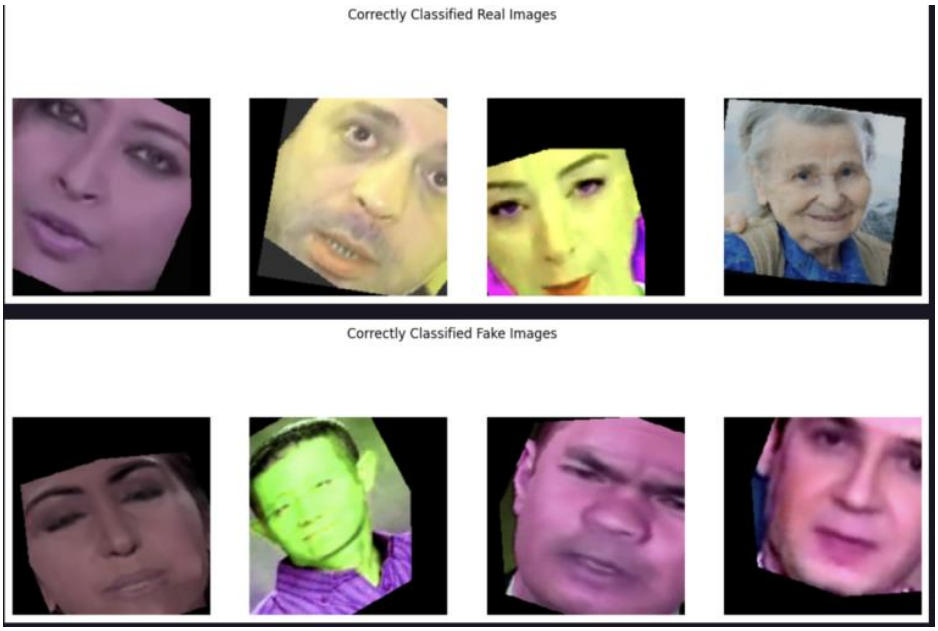

Fig 15. Correctly classified samples

Correctly classified real images show natural textures and consistent lighting, while correctly classified fake images exhibit artificial textures and exaggerated features as shown in Figure 15. Misclassified real images may suffer from poor lighting or resolution, leading to confusion as shown in Figure 13. Misclassified fake images have high-quality details and subtle color distortions that make them appear real as shown in Figure 14.

3.3. Inference Pipeline for Predicted Images

In our current analysis, the predicted images generated by both the ViT and CNN models were obtained after applying data augmentation techniques. This approach allowed us to assess the models' performance in a more robust manner, as the augmented data simulated a variety of real-world scenarios and enhanced the models' ability to generalize. However, to further evaluate the effectiveness of our models, we will establish an inference pipeline in the next phase of our project. This pipeline will enable us to obtain predictions from the models using the original, un-augmented images.

3.4. Significance of the Results

The outcomes of our deepfake detection project are essential for addressing the increasing threat of manipulated media in digital content and social platforms. By evaluating models like CNN and ViT, we provide crucial insights into effective detection techniques that can mitigate misinformation in high-stakes contexts, such as political campaigns and social media discourse.

Our findings show that the ViT model enhances deepfake detection, making it valuable for media organizations and social networks that aim to maintain information integrity and counter misinformation. Implementing reliable detection systems is vital for building trust in digital media. Accurate detection can help protect reputations and promote a well-informed public.

Based on our results, we plan to incorporate additional GAN-generated images to enhance the training dataset for the ViT model. This approach aims to improve the model's performance in detecting deepfakes and ensure its robustness. Once optimized, we will deploy the ViT model to provide effective real-time detection of manipulated media.

4. Discussions

4.1. Strengths and Limitations

4.1.1. CNN Model

Strengths: The CNN model demonstrates a **precision of 87.2%**, indicating its reliability in identifying fake images while minimizing the number of false positives. Additionally, the **ROC AUC score of 0.95** reflects its strong capability to effectively distinguish between real and manipulated content.

Limitations: Despite these strengths, the CNN model has a **recall of 86.1%**, which is lower than that of the Vision Transformer (ViT) model, resulting in some missed detections of fake images. Furthermore, with **180 real images misclassified** as fake, the model's false positives can potentially undermine trust in authentic content, emphasizing the need for improvement in accuracy.

4.1.2. ViT Model

Strengths: The ViT model excels with a **recall of 92.1%**, demonstrating its effectiveness in capturing a higher number of fake images compared to the CNN. Its **F1-score of 89.2%** indicates a balanced performance between precision and recall, making it a reliable choice for deepfake detection. The **ROC AUC score of 0.9642** further underscores its strong performance in distinguishing between real and fake images.

Limitations: However, the ViT model has a **precision of 86.4%**, which is slightly lower than that of the CNN, resulting in a higher number of false positives. Specifically, **288 real images were misclassified** as fake, highlighting the model's limitations in accurately identifying genuine content.

4.2. 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

Fig 16. Table showing comparison between CNN and ViT

Figure 16 consists of a table comparing both the models on various metrics, we can say that while the CNN model shows strengths in precision, the ViT model outperforms in recall, F1-score, and ROC-AUC, making it the better model for comprehensive deepfake detection in this comparison.

5. Conclusion

In conclusion, our deepfake detection project underscores the critical need for advanced methodologies in combating manipulated media. By evaluating and comparing the performance of CNN and Vision Transformers (ViT), we have identified effective strategies for enhancing deepfake detection capabilities. The promising results, particularly from the ViT model, highlight its potential to significantly improve the accuracy of detecting misinformation across various domains.

As we move forward, our commitment to incorporating more GAN-generated images into the training dataset aims to further refine the model's effectiveness. Additionally, we will delve deeper into understanding the specific features the model focuses on during classification. By analyzing prediction patterns and identifying factors behind any misclassifications, we can enhance the model's accuracy, especially in challenging cases. Ultimately, deploying

a robust detection system will be essential for preserving the integrity of digital content and supporting a well-informed public in an era increasingly challenged by deepfakes.

6. Statement of Contributions

Parth Malik : Implementation of Vision Transformers.
Shushant Ghosh : Implementation of GANs for image generation.
Suprajasai Konegari : Implementation of the CNN model.
Supriya Konegari : Responsible for data preparation and contributed to the CNN implementation.

7. References

[1] M. Alben Richards, E. Kaaviya Varshini, N. Diviya, P. Prakash, P. Kasthuri and A. Sasithradevi, "Deep Fake Face Detection using Convolutional Neural Networks," 2023 12th International Conference on Advanced Computing (ICoAC), Chennai, India, 2023, pp. 1-5, doi: 10.1109/ICoAC59537.2023.10250107.

[2] Belhassen Bayar and Matthew C. Stamm. 2016. A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer. In Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '16). Association for Computing Machinery, New York, NY, USA, 5–10. https://doi.org/10.1145/2909827.2930786.

[3] Wodajo, D. and Atnafu, S., 2021. Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126.

[4] Tolosana, R., Vera-Rodríguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection. ArXiv, abs/2001.00179.

[5] Preeti, Manoj Kumar, Hitesh Kumar Sharma,A GAN-Based Model of Deepfake Detection in Social Media,Procedia Computer Science,Volume 218,2023,Pages 2153-2162,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2023.01.191.